Semantic Adaptive Microaggregation of Categorical Microdata

Sergio Martínez*, David Sánchez, Aida Valls

Department of Computer Science and Mathematics. Universitat Rovira i Virgili

Intelligent Technologies for Advanced Knowledge Acquisition (ITAKA) research group

Av. Països Catalans, 26. 43007. Tarragona, Catalonia (Spain)

Abstract

In the context of Statistical Disclosure Control, microaggregation is a privacy preserving method aimed to mask sensitive microdata prior to publication. It iteratively creates clusters of, at least, *k* elements, and replaces them by their prototype so that they become *k*-indistinguishable (anonymous). This data transformation produces a loss of information with regards to the original dataset which affects the utility of masked data, so, the aim of microaggregation algorithms is to find the partition that minimises the information loss while ensuring a certain level of privacy. Most microaggregation methods, such as the MDAV algorithm, which is the focus of this paper, have been designed for numerical data. Extending them to support non-numerical (categorical) attributes is not straightforward because of the limitations on defining appropriate aggregation operators. Concretely, related works focused on the MDAV algorithm propose grouping data into groups with constrained size (or even fixed) and/or incorporate a basic categorical treatment of non-numerical data. This approach affects negatively the utility of the protected dataset because neither the distributional characteristics of data nor their underlying semantics are properly considered. In this paper, we propose a set of modifications to the MDAV algorithm focused on categorical microdata. Our approach has been evaluated and compared with related works when protecting real datasets with textual attribute values. Results show that our method produces masked datasets that better minimises the information loss resulting from the data transformation.

Keywords: Privacy protection, anonymity, microaggregation, MDAV, ontologies, semantic similarity.

---

* Corresponding author. Address: Departament d'Enginyeria Informàtica i Matemàtiques. Universitat Rovira i Virgili. Avda. Països Catalans, 26. 43007. Tarragona. Spain. Tel.: +34 977 256563; Fax: +34 977 559710; E-mail: sergio.martinezl@urv.cat.

1. Introduction

The publication of databases containing personal microdata about individuals is needed in some domains to facilitate research initiatives that, by means of data analysis tools, aim creating new valuable knowledge. For example, some medical organisations such as the Office of Statewide Health Planning and Development in California[i] provide patient data to perform market share analysis to define new health plans and insurance coverage or to study the quality of care. In a more general context, National Statistical Offices commonly publish responses collected in surveys that can be used for social or business policy-making.

Since most of these data refer to individuals, anonymisation methods are required to guarantee their privacy prior publication. The main goal of any anonymisation method is to hide or distort potentially identifying information found in input data to minimise the risk of disclosing the identity of individuals while retaining, as much as possible, the utility of data (similar conclusions can be extracted from the analysis of the original and anonymised datasets) (Domingo-Ferrer, 2008). In the past, most of the published datasets consisted exclusively of numerical data. Nowadays, motivated by the growth of the Information Society, non-numerical data, such as categorical attributes or even textual responses, are commonly collected and published for example medical conditions and treatments or personal preferences. The anonymisation of this last kind of data which is the focus of this paper, has been less studied and presents additional challenges in comparison to numerical datasets (Torra, 2011).

In the literature, we can find several families of methods dealing with the anonymisation of non-numerical data, which can be distinguished according to the degree of structuration of the input data. First, *document sanitisation* methods (Chakaravarthy et al. , 2008; Wei et al. , 2009) deal with unstructured textual documents, which are individually anonymised by hiding potentially identifying terms, for example addresses or names. Other methods (He and Naughton, 2009; Terrovitis et al. , 2008) deal with *set-valued data*, consisting of textual records of transactions collected about individuals as for example query logs or customer's commodities. In this context, data is horizontally

aggregated to avoid unique combinations of values that may disclose the identity of an individual. Finally, works like ours, framed in the area of *Statistical Disclosure Control* (SDC) of databases, deal with structured databases, in which rows correspond to records and columns to single-valued attributes (Domingo-Ferrer, 2008; Herranz et al. , 2010; Jin et al. , 2011; Oliveira and Zaïane, 2007; Shin et al. , 2010; Willenborg and Waal, 2001). Non-numerical values usually correspond to categorical attributes, taking values from a finite set of modalities, for examplecity of living or job. In general, the attributes are independent so that they can be treated separately and can be classified according to the security point of view. First, *identifier* attributes such as ID-card numbers are removed from the dataset. Second, *non-confidential quasi-identifier attributes* as for example job, city of living or age, even though they do not link to specific individuals separately, are modified to avoid privacy disclosure when they are considered in groups such as job + city of living may provide unique -identifying- value tuples. The performed modifications pursue to minimise the trade-off between the *information loss* resulting from the masking process, so that the utility of data can be preserved up to a degree, and the minimisation of the re-identification risk, typically tackled by fulfilling the k-anonymity property (Domingo-Ferrer, 2008). A masked dataset is considered *k*-anonymous if any record is indistinguishable from, at least, *k-1* other ones (Sweeney, 2002a; b).

Among the plethora of anonymisation methods, *microaggregation* stands as a natural approach to satisfy the *k*-anonymity property in statistical databases (Domingo-Ferrer and Torra, 2005). 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. Due to the search for the optimal partition when considering multivariate data is NP-hard (Oganian and Domingo-Ferrer, 2001) sub-optimal heuristic methods have been proposed. One the most popular ones is the MDAV (Maximum Distance Average Vector) method (Hundepool et al. , 2003), because it provides high

[i] http://www.oshpd.ca.gov/HID/Dataflow/index.html

quality aggregations without being constrained by some configuration parameters, as other methods do (Domingo-Ferrer et al. , 2006).

Even though the MDAV method has been applied/adapted to non-numerical categorical datasets (Abril et al. , 2010), the fact that it was originally designed to deal with numerical values imposes some limitations that, as it will be discussed in the next section, negatively affect the utility of the output data. In this paper, we propose several algorithmic and design modifications that, by considering the distributional characteristics of categorical attributes, aim to minimise the information loss resulting from the anonymisation process. 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 the *k*-anonoymity property. Our approach has been evaluated from different perspectives using real datasets. Results show that the proposed modifications better minimise the loss of semantics of the masked data than related works, while retaining, or even improving the computational scalability with large datasets.

The rest of the paper is organised as follows. In Section 2, a review of related literature is done, presenting the main anonymisation approaches dealing with non-numerical data, the different methods used to build clusters in microaggregation methods and, finally, works related to the MDAV method, discussing their limitations. Section 3 presents and formalises the proposed modifications to the MDAV method aimed at minimising the information loss of the anonymisation of categorical data. Section 4 tests and compares our proposal with the related works using two datasets with different characteristics: a set obtained from a survey at the Delta de l'Ebre National Park in Catalonia (Spain) with around 1,000 registers with a large diversity on the values, and the Adult Census dataset from the UCI repository, which is larger (around 30,000 registers) but more homogenous. The final section presents conclusions and future work.

2. Related Work

In this section, we review related works on the anonymisation of non-numerical data, stating the starting point for our research. First, we briefly outline the main approaches to text anonymisation, that is sanitisation, which focus on avoiding the identification of individuals' private data in documents or in transactional data set-valued data. For each one, we state the main differences against the methods framed on SDC in databases, which is the focus of this paper. Then, different micro-aggregation approaches available in the literature are discussed and compared against the MDAV algorithm, which centres the attention of our work. After that, the MDAV algorithm is presented in detail, discussing the limitations of the existing versions when dealing with categorical data.

2.1. Anonymisation of unstructured textual data

Methods dealing with unstructured textual data usually aim at finding and hiding personal identifiable information in narrative text that is individual documents (Meystre et al. , 2010). The goal is to hide sensitive parts of text while avoiding unnecessary distortion, so that the document continues being readable and useful after the modifications. To keep the meaning of the hidden text, sanitisation methods have been proposed (Wei et al. , 2009) based on generalising the sensitive words. Authors rely on a knowledge structure that represents the concepts of the domain for example an ontology, which is used to change the sensitive values by other that are more general. All the possible generalisations are generated and a suitable combination is selected based on a pruning strategy. A *t*-Plausibility measure is used as stopping criterion. A sanitised document *d* is *t*-plausible if at least *t* documents, including the original one can be generalised to *d*, based on the same reference ontology. In (Chakaravarthy et al. , 2008), it is assumed that an adversary knows a set of context terms associated to some individual. A sanitised document is secured if the adversary cannot match the terms in the document with the context terms that he knows about the protected entity.

In this type of problems, the goal is to protect a *single* text document referring to some particular individual. This is an important difference with respect to the problem faced in SDC, which deals with

structured databases of records from *different* individuals. Therefore, even though the masking of individual values follows similar principles, that is a knowledge base is exploited to propose data transformations that minimises the loss of semantics, the masking schemas are different: text sanitisation hides individual sensitive values at *document level*, whereas SDC methods *aggregate* data of *different* individuals to make them indistinguishable.

2.2. Anonymisation of set-valued data

Sometimes, published contents consist of a collection of transactional data corresponding to specific individuals. This is known as set-valued data, in which each record contains variable-length multi-valued data corresponding to an individual, such as lists of commodities bought by a customer (Terrovitis et al. , 2008), query logs performed by a user of a Web search engine (He and Naughton, 2009) or outcomes of a clinical record (He et al. , 2008). Even after removing all the personal characteristics for example names or ID-card numbers from the dataset, the publication of such data is still risky on attacks from adversaries who have partial knowledge about the individual's actions.

In this context, unique combination values are transformed to reduce the risk of re-identification of individuals while keeping, as much as possible, the utility of the data. To anonymise this type of datasets, authors base their works on a generalised definition the *k*-anonymity property, known as the *km*-anonymisation model (Terrovitis et al. , 2008). Assuming that the maximum knowledge of an adversary is at most *m* items in a given transaction, the goal is to build sets of *k* transactions that cannot be distinguished on the basis of these *m* items. In (Terrovitis et al. , 2008) authors present a method based on generalising the original values according to Value Generalisation Hierarchies (VGH) constructed ad hoc for the considered domain. As acknowledged by the authors, the optimal generalisation that minimises the information loss, according a metric based on the number of generalisation steps is NP-hard. To provide a more scalable solution, authors propose a greedy heuristic sub-optimal. In (He and Naughton, 2009), authors present a similar solution but starting from the most abstract generalisation, the one with the highest information loss and progressively specialising it in performing sub-optimal data partitions.

Even though set-valued data is semi-structured, the problem tackled in this framework is quite different to the one faced in the context of SDC (Terrovitis et al. , 2008). Set-valued datasets consist of variable-length lists of values describing the same feature, which are anonymised as a unique feature describing the corresponding individual. On the contrary, databases are organised as a *set* of attributes, each one corresponding to a *different* feature of the described entity. The anonymisation process, in consequence, focuses on attribute's value *combinations* that can identify individuals *quasi-identifiers*. Attributes are usually single-valued and independent between them, so that the anonymisation process can manage each attribute independently. Thanks to the availability of structured data in the form of attribute-value pairs for each individual, anonymisation methods based on clustering or microaggregation are naturally applicable in SDC scenarios as it will be discussed in the next section.

2.3. Anonymisation of structured databases

Privacy preservation in structured databases is framed in the Statistical Disclosure Control discipline (Domingo-Ferrer, 2008; Herranz et al. , 2010; Jin et al. , 2011; Oliveira and Zaïane, 2007; Shin et al. , 2010; Willenborg and Waal, 2001). From the diverse methodologies tackling the masking of quasi-identifier attribute values in a database, *microaggregation* methods stand out as a natural solution to group structured data as a means to anonymise them, achieving good results in data utility and disclosure risk when compared to other approaches (Domingo-Ferrer, 2008; Herranz et al. , 2010). In fact, microaggregation methods satisfy the *k*-anonymity property (Domingo-Ferrer and Torra, 2005) per se, because they build clusters of at least *k* records that are substituted by their centroid becoming indistinguishable. Fulfilling the anonymisation property, the goal of privacy-preserving microaggregation methods is to find the record partition that minimises the information loss measured as the relative distance between cluster members and their centroid using the Sum of Square Errors, SSE (Abril et al. , 2010; Domingo-Ferrer et al. , 2006; Domingo-Ferrer and Mateo-Sanz, 2002; Lin et al. , 2010; Torra and Miyamoto, 2004). Note that this optimisation goal is different to grouping data mining algorithms such as clustering that focus on discovering the underlying classification of the

objects according to their common features that is, in clustering methods, intra-group distances must be minimised, while inter-group distances must be maximised.

Again, because the search for the optimal partition when considering multivariate data is NP-hard (Oganian and Domingo-Ferrer, 2001), sub-optimal heuristic methods have been proposed. First, we can find methods that adapt existing clustering algorithms, mainly framed in the Data Mining field to the constrained-sized clusters needed to solve the *k*-anonymous microaggregation problem. In general, these methods first apply a clustering strategy and, after that, re-arrange resulting clusters not fulfilling the *k*-anonymity, those with cardinality below *k*. Since the re-organisation of clusters in an optimal way is again NP-hard (Lin and Wei, 2008), authors propose different sub-optimal heuristics to perform this task. In (Chiu and Tsai, 2007), authors first randomly select n/k records and assign all records to their closest clusters minimising the information loss. Then, those clusters with less than *k* records are merged until they fulfil the *k*-anonymity. In (Lin and Wei, 2008), authors propose an approach derived from the K-Means algorithm. During the clustering step, the algorithm randomly selects n/k records to build clusters. The closest records are then clustered, based on the distance to the cluster centroid. After that, to fulfil the *k*-anonymity, records belonging to clusters with more than *k* records are moved to clusters with less than *k* records.

The quality, that is the information loss, of the above methods is hampered by the random selection of cluster seeds and the cluster re-arrangement strategy. To minimise their influence, microaggregation methods that directly cluster data in groups of at least *k* records have been proposed (Byun et al. , 2007; Hundepool et al. , 2003; Loukides and Shao, 2007). From these, the MDAV (*Maximum Distance Average Vector*) method (Hundepool et al. , 2003), on which this paper focuses, is one of the most popular ones because it provides high quality results (Domingo-Ferrer et al. , 2006) without relying on random cluster seeds.

2.4. The MDAV microaggregation method

The MDAV method is based on generating clusters of *k* elements around records selected according to their distance with respect to the global centroid of the dataset, which avoid relying on random

cluster seeds, that might provide less accurate results. The centroid is calculated by using an averaging operator on the values of the records. Sinceclusters already fulfil the *k*-anonymity property, no record re-arrangement is needed at the end. For the numerical case, the Euclidean distance and the arithmetic average are the usual operations applied in MDAV (Domingo-Ferrer and Torra, 2005; Hundepool et al. , 2003). MDAV has been commonly used in the past to protect microdata due to its performance in comparison with other methods (Abril et al. , 2010; Domingo-Ferrer et al. , 2006; Domingo-Ferrer et al. , 2008; Domingo-Ferrer and Torra, 2005; Erola et al. , 2010; Huang et al. , 2010; Lin et al. , 2010; Nin et al. , 2008a).

The behaviour of the MDAV method is detailed in algorithm 1. First, the centroid of the dataset is calculated and the most distant object *r* (selected by means of a distance measure appropriate for the type of data) is selected. 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.

Algorithm 1. MDAV

```
Inputs: D (dataset), k (level of anonymity)
Output: D^A (a transformation of D that satisfies the k-anonymity level)
1   D^A = D
2   while (|D| ≥ 3*k) do
3      Compute the centroid X̄ of all records in D;
4      Consider the most distant record r to the centroid X̄
5      Find the most distant record s to the record r
6      Form a cluster in D^A with r and the k-1 closest records to r
7      Remove these records from D
8      Form a cluster in D^A with s and the k-1 closest records to s
9      Remove these records from D
10  end while
11  if (|D| ≥ 2*k) then
12     Compute the centroid X̄ of remaining records in D;
13     Find the most distant record r to the centroid X̄
```

```
    Form a cluster in D^A with r and the k-1 closest records to r
    Remove these records from D
end if
Form another cluster in D^A with the remaining records
Compute the centroid X̄_i for each cluster in D^A
Replace all original values in each cluster in D^A with its centroid X̄_i
Output D^A
```

2.5. Limitations of MDAV when dealing with categorical data

The basic MDAV method, however, presents some limitation that may hamper the utility of anonymised 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 SSE and, hence, the information loss resulting from the replacement of the individual records by the cluster centroid (Domingo-Ferrer and Mateo-Sanz, 2002). The possibility of varying the size of the clusters ensuring a minimum cardinality of k to fulfil the k-anonymity property, would be preferable because it allows a better adaptation of the clusters to the data distribution. This is especially relevant for *categorical* data as for examplejob 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.

Fig. 1 shows the advantage of variable-sized clustering in a two-dimensional space. Using a fixed-size microaggregation with $k = 3$, the data is grouped in three clusters, Fig. 1-A, one of them, cluster 2 composed by three very distant elements, which have been put together just to have clusters of size equal to 3. Moreover, allowing variable-sized clusters, two clusters are formed with five elements on the left and four elements on the right, Fig. 1-B. This second clustering seems much more natural, with more homogeneous clusters.

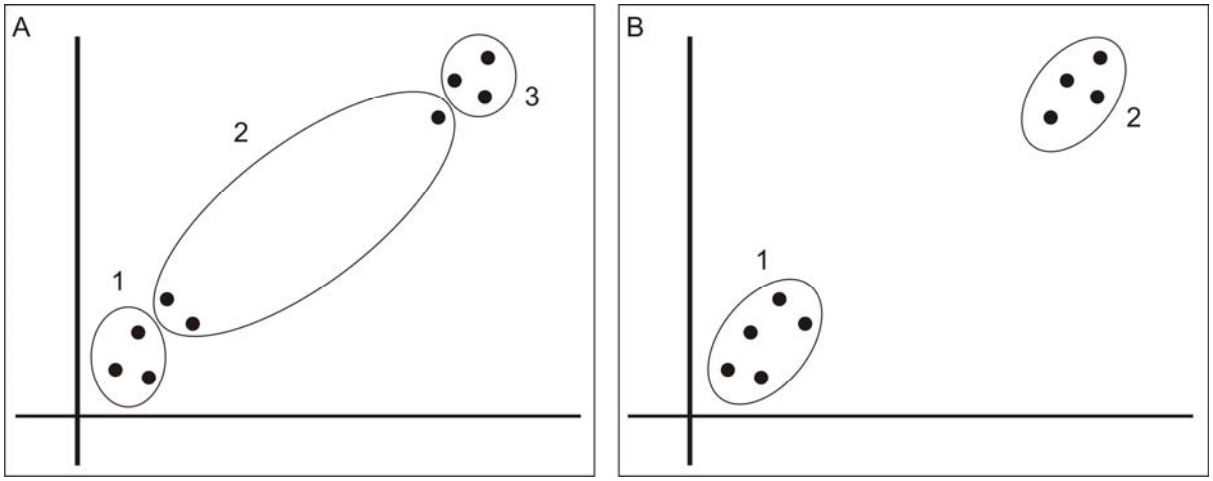

Fig. 1. A comparative example of microaggregation with $k = 3$. **A.** Fixed-sized clustering. **B.** Variable-sized clustering.

Some authors have proposed modifications of the MDAV algorithm to support variable-sized clusters (Domingo-Ferrer and Mateo-Sanz, 2002; Laszlo and Mukherjee, 2005; Lin et al. , 2010). However, on one hand, all of them focus on continuous numerical data and, on the other hand, the maximum size of the clusters is constrained to $2k - 1$.

In addition to the restrictions regarding the size of the clusters, the results are influenced by the two operators needed during the microaggregation: the *distance measure*, used to compare records and centroids (lines 4, 5, 6, 8, 13 and 14 of algorithm 1), and the *centroid construction,* needed to calculate the global centroid at each iteration and to select the representative record for each cluster (lines 3, 12 and 18 of algorithm 1). The first applications of MDAV considered only numerical attributes. Numbers define a continuous scale of infinite values, which can be compared and transformed by means of mathematical operators. This facilitates the processing of input data, so that distortions needed during data microaggregation can be introduced while maintaining the statistical properties of the dataset.

Non-numerical categorical data, on the contrary, take values from a discrete, finite and typically reduced list of modalities which are commonly expressed by words. Since arithmetic functions cannot be applied to this kind of data, a simple method to apply MDAV to categorical data consists on using Boolean equality/inequality operators (Domingo-Ferrer and Torra, 2005; Torra, 2004). Thus, the

distance between two values is defined as 0 if the attribute values are identical and 1 otherwise. The original records are substituted at the end of the algorithm by the most frequently occurring value in the cluster, mode.

This simplistic treatment of data, neglect one of the most important dimensions of non-numerical data: semantics. In fact, the preservation of data semantics plays a crucial role to ensure the utility of the masked results (Martinez et al. , 2011; Torra, 2011). Hence, considering the semantics of concepts when evaluating similarity to construct clusters and when selecting centroids may improve the quality and hence, the utility of the masked data file.

To make a semantic interpretation of textual values, algorithms require some sort of structured knowledge source that represents the relations between words at a conceptual level. As stated in the introduction, some authors (He and Naughton, 2009; Terrovitis et al. , 2008) rely on ad-hoc Value Generalisation Hierarchies (VGH) to implicitly interpret and transform, by means of taxonomical generalisations, non-numerical values. Other authors (Abril et al. , 2010) exploit more general knowledge structures (ontologies) to assist the semantic interpretation of data without depending on potentially biased or overspecified VGHs (Martínez et al. , 2010). By exploiting these knowledge bases, semantic similarity measures can be defined to assess the alikeness of a pair of terms according to the semantic evidences extracted from them (Batet et al. , 2011; Rada et al. , 1989; Sánchez and Batet; Sanchez et al. , 2011; Sanchez et al. , 2010; Wu and Palmer, 1994).

Recently, some authors have considered the semantics of textual data during the microaggregation process. In (Abril et al. , 2010), the MDAV algorithm is applied to textual attributes, computing the distance between records using the Wu & Palmer similarity measure (Wu and Palmer, 1994) and WordNet (Fellbaum, 1998) as the ontology. The Wu & Palmer measure evaluates the similarity between two concepts (c_1 and c_2) as the inverse of the number of semantic relationships needed to go from c_1 to c_2 in the background ontology (Eq. 1). This is normalised according to the depth of their Least Common Subsumer (LCS), the most specific ancestor that generalises the two concepts. They also take into account that the depth of the concepts represents different degrees of generality.

$$similarity_{w\&p}(c_1, c_2) \frac{2 \times N_3}{N_1 + N_2 + 2 \times N_3} \quad (1)$$

where N_1 and N_2 are the number of is-a links (taxonomical relations between specialisations and generalisations from c_1 and c_2, respectively, to their LCS, and N_3 is the number of is-a links from the LCS to the root of the ontology. This ranges from 1 for identical concepts to 0. Hence, this similarity measure is converted into a distance function as follows:

$$dis_{w\&p}(c_1, c_2) = 1 - similarity_{w\&p}(c_1, c_2) \quad (2)$$

In (Abril et al. , 2010) this measure is used to select the most distant record to the centroid, lines 4, 5 and 13 in the MDAV algorithm, and to create a cluster with the *k-1* nearest ones, lines 6, 8 and 14. As a result, terms are grouped into clusters according to their semantic similarity.

Regarding the centroid calculus, in (Abril et al. , 2010) the centroid of both the whole dataset, lines 3 and 12 and the resulting clusters, line 18 is the LCS subsuming all the values on the cluster. The rationale is that the LCS represents the semantic content that all the concepts in a cluster have in common. This is similar to the strategy proposed by approaches based on value generalisations (He and Naughton, 2009; Terrovitis et al. , 2008), in which values are partitioned and replaced by their common generalisation. This contrasts to approaches dealing with textual data in a categorical fashion (Domingo-Ferrer and Torra, 2005; Torra, 2004), which select the centroid according to the distributional – rather than semantic – characteristics of data. Even though the terms semantics are considered, the use of the LCS as centroid has some drawbacks. First, the presence of outliers, terms referring to concepts which are semantically far to the majority of elements in the cluster will cause the LCS to become a very general concept, that is in the worst case, the root of the taxonomy. Moreover, the substitution of the values of a cluster by such as general concept implies a high loss of semantic information. Finally, the frequency of appearance of words is not considered during the centroid selection and hence, a scarce term is considered as important as the common ones, biasing the results.

3. A new proposal: Semantic Adaptive MDAV (SA-MDAV)

As discussed above, related works adapting the MDAV algorithm to categorical data presented several limitations that negatively affected the information loss resulting from the microaggregation process. To overcome some of them, in this section, we propose a set of modifications to both the MDAV algorithm and the underlying methods (centroid and distance calculus). These changes are based on the intrinsic and distributional properties of categorical data. The goal is to microaggregate data into highly cohesive clusters, in order to minimise the information loss resulting from the masking process. Concisely, the proposed modifications focus on the following aspects:

- Adaptive microaggregation: as stated in section 2, due to the discrete nature of categorical data, it would be desirable to consider their distribution to create cohesive clusters. In section 3.1, we propose a modification of the MDAV algorithm that, while ensuring the *k*-anonymity property, creates clusters of different size according to the data distribution.
- Semantic weighted distance: considering that textual data should be interpreted according to their underlying semantics (Torra, 2011), in section 3.2, we propose a semantic weighted distance to guide both the centroid calculus and the cluster construction process. It considers both the meaning of the values given by a background knowledge base and the distribution of those values.
- Semantic centroid: as stated in section 2, the construction of an appropriate and representative centroid is crucial to guide the microaggregation process and to minimise the information loss. In section 3.3, we propose a semantically-grounded method to calculate the centroid of multivariate categorical datasets, exploiting a knowledge base as well as considering the data distribution.

3.1. Adaptive microaggregation

As stated in section 2, clusters with adaptable size are desirable to better cope with the data distribution. Due to the discrete nature of categorical data (gender or city-of-living), values usually

define a limited set of modalities that tend to repeat. Because the distance between identical values is zero, it would be very convenient to include all of them in a single cluster to improve its cohesion and hence, minimise the information loss as it will be shown in the evaluation section. This is done even though the number of repetitions could be higher than the usual upper bounds, like k, for fixed-sized microaggregation approaches (Abril et al. , 2010; Domingo-Ferrer and Torra, 2005) or $2k-1$, for variable-sized ones (Domingo-Ferrer and Mateo-Sanz, 2002; Lin et al. , 2010). Following this premise, the proposed microaggregation algorithm will focus on putting all the records that have 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, the frequency of appearance of the values. This incorporates the benefits of variable-sized cluster-based anonymisation methods (Chiu and Tsai, 2007; Lin and Wei, 2008) discussed in the introduction, but without being hampered by the random selection of cluster seeds or the posterior cluster re-arrangement stage. To do this, we manage the original data set as follows.

Let us take a univariate input dataset with a single categorical attribute V. We will represent the information as a tuple of the form: $V = \{<v_1, \omega_1>,\ldots,<v_p, \omega_p>\}$, where $<v_i, w_i>$ define a *value tuple* in which v_i is each distinct term found in the dataset and ω_i is its number of repetitions.

Example 1. Given the dataset $V = \{v_1, v_2, v_1, v_3, v_3, v_1, v_4\}$, we represent it as $V = \{<v_1, 3>, <v_2, 1>, <v_3, 2>, <v_4, 1>\}$.

This can be generalised for multivariate datasets as follows. Let us take MV a multivariate dataset with p indistinguishable records and m attributes. We represent it as $MV = \{<\{v_{11},\ldots,v_{1m}\}, \omega_1>,\ldots,<\{v_{p1},\ldots v_{pm}\}, \omega_p>\}$, where each value tuple $\{v_{i1},\ldots,v_{im}\}$ represents a distinct combination of m attribute values and ω_i is its number of occurrences in the dataset.

Example 2. Given the dataset with two attributes and three different tuples, such as $MV = \{\{v_{11}, v_{21}\}, \{v_{13}, v_{23}\}, \{v_{13}, v_{23}\}, \{v_{11}, v_{21}\}, \{v_{12}, v_{23}\}, \{v_{13}, v_{23}\}, \{v_{11}, v_{21}\}, \{v_{11}, v_{21}\}, \{v_{13}, v_{23}\}, \{v_{11}, v_{21}\}, \{v_{13}, v_{23}\}, \{v_{11}, v_{21}\}, \{v_{11}, v_{21}\}\}$, we will represent it as $MV = \{<\{v_{11}, v_{21}\}, 7>, <\{v_{12}, v_{23}\}, 1>, <\{v_{13}, v_{23}\}, 5>\}$.

Following Example 2, Fig.1 shows the advantage, compared with related works, of using the MDAV method, including the adaptation of the size of the clusters by considering only a lower bound -for *k*-anonymity- but not an upper bound. First, given a *k*-anonymity level of $k=3$ and using a fixed-size microaggregation, Fig. 2-A, the individuals are grouped in four clusters. Due to having an upper bound of 3, individuals with the same values are separated into different clusters, such as $\{v_{13}, v_{23}\}$. As a result, we obtain less cohesive clusters, such as cluster 3, which implies a high information loss when the original values are replaced by their centroid, at the end of the algorithm. Using variable-sized clusters with up to $2k-1$ elements, Fig. 2-B, the cluster size can vary between 3 and 5. In this case, clusters may incorporate a higher amount of repetitions, like clusters 1 and 3 but, if $\omega_i > 2k-1$, which is the case of the tuple $<\{v_{11}, v_{21}\}, 7>$, again, some of the indistinguishable records will be placed in another cluster. As a result, a less cohesive cluster, cluster 2, is obtained with the remainder elements. Finally, adapting the cluster size to the data distribution, Fig. 1-C, the two resulting clusters are the most cohesive because they can accommodate all the repetitions into the same cluster ($<\{v_{11}, v_{21}\}, 7>$ and $<\{v_{13}, v_{23}\}, 5>$). Note that, for value tuples with a $\omega_i < k$ (which is the case of $<\{v_{12}, v_{23}\}, 1>$), those are included in the closest cluster, cluster 2 in this case, as stated in the MDAV algorithm.

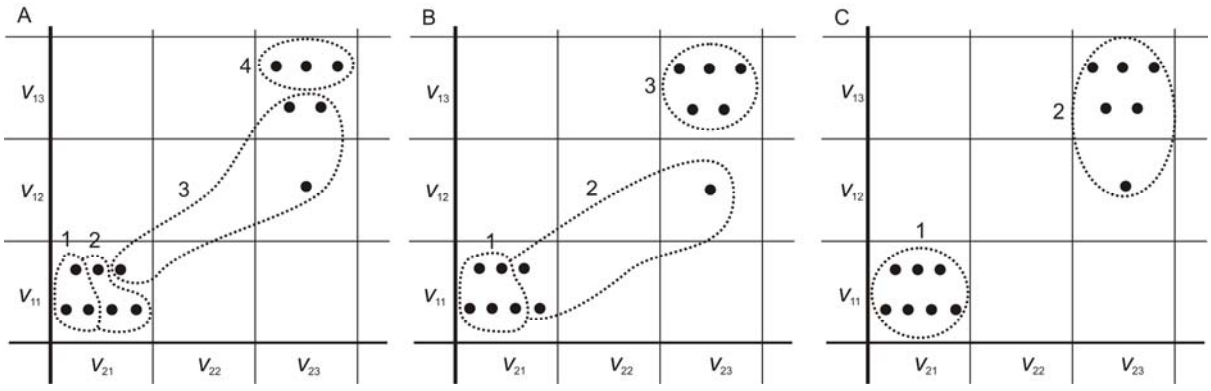

Fig. 2. An example of microaggregation with $k = 3$. **A.** Fixed-sized clustering. **B.** Variable-sized clustering with a maximum size of $2k-1$. **C.** Adaptive clustering without maximum size restriction.

To incorporate this adaptive behaviour during the clustering construction, the MDAV method, shown in Algorithm 1, has been modified as shown in Algorithm 2. We called the algorithm *Semantic Adaptive MDAV* (SA-MDAV).

Algorithm 2. SA-MDAV

```
Inputs: D (dataset), k (level of anonymity)
Output: D^A (a transformation of D that satisfies the k-anonymity level)
1   D^A = D
2   while (|D| ≥ k) do
3    Compute the centroid X̄ of all tuples in D;
4    Consider the most distant tuple r to the centroid X̄
5    Form a cluster C in D^A with the tuple r. Calculate centroid C̄
6    Remove this tuple from D
7    while (|C| < k) do
8      Add to cluster C the closest tuple in D^A to the cluster centroid C̄
9      Remove this tuple from D
10     Calculate the new centroid C̄ of cluster C
11   end while
12   if (|D| ≥ k) then
13     Find the most distant tuple s to tuple r
14     Form a cluster C in D^A with the tuple s. Calculate centroid C̄
15     while (|C| < k) do
16      Add to cluster C the closest tuple in D^A to the cluster centroid C̄
17      Remove this tuple from D
18      Calculate the new centroid C̄ of cluster C
19     end while
20   end if
21  end while
22  Add each remaining tuple in D to their closest cluster in D^A
23  Output D^A
```

While the core of the algorithm remains as in the original MDAV, our proposal incorporates several modifications.

First, to support adaptive-sized clusters without a maximum size bound, it is required to check if enough records are available at the beginning of each aggregation step (lines 2 and 12) to create a new *k*-anonymous cluster with a minimum of *k* elements. When there are not enough remaining elements,

as done in the original method, each is added to the closest cluster (line 22). Second, we propose a modification in the procedure of creation of a cluster. In the original MDAV algorithm, once the most distant record *r* to the centroid is found (line 4 in Algorithm 1), the closest records to *r* are iteratively joined to create a cluster, considering *r* as a static centroid. However, when new records are joined into a cluster, the real centroid should change according to the distribution of the objects in the space. To tackle this issue, our algorithm recalculates the centroid of the cluster being constructed whenever a new element is added (lines 10 and 18). This behaviour has been implemented in classical clustering algorithms such as the *K-means* (Macqueen, 1967) and implies that the centre of the cluster is displaced in each iteration, creating more cohesive clusters. This will further contribute to minimise the information loss when replacing cluster elements by their centroid, as it will be shown in the evaluation section.

Note also that, as formalised above, input data is managed according to distinct value tuples with associated frequencies of appearance. Considering the algorithm design, this results in a computation cost of $O(p^2)$ where p is the number of distinct tuple values. On the contrary, the classic MDAV algorithm manages data according to individual records, resulting in a cost of $O(n^2)$, where n is the number of records. By definition, $p \leq n$; considering that, as stated in the introduction, categorical data is characterised by a limited and typically reduced set of modalities, in a real scenario it would be very common that $p << n$. As a result, even considering the overhead of the algorithmic refinements proposed below, the scalability of our method is ensured and even improved in comparison to the basic MDAV.

Algorithmically, the other steps of SA-MDAV are the same as in the original method. The underlying similarity measure and the centroid construction method, however, have been modified to better consider the characteristics of categorical data. In short, we propose to use a semantic weighted distance function to discover the most distant (lines 4 and 13) and closest values (lines 8, 16 and 22), and a semantically-grounded method for constructing the centroid, both as a centre of the dataset, line 3, and as the centre of each cluster, lines 10 and 18. The following sections formalise the semantic distance measure, section 3.2, and the centroid construction method, section 3.3.

3.2. Weighted semantic distance for categorical data

Semantics should be considered to properly interpret non-numerical attributes (Torra, 2011) so that masked data may preserve, as much as possible, the meaning of original data. Moreover, data distribution should be preserved when transforming data in privacy-preserving methods (Domingo-Ferrer, 2008). Until now, either the distribution (Torra, 2004) or the semantics (Abril et al. , 2010) of data have been considered when evaluating the distance between records in the MDAV algorithm.

In this section, we present a distance measure that integrates both aspects to properly evaluate and manage categorical data. Due to our goal of minimising the information loss by creating highly cohesive clusters, we propose to semantically compare two tuples of values, weighting their semantic distance by their number of repetitions. In this manner, as shown in Algorithm 2, both the most semantically distant and/or the most frequent tuple to the centroid of the dataset is selected in the first place (line 4). Moreover, this procedure permits to initiate the construction of a cluster incorporating all value repetitions, assuring a maximum cohesion. After that, value tuples with the closest distance and lowest amount of repetitions are selected to further extend the cluster if necessary (lines 8 and 16), minimising the cluster heterogeneity.

As introduced in section 2, to interpret data semantics, we rely on semantic similarity measures, which evaluate the taxonomical resemblance of terms according to the knowledge provided by a background ontology. As in (Abril et al. , 2010), we use the Wu and Palmer similarity measure (Wu and Palmer, 1994) and WordNet as the ontology, so that our results can be objectively and unbiasedly compared to related works.

Definition 1. The weighted semantic distance (wsd_O) between two univariate tuples ($<v_1, \omega_1>$, $<v_2, \omega_2>$), computed from the ontology O, is defined as:

$$wsd_O(<v_1,\omega_1>,<v_2,\omega_2>) = \sum_{i=1}^{\omega_1}\sum_{j=1}^{\omega_2} dis_{w\&p}(v_1,v_2) = (\omega_1 \cdot \omega_2)\cdot dis_{w\&p}(v_1,v_2) \qquad (3)$$

where the function $dis_{w\&p}$ is the semantic distance expressed in the Eq. 2 (based on Wu & Palmer similarity, Eq. 1) and ω_1 and ω_2 are the number of repetitions of v_1 and v_2 respectively. Note that, to

aggregate all individual distance values between elements of value tuples with multiple repetitions (so that data distribution is also considered), their respective appearance frequencies are multiplied.

This measure can be generalised to multivariate data as follows:

Definition 2. The distance between two multivariate objects ($<\{v_{11},\ldots,v_{1m}\},\omega_1>$, $<\{v_{21},\ldots v_{2m}\},\omega_2>$), computed from the ontology O, is defined as the average of the distances between individual values, as follows:

$$wsd_O(<\{v_{11},...,v_{1m}\},\omega_1>,<\{v_{21},...,v_{2m}\},\omega_2>)=\sum_{j=1}^{m}\frac{wsd_O(<v_{1j},\omega_1>,<v_{2j},\omega_2>)}{m} \quad (4)$$

where v_{1j} and v_{2j} are the values of the jth attribute of the objects $\{v_{11},\ldots,v_{1m}\}$ and $\{v_{21},\ldots,v_{2m}\}$, respectively.

In addition to guide the clustering construction, as it will be shown in the next section, the distance measure will be used to assist the centroid construction so that both the semantics and the data distribution are considered.

3.3. Constructing the centroid for categorical data

The MDAV algorithm and the SA-MDAV version rely on centroids to guide the microaggregation process and to replace groups of data by their corresponding representative values, achieving a *k*-anonymous dataset. Hence, the selection of an accurate centroid is crucial to construct cohesive clusters and to minimise the information loss resulting from the data transformation. Centroid construction for categorical data is challenging because the definition of appropriate averaging operators on non-numerical data is not straightforward. In section 2, several approaches to centroid construction for categorical data were discussed. On one hand, centroids computed solely according data distribution, such as the mode (Domingo-Ferrer and Torra, 2005; Torra, 2004), omit the semantics of data and, hence, a crucial dimension of data utility. Moreover, centroids are constrained

to values appearing in the input dataset. On the other hand, pure semantic centroids that are based on the ontological concept that subsumes all the values in the cluster, the LCS (Abril et al. , 2010), are affected by outliers and thus, they commonly suffer from too abstract generalisations, resulting in a high information loss.

In this section, we propose a centroid calculation method for multivariate categorical data that considers, in an integrated manner, both the semantics and the distribution of the data. Moreover, the background knowledge base is exploited not only to assess the semantic distance between terms, but to retrieve centroid candidates.

To calculate the centroid, we assume the classical centroid definition as the value (or tuple, in the case of multivariate data) that minimises the distance against all the elements in a set. Formally, given a distance function d, the centroid of a set of values $\{v_1, v_2, \ldots, v_n\}$ is defined as:

$$centroid(v_1, v_2, \ldots, v_n) = arg\ min_c \left\{ \sum_{i=1}^{n} d(c, v_i) \right\} \quad (5)$$

where c is a centroid candidate for the set of arguments.

This is a relevant difference to related works, which do not incorporate the notion of *distance* during the centroid construction.

A second relevant difference of our approach concerns the search space for constructing the centroid. When selecting the centroid according to the frequency of values, mode, the number of centroid candidates is limited to the set of different values that appear in the cluster. On the contrary, using an, ideally detailed, ontology such as WordNet, the search space can extended to all concepts modelled in the ontology and hence, the centroid can be constructed from a finer grained set of candidates. The search can be limited to the hierarchical tree to which input values belong, and retrieve some possible centroid candidates as for example sets of taxonomical ancestors. This strategy, combined with the semantic weighted distance proposed in section 3.2, will help to propose more accurate centroids.

First, we formalise our centroid construction method for univariate data.

Let us take $V=\{<v_1,\omega_1>,\ldots,<v_n,\omega_n>\}$ as an input dataset with a single categorical attribute. Let us take an ontology O containing and semantically modelling all v_i in V. The first step of our method consists of mapping the values in V to concepts in O, so that semantically related concepts can be extracted from O following the semantic relationships. We assume that taxonomical subsumers of a term, including itself, are valid representatives of the term. The set of candidates is given in the *minimum subsumer hierarchy*, $H_O(V)$ that goes from the concepts corresponding to the values in V to the Least Common Subsummer of all these values. The *Least Common Subsumer* (LCS) of a set of categorical values V in an ontology O ($LCS_O(V)$) is the deepest taxonomical ancestor that the terms in V have in common for the ontology O. We omitted taxonomical ancestors of the LCS because those will always be more general, that is more semantically distant than the LCS and hence, worse centroid candidates.

Definition 3. The *set of taxonomical subsumers*, $S_{LCSO}(v_i)$ between a certain v_i in V and $LCS_O(V)$ is defined as the set of concepts found in the ontology O that connect via taxonomic relationships v_i and $LCS_O(V)$, including themselves. On ontologies with multiple taxonomical inheritance, several paths can be found between v_i and $LCS_O(V)$; all of them are included in $S_{LCSO}(v_i)$.

Definition 4. The *minimum subsumer hierarchy* ($H_O(V)$) extracted from the ontology O corresponding to all the values in V is defined as the union of all the concepts in $S_{LCS_O}(v_i)$ for all v_i.

$$H_O(V) = \bigcup_{i=1}^{n} \{ S_{LCS_O}(v_i) \} \tag{6}$$

where n is the cardinality of V.

Example 3. As an illustrative example, let us consider a univariate dataset where the attribute refers to the preferred sport: V_1 = {<*boxing*, 1>, <*soccer*, 2>, <*rugby*, 2>, <*contact_sport*, 1>, <*swimming*, 1>, <*surfing*, 3>}. By mapping these values to concepts found in the background ontology O

(WordNet) and applying Definition 4, we are able to extract the minimum hierarchy H_{WN}, shown in the Fig. 3.

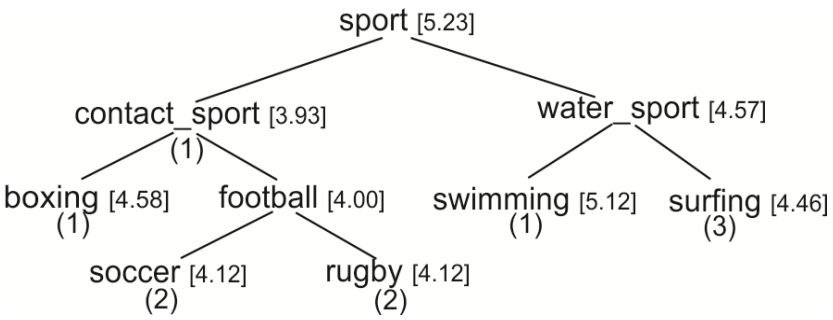

Fig. 3. The minimum subsumer hierarchy H_{WN} for the set V_1, extracted from WordNet. Numbers in parenthesis represent the number of repetitions of each value in the dataset. Numbers in brackets represent the accumulated distance of each centroid candidate.

The LCS of the entire set V_1 is $LCS_{WN}(V_1)$=*sport*. For example by applying Definition 3, the ancestors of *soccer* are S_{LCSWN}(*soccer*)={*soccer*, *football*, *contact_sport*, *sport*}.

Definition 5. All the concepts c in H_O are the centroid candidates for V.

Following Example 3, the centroid candidates of V_1 are those in H_{WN}: {*boxing*, *soccer, rugby*, *swimming*, *surfing*, *contact_sport*, *football*, *water_sport*, *sport*}.

From the set of centroid candidates, and following the centroid definition (Eq. 5), the term c in H_O that minimises the semantic distance to all the v_i in V will be selected as the final centroid. In order to consider both semantics and distribution of data, the distance measure presented in Eq. 4 (section 3.2) is applied to each centroid candidate.

Definition 6. The centroid of a set of textual values V in the ontology O is defined as the concept c_j belonging to $H_O(V)$ that minimises the weighted semantic distance wsd_O with respect to all the values of in V.

$$centroid_O(V) = \{ argmin(\sum_{i=1}^{n} wsd_O(< c_j, 1 >, < v_i, \omega_i >) \}, \forall c_j \in H_O(V), \forall v_i \; in \; V \qquad (7)$$

If more than one candidate minimises the distance, all of them would be equally representative, and any of them can be selected as the final centroid.

To illustrate the procedure, let us take Example 3. Taking the values in V_1, we obtain the weighted semantic distances for each centroid candidate in H_{WN}. The candidate that minimises the distance against all input values is *contact_sport* with wsd_{WN} (*contact_sport*, V_1) = 3.93. So, by Definition 6, $centroid_O(V_1)$ = *contact_sport*.

The fact that all the centroid candidates are evaluated to minimise the distance to all values in V produces optimal results with respect to the background ontology. It is important to note that, as shown in Example 3, neither the LCS of V (*sport*) nor the most frequently appearing value in V (*surfing* with 3 appearances) necessarily minimise that distance. In fact, the use of the LCS as centroid for non-uniformly distributed data values, both with respect to their frequency of appearances, but also to their distribution through the hierarchy H_{WN}, typically results in a high semantic distance wsd_{WN}(*sport*, V_1) = 5.23. In Example 3, the optimal centroid (*contact_sport*) balances the frequency of appearance of the terms and the unbalanced distribution of those terms within the hierarchy, i.e. the dataset has more *contact_sport* branch than *water_sports*.

The method can be generalised for multivariate data considering independent attributes and applying the proposed method individually for each attribute. Note that, in this manner, the centroid construction is optimised at an attribute level, but not at a global level. In this last case, a global centroid selected from the evaluation of all the possible value tuple combinations will be necessary to provide optimal results. However, this will hamper the scalability of our method, requiring the evaluation of an exponentially-large number of value combinations, generated according to the input values and the taxonomical ancestors modelled in the background ontology.

Definition 7. The centroid of a set of multivariate data MV in the ontology O is defined as:

$$centroid_O(MV) = \{centroid_O(A_1), centroid_O(A_2), \ldots, centroid_O(A_m)\} \quad (8)$$

where A_j the set the set of distinct values for the jth attribute in MV and m is the number of attributes in MV.

4 . Evaluation

In this section, the evaluation of the proposed SA-MDAV algorithm is detailed. The datasets and measures used in the evaluation are introduced in sections 4.1 and 4.2, respectively. Afterwards, we first present a study of the contribution of each of the modifications proposed with regards to the minimisation of the information loss, section 4.3. In section 4.4, the performance of the SA-MDAV method is evaluated and compared against those of related works under different perspectives, information loss, disclosure risk and runtime.

4.1. Evaluation data

In all the tests the knowledge base used is WordNet 2.1 (Fellbaum, 1998), both to compute semantic similarity as well as to assist the centroid construction process. WordNet offers several advantages in comparison to ad-hoc knowledge structures used in other works that is Value Generalisation Hierarchies (He and Naughton, 2009; Terrovitis et al. , 2008), which are specially constructed and tailored for the input data. First, WordNet is a standard knowledge source created from the consensus of the contributing community. Hence, it represents the knowledge in a more objective way in comparison to ad-hoc VGHs that could be biased toward author's point of view and over-specified to the concrete problem (Martínez et al. , 2011). Relying on an independent knowledge base abstracts the evaluation of the results from the influence of the design of problem-specific knowledge structures. Moreover, the related works proposing semantically-grounded MDAV applications (Abril et al. , 2010), to which our method will be compared, also use WordNet in their evaluations. Due to these reasons, the use of WordNet enables a more objective evaluation and comparison of the results. Finally, as argued (Martínez et al. , 2010), a more accurate anonymisation with lower information loss is achieved with the use of a large knowledge base like WordNet, because this kind of ontologies have

a finer-grained taxonomical structure than the usually small VGHs and, hence, more accurate generalisation values can be found. Obviously, due to the fact that WordNet is a general purpose knowledge base, in some cases, value-concept mappings could not be directly found for example when input dataset includes specially tailored terms or ad-hoc abstractions. In these cases, we include the specific mapping used for evaluation tests.

As evaluation dataset, we have taken two databases with different characteristics, which permits us to evaluate our methods under well-differenced scenarios.

Dataset 1 consists on answers to polls made by the "Observatori de la Fundació d'Estudis Turístics Costa Daurada" at the Catalan National Park "Delta de l'Ebre". It comprises 975 records, each one corresponding to an individual, storing the answers given to the questions of the poll, where each question is treated as a different attribute. For evaluation purposes, we consider two non-numeric categorical attributes as *quasi-identifiers* that should be anonymised. These correspond to answers to the questions "Reason for visiting the park" and "Main activity done during the visit to the park". The answers to each of these questions take values from a set of 22 modalities, see Fig. 8. These correspond to common concepts related to hobbies, interests and activities which can be directly found in WordNet. Concretely, WordNet 2.1 has been used to support the anonymisation process. Fig. 4 shows the distribution of the data according to the number of repetitions, frequency of appearance of each different pair of values, that is the answers to the two questions considered. The variety of possible answers generates a large set of combinations, that is tuples, for the two attributes considered. Moreover, the frequency of appearance of each tuple of answers is not uniform, as it can be seen in Fig. 4. In fact its shape suggests a *long tail* distribution in which some of the tuples tend to repeat many times, whereas others appear very rarely. In this case, we have a total of 211 different tuples, representing a 21.6% of the total, where the tuple with the highest amount of repetitions appears 115 times, <"*nature*", "*relaxation*">, which corresponds to the 11.8% of the answers. Otherwise, many tuples, 118, a 12.1% of the total answers, related to rarer motivations, such as "*business*" or "*loyalty*", are unique, showing a relatively high amount of outliers. This dataset is

especially interesting from the privacy-preserving perspective due to the presence of a large amount of easily identifiable records; around a 31% of them are 5- of lower *k*-distinguishable.

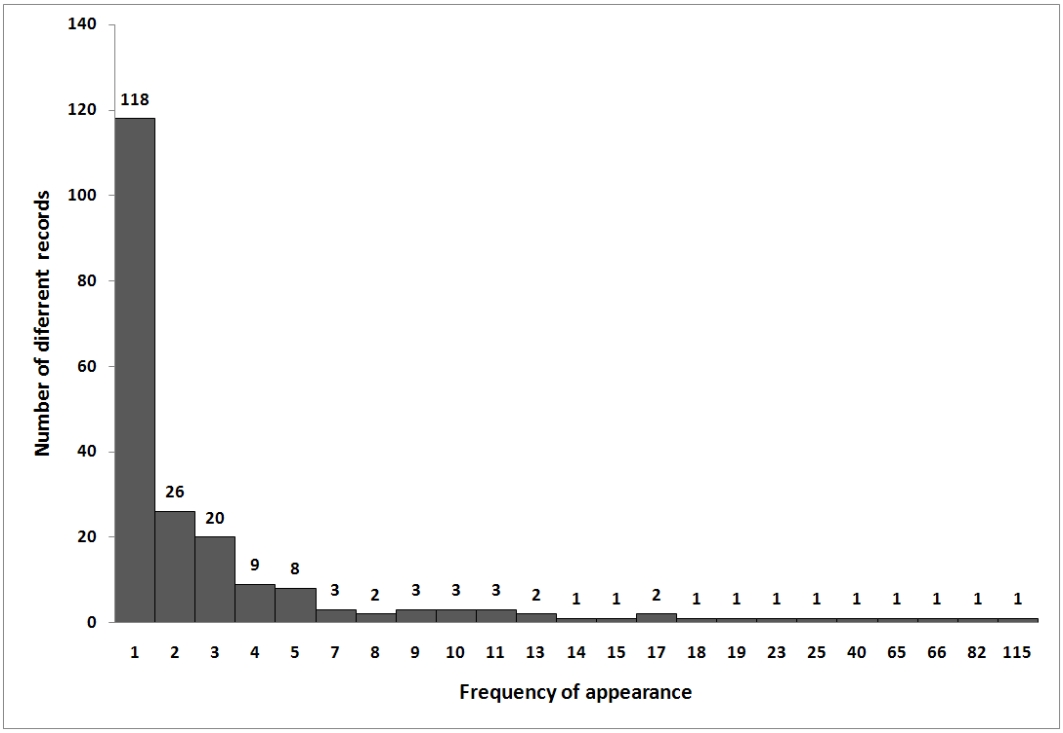

Fig. 4. The Dataset 1 frequency distribution of distinct value tuples

Dataset 2 is the well-known *Adult Census* (Hettich and Bay, 1999), which is publicly available in the UCI repository[ii] and has been often used in the past for evaluating privacy-preserving methods (Domingo-Ferrer et al. , 2006; Fung et al. , 2005; Iyengar, 2002; Lin and Wei, 2008; Lin et al. , 2010). In the same manner as above, we have considered two categorical attributes as *quasi-identifiers* corresponding to "occupation" (14 distinct modalities) and "native-country" (41 modalities). Because some of these modalities cannot be directly found in WordNet, due to its ad-hoc linguistic label, they have been mapped to WordNet concepts as shown in Table 1. For evaluation purposes, we have used the training set consisting on 30,162 records, after removing rows with missing values. Fig. 5 shows the data distribution. In it, 388 different responses exist, which represent only a 1.28% of the total, in comparison with the 21.6% of Dataset 1, with barely 83 of them being unique. Even though the

[ii] http://archive.ics.uci.edu/ml/datasets/Adult

dataset also follows a long tail distribution, here the data distribution is considerably less heterogeneous than for Dataset 1. In fact, the tuple with the highest amount of repetitions appear 3,739 times and there are 9 tuples with more than 1,000 repetitions, representing the 83.2% of the size of the dataset. On the contrary, responses with 5 repetitions or less that is, those that should be clearly protected represent a 1.9% of the total, compared to the 31% for Dataset 1. This difference interestingly shows the behaviour of our method in two well-distinguished scenarios, regarding the privacy/utility evaluation measures, but also with respect to scalability.

Table 1. Value mapping between *Adult Census* dataset and WordNet

Original values	**WordNet values**
Tech-support	Technician
Craft-repair	Craftsman
Other-service	Worker
Exec-managerial	Executive
Prof-specialty	Specialist
Handlers-cleaners	Cleaner
Machine-op-inspct	Operator
Adm-clerical	Clerk
Farming-fishing	Skilled_worker
Transport-moving	Carrier
Priv-house-serv	Housekeeper
Protective-serv	Guard
Armed-Forces	Soldier
Outlying-US(Guam-USVI-etc)	American_State
Holand-Netherlands	Netherlands
Hong	Hong-Kong

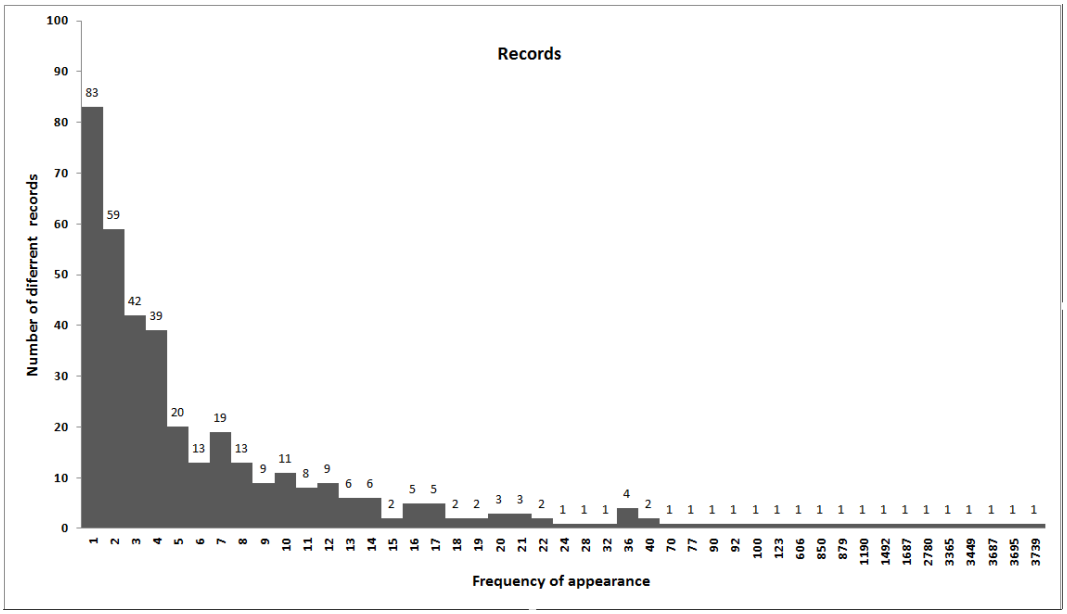

Fig. 5. The Dataset 2 frequency distribution of distinct value tuples

Due to the differences in data distribution, the *k*-anonymity level tested for each one is also different. For Dataset 1, it ranges between 2 and 15, which are common *k*-anonymity values for heterogeneous datasets (Abril et al. , 2010; Domingo-Ferrer et al. , 2006; He and Naughton, 2009; Loukides and Shao, 2007) and masks up to a 50% of the total records. For Dataset 2, as done in related works also employing this set for evaluation purposes (Fung et al. , 2005; Lin and Wei, 2008), we have increased the *k* level to obtain more representative results: in our case up to 1,800, so that up to a 31% of the dataset is protected.

4.2. Evaluation measures

As stated in the introduction, any disclosure control method should balance two opposite dimensions: data utility, as a function of information loss, and disclosure risk. Several standard measures have been proposed in the literature to evaluate these perspectives. In this section, we detail the measures used to evaluate our method.

As stated in section 2, the *information loss* resulting from the microaggregation process is the direct consequence of replacing cluster values by the cluster centroid. Hence, within-cluster homogeneity is the critical dimension that a microaggregation method should optimise. This

dimension is measured as the *Sum of Square Errors* (SSE), which is the optimisation/evaluation criterion commonly used in privacy-preserving microaggregation methods (Abril et al. , 2010; Domingo-Ferrer et al. , 2006; Domingo-Ferrer and Mateo-Sanz, 2002; Lin et al. , 2010; Torra and Miyamoto, 2004). It is defined as the sum square of distances between each element of each cluster and their corresponding centroid (9). Hence, the lower the SSE is, the higher the within-group homogeneity and the lower the information loss resulting from the replacement of values by cluster centroids.

$$SSE = \sum_{i=1}^{g} \sum_{j=1}^{n_i} (dis_{w\&p}(x_{ij}, \overline{x}_i))^2 \quad , \tag{9}$$

where g is the number of clusters, n_i is the number of elements in the ith cluster, x_{ij} is the jth element of the cluster i and $\overline{x}_i$ denotes the centroid of the cluster ith cluster. As a distance measure, we employ $dis_{w\&p}$ defined in Eq 2, as it is done in related works.

To enable the normalisation of SSE values according to the distribution of each particular dataset, the *Total Sum of Squares* (SST) evaluates the sum square of distances between each individual element and the centroid $\overline{x}$ of the whole dataset:

$$SST = \sum_{i=1}^{g} \sum_{j=1}^{n_i} (dis_{w\&p}(x_{ij}, \overline{x}))^2 \tag{10}$$

Hence, the information loss (L) of a microaggregated dataset is measured as the ratio, in percentage, between SSE and SST (Abril et al. , 2010; Domingo-Ferrer, 2008; Domingo-Ferrer and Mateo-Sanz, 2002; Torra and Miyamoto, 2004):

$$L = \frac{SSE}{SST} \times 100 \tag{11}$$

The opposite dimension to *information loss* in a privacy-preserving method is the *Disclosure Risk* (DR), that is the chance of an intruder to disclosure the identity of an individual. This dimension is commonly evaluated by means of *Record Linkage* (RL) as defined in (Torra and Domingo-Ferrer, 2003). RL is the task of finding matches in the original data from the anonymised results. Hence, the

disclosure risk of a privacy-preserving method can be measured as the difficulty in finding correct linkages:

$$RL = \frac{\sum_{i=1}^{m} P_{rl}(r_i^A)}{m} \cdot 100 \tag{12}$$

The record linkage probability of an anonymised record $P_{rl}(r_i^A)$ is calculated as follows:

$$P_{rl}(r_i^A) = \begin{cases} 0 & if\ r_i \notin G \\ \frac{1}{|G|} & if\ r_i \in G \end{cases} \tag{13}$$

where r_i is the original record, r_i^A is its anonymised version and G is the set of original records that have been linked to r_i^A. When dealing with categorical attributes, record matching is performed by terminologically matching textual values of each attribute. Therefore, each r_i^A is compared to all records of the original dataset, thus obtaining the G set of matched records. If r_i is in G, then the probability of record linkage is computed as the probability of finding r_i in G. Otherwise, the record linkage probability is 0.

To evaluate the *balance score* between the information loss and the disclosure risk, which is the final goal that any SDC method pursuits, it is proposed to compute the weighted average of these two complementary measures (Eq 14). The parameter α is used to adjust the interest of the user on data utility versus privacy. A value of α=0.5, which results in a standard arithmetic mean, is commonly considered in the related works (Domingo-Ferrer, 2008; Domingo-Ferrer and Torra, 2001; Torra, 2004; Yancey et al. , 2002). The overall score should be minimised since, the lower the score is, the higher the quality of the method because we achieved both low information loss and low record linkage, de-identification.

$$score = \alpha L + (1-\alpha)RL \tag{14}$$

4.3. Analysis of SA-MDAV

As stated in Section 3, the main aim behind the modifications introduced in the MDAV algorithm when dealing with categorical data is the minimisation of the information loss resulting from the

microaggregation process. In this section, we evaluate the contribution of the algorithmic modifications introduced to MDAV from that perspective.

To do so, we configured three settings, each one incorporating or not some of the modifications proposed in Section 3. To focus only on the algorithmic differences between MDAV and SA-MDAV, in all the three settings, the distance between records was calculated as proposed in section 3.2 (Eq. 4) and the selection of centroids was done as proposed in Section 3.3 (Eq. 8). The different versions are:

- *S-MDAV (Semantic MDAV)*: The dataset is microaggregated using the basic MDAV (Algorithm 1, in Section 2). Thus, clusters are bounded to a fixed size of k, except the last one and each cluster is constructed from the first selected record (r), rather than from the centroid computed at each aggregation step, as proposed in Section 3.1. The difference with the classical MDAV relies on the use of the semantic similarity to compare tuples, instead of the equality predicate and on the selection of centroids.
- *SA-MDAV-static (Semantic Adaptive MDAV with static centroids)*: The dataset is microaggregated using the adaptive method proposed in Section 3.1. However, as in the original method, each cluster is created from the first selected record (r, static centroid), instead of computing, at each aggregation stage, the cluster centroid. Comparing to S-MDAV version, one can quantify the influence in information loss of the adaptation of cluster size according to the data distribution.
- *SA-MDAV (Semantic Adaptive MDAV)*: The dataset is microaggregated using our complete proposal. Comparing the previous version, it is quantified the contribution of cluster centroid recalculation at each microaggregation step.

The three versions have been applied to Dataset 1, for k-values between 2 and 15, and Dataset 2, for k-values between 2 and 1,800. To compare the quality of the results, the information loss (L) measure presented in Section 4.2 was computed; results are shown in Fig. 6.

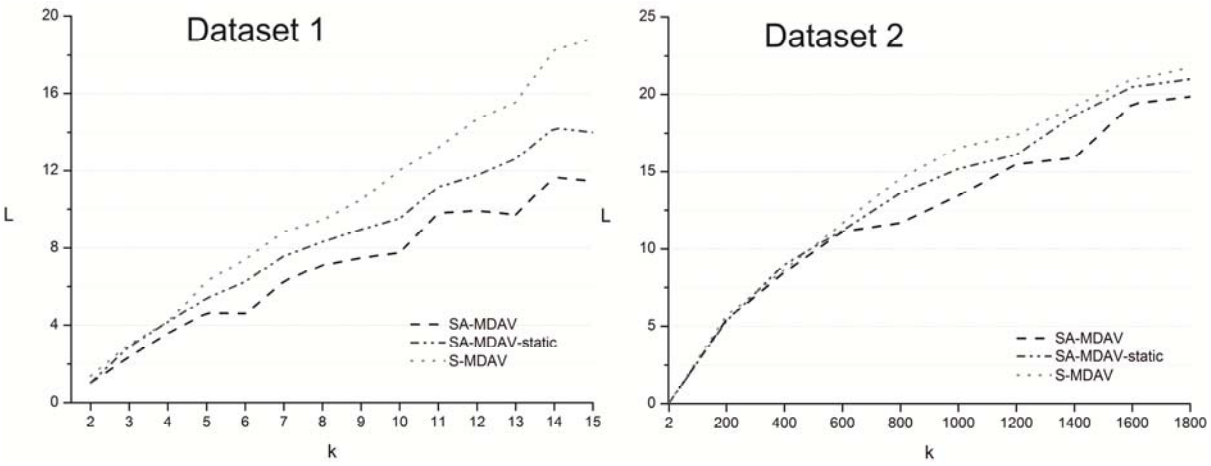

Fig. 6. A comparison of the three algorithm versions according to Information Loss (L)

Results shown in Fig. 6 are coherent to what it is expected from the design of the contributions proposed in this paper. In general, each modification introduced to the MDAV algorithm resulted in a progressive decrease of the *L* value, with a tendency to maximise the difference as the *k*-value grows-up. First of all, the recalculation of cluster centroid at each aggregation step (SA-MDAV-static vs. SA-MDAV) leads to subtle improvement of the within cluster homogeneity, and also of the *L* measure, because clusters are optimally constructed with regards to the minimisation of intra-cluster distances. Differences are more evident for Dataset 1 because its higher heterogeneity makes the construction of appropriate clusters more prone to incorporate outliers. Dataset 2, on the contrary, provides large and cohesive sets of equal input records, so that the cluster construction does not depend on the dynamically computed centroid, especially for low values of *k*.

Secondly, when cluster size is adapted to data distribution (SA-MDAV vs. S-MDAV), the minimisation of information loss (*L*, Fig. 6) is more noticeable. This shows that more cohesive clusters can be obtained if their size can be adapted to the data distribution. Differences become more noticeable for high values of *k* (above 5 for Dataset 1 and above 800 for Dataset 2) because, as the value of *k* grows up, the cardinality of sets with identical value tuples become hardly *k*-divisible. Hence, fixed-size aggregation is forced to join residual records of several tuples together, resulting in highly heterogeneous clusters, as illustrated in Fig. 6. On the contrary, our adaptive method only joins records with different values when their individual cardinalities are lower than *k* to fulfil the *k*-anonymity property.

4.4. Evaluation and comparison with related works

In this section, we compare the SA-MDAV method against those proposed by two representative related works using the MDAV algorithm to deal with categorical data. On one hand, the proposal by (Domingo-Ferrer and Torra, 2005), as detailed in Section 2, propose a fixed-size microaggregation using the equality predicate, 0 for identical tuples, 1 otherwise. Centroids are computed as the most frequent value, mode. Hence, this approach does not consider the semantics of concept in any way. On the other hand, the method by (Abril et al. , 2010) has been also tested. In this case, the distance between tuples is computed using the Wu & Palmer similarity measures (Eq. 1) and WordNet as background ontology. Centroids are selected as the concept in WordNet that subsumes all values (LCS). In both cases, input data is analysed record by record, instead of by distinct value tuples, as proposed in our method.

Masked datasets obtained by the three methods (SA-MDAV and the two related works) have been evaluated and compared by means of the information loss measure (*L*, as shown in Fig. 7), disclosure risk (*RL*, as shown in Fig. 9), quality score (Fig. 10 and Fig. 11) and also runtime (Fig. 12).

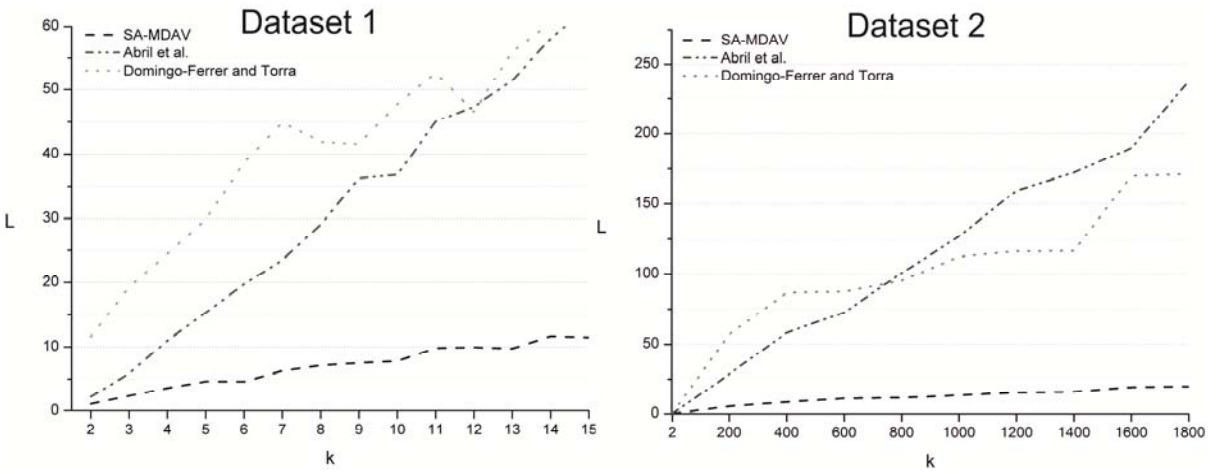

Fig. 7. A comparison of Information Loss (L) values for the evaluated methods.

With regards to the information loss metric (Fig. 7), our approach was able to improve the related works on the two datasets. First, we observe that when no semantics are considered during the microaggregation of categorical data (Domingo-Ferrer and Torra) the information loss is high, even for low *k*-values. In fact, managing and transforming categorical data only by means of their

distributional features, both when comparing records and when computing centroids, hardly preserves their meaning, a dimension that, as discussed in (Martinez et al. , 2011), is closely related to their utility. Regarding the ontology-based method proposed by (Abril et al. , 2010), we observe that, even though it retains more information than the non-semantic method for low *k*-values, it produces lower quality results when a higher degree of anonymity is required, especially for the more heterogeneous Dataset 1. This is closely related to the fact that semantic management of data does not consider their distribution. Hence, as discussed in section 2, the centroid selection, consisting on picking up the LCS of all the considered values, easily results in very general abstractions due to the need of generalising outliers. This results in a high information loss that is accentuated when clusters become less homogeneous, which is the case of fixed-sized microaggregation methods for high *k*-values.

In comparison, our method is able to retain a lower information loss for both datasets through all the *k*-values, showing an increasing tendency with a smoother slope, while keeping the information loss at the lowest level. This is the result of carefully considering both the semantics of data and their distribution in all stages of the microaggregation process, distance calculus, centroid selection, cluster construction and data transformation. Resulting clusters are hence more cohesive thanks to the less constrained aggregation, and due to the optimisation of both their semantic and distributional features.

For illustration purposes, in Fig. 8 we show the knowledge structure evaluated from WordNet when computing the first initial centroid, first execution of line 3 in Algorithm 1. In this case, the whole set of values of an attribute of Dataset 1 is considered to compute the centroid. Analysing the appearance frequencies, we observe that the most appearing term, mode is “relaxation” with 249 appearances. The approach by (Domingo-Ferrer and Torra) will select this term as the centroid for this attribute. Moreover, the LCS for the complete set of values is “entity”, which is also the root node in the WordNet’s taxonomy. The approach by (Abril et al. , 2010) would select this term as the centroid. Applying our method described in Section 3.3, the sum of distances with respect to “relaxation” is 322.74, whereas the sum of distances to “entity” is 746.31. This shows that using the LCS as the centroid poorly optimises the minimisation of semantic distances of the elements on the dataset. In fact, the most adequate centroid is “inactivity”, for which the sum of distances is 257.97, the

minimum, representing the centre of the dataset. Coherently, this term is near to the dataset's "relaxation" mode, but also, it is closer than the mode to other terms with special relevancy, such as "nature", for which the sum of distances is 348.72. This shows why, even with datasets with a clear prevalence of certain terms, "relaxation" and "nature" in this dataset, the mode is not necessarily the most adequate centroid.

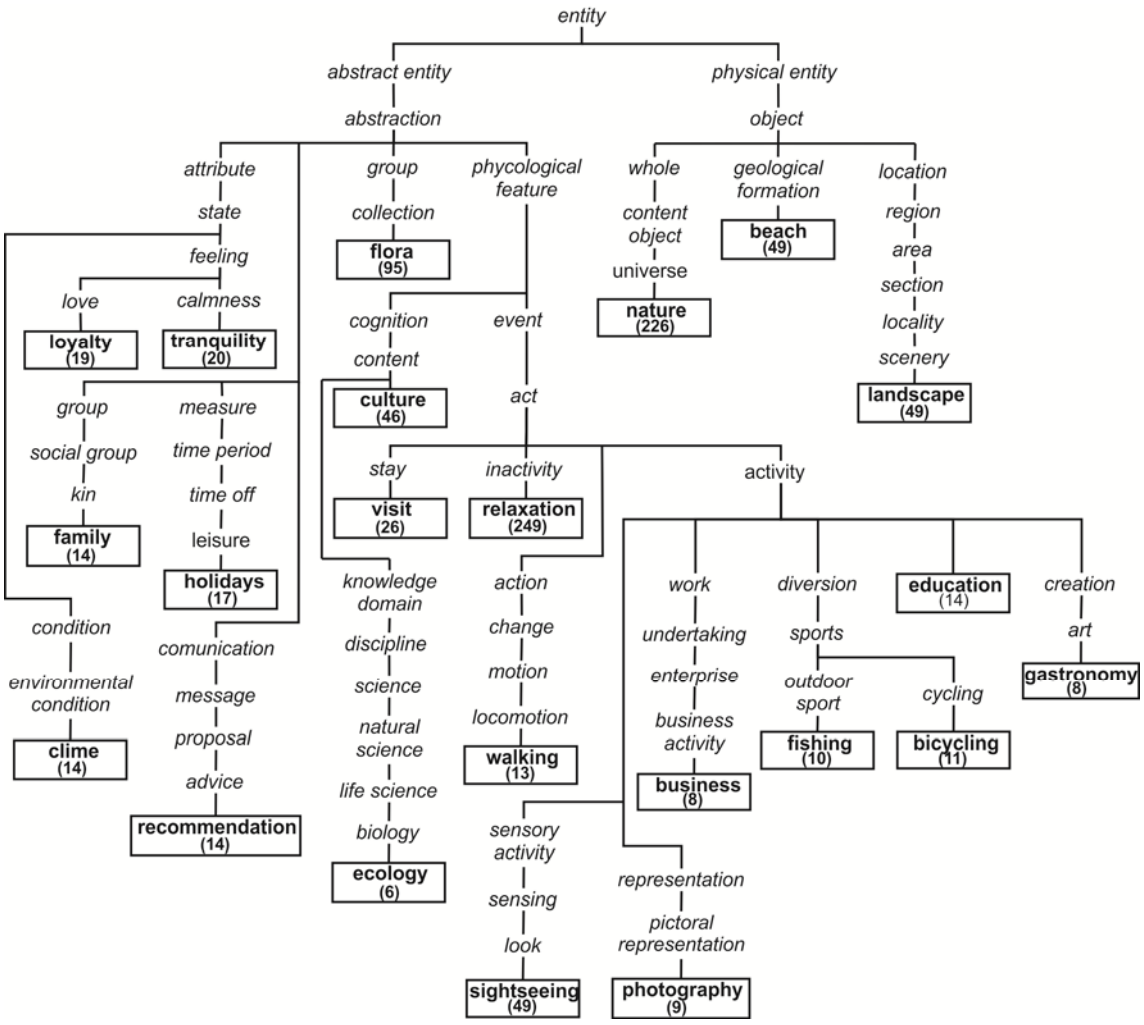

Fig. 8. The knowledge structure extracted from WordNet for input values of an attribute of Dataset 1 (in bold). The numbers in brackets represent the amount of appearances of each value.

Even though data utility is of utmost importance when masking data, the disclosure risk should be also minimised. Analysing the *RL* results on the protected Datasets 1 and 2 for the three methods (Fig. 9) several conclusions can be extracted. First the method with the lowest percentage of record

linkages in both evaluations is the one by Abril et al. Their method replaces aggregated values by their LCS (retrieved from WordNet). As a result, especially for high *k*-values, most values in the original dataset are replaced by abstract generalisations. In this case, because RL quantifies the amount of terminological matchings between the original and masked datasets, the chance of proposing a correct one is very low. Hence, the RL values tend to be very low. The opposite case applies for the approach by Domingo-Ferrer and Torra. When centroids are calculated as the mode of the obtained groups, records are replaced by values already present in the original dataset. This significantly increases the chance of proposing a correct linkage using a terminological matching.

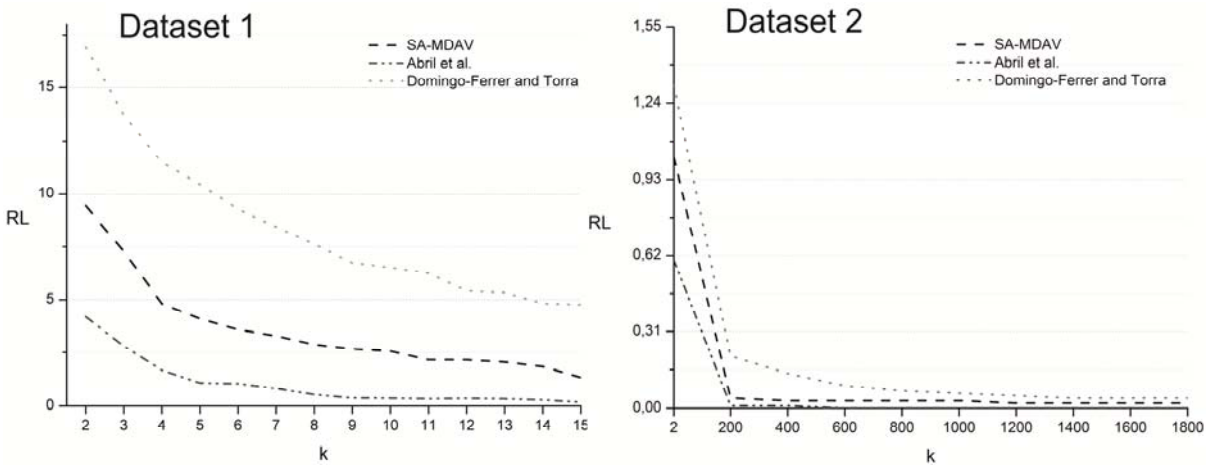

Fig. 9. A comparison of RL percentage for the evaluated methods.

In comparison, our method presents an average behaviour, even though it approximates more to Abril et al. than Domingo-Ferrer and Torra. This is because the centroids are selected according to both distributional and semantic features of data, as detailed in Section 3.3. Centroids aim to minimise the sum of semantic distances between all the aggregated values, considering also their distribution, as a weighting factor. This results in centroids that can be either more new general concepts retrieved from WordNet, in case of homogenously distributed values across the hierarchy of concepts or values already found in the original dataset, if they appear predominantly. The number of record linkages is always between the results obtained by the approaches in which data is almost completely replaced by new values (Abril et al.) and those in which the same values are maintained (Domingo-Ferrer and Torra).

To evaluate and compare the quality of the different methods as a whole, the overall score integrating the information loss metric and the disclosure risk measure is studied (Eq 14). First, as shown in Fig. 10, we consider an equal balance between the data utility and the disclosure risk, an average with α=0.5, as done in the related works (Domingo-Ferrer, 2008; Domingo-Ferrer and Torra, 2001; Torra, 2004; Yancey et al. , 2002).

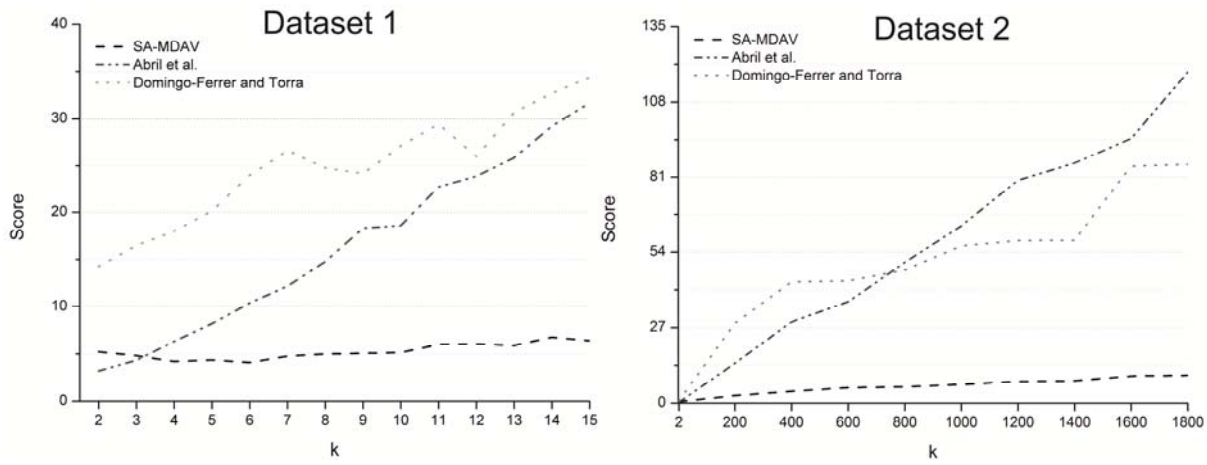

Fig. 10. Score with an equilibrated balance (α=0.5) between information loss and disclosure risk.

The conclusion is that, even though our method resulted in higher RL values than the one by Abril et al., it provides the best balance in both datasets between information loss and disclosure risk, a circumstance that, as discussed in the introduction, is the main aim of a privacy preserving method. It is also relevant to note that the score is maintained almost constant as *k*-values grow, this behaviour is remarkable in both evaluations and more evident in the Dataset 2, stating that the quality of our method scales well as the privacy requirements increase. In contrast, for related works, the *score* grows almost linearly with respect to *k*. The approach by Domingo-Ferrer and Torra resulted in a high score even for low *k*-values due to the high information loss resulting from the non-semantic management of data. The approach by Abril et al., on the contrary, provided quality results for low *k*-values due its low disclosure risk and controlled information loss. As *k*-values grow, however, the score follows the same tendency as the information loss because due to the disclosure risk can be hardly minimised when most of the values have been replaced.

Second, we have studied the behaviour of the overall score when varying the α parameter between 0 to 1. With α=0, the score is based solely on the disclosure risk measure, while with α=1, the score is based only on the information loss. In this analysis, an intermediate level of anonymity (k value) has been fixed in both datasets. As it can be seen in Fig. 11, our method achieves the best results, minimal score for almost all the cases in both datasets. The results of Domingo-Ferrer and Torra are significantly higher, while the method of Abril et al. is only able to surpass our results when a highest weight is given to disclosure risk and data utility is not taken into account. This effect has been previously observed in Fig. 9.

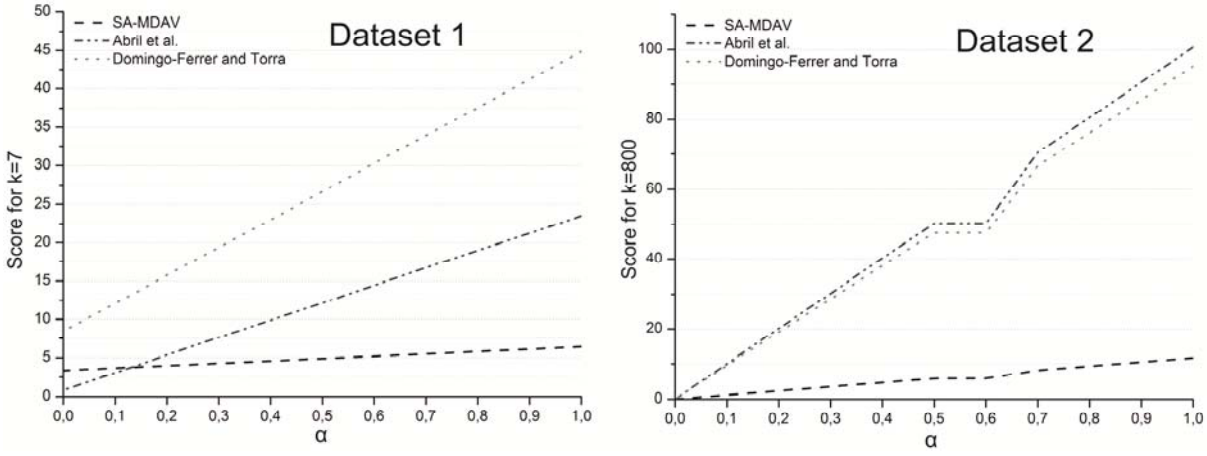

Fig. 11 Score values when varying the relative weight between information loss and disclosure risk.

Finally, an analysis of the execution time has been done. To understand the runtime results shown in Fig. 11, obtained with a 2.4 GHz Intel Core processor with 4 GB RAM, let us first analyse the cost of the different microaggregation methods. As introduced in Section 2, in the basic MDAV algorithm, for each generated cluster, it is necessary to calculate the centroid, the farthest record and the k-1 closest records, which implies a computational cost of $O(n^2)$, where n is the number of *records* in the dataset. Our SA-MDAV proposal adds a computational overhead in the optimisation of the centroid calculation, which results in $O(c \cdot k)$ for each cluster, where c is the number of centroid candidates, see Section 3.3. Furthermore, as stated in section 3, our method manages input data according to the number of *distinct tuple values* (p) instead of *total records* (n). Hence, the computation cost of our

method would be $O(p^2)\cdot O(c\cdot k)$. Since k should be usually kept small, the computational cost depends on the number of distinct records in the dataset, being $p<n$ as seen in the data distribution analysis shown in Fig. 4 and Fig. 5, and on the number of centroid candidates, which depends on the ontology.

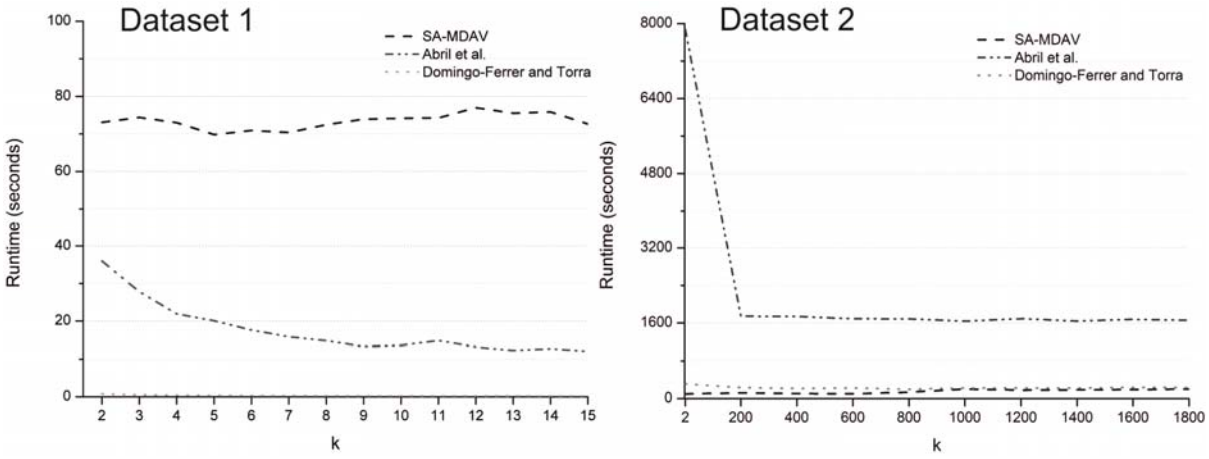

Fig. 12. A runtime comparison of the evaluated methods.

In this case, the use of these two distinct datasets permits to illustrate the runtime behaviour in two different scenarios. For Dataset 1, our method requires the highest runtime to mask data. On the contrary, for Dataset 2, our proposal obtains the best performance. This is caused by the different magnitudes and data distributions. For the first dataset n=975 and p=211. In this case, since the set distinguishable tuples represent a significant 21.6% of the total records, the difference between n and p is too low with respect to the overhead of querying the ontology for distance calculus and concept retrieval, recalculating cluster centroids at each aggregation step and optimising the centroid selection. Even though, such tasks only depend on p, our method produces an almost constant runtime regardless of the k-value among 70 to 80 seconds. In comparison, thanks to the low number of records (n) the approach by Domingo-Ferrer and Torra, which is based on simple operators (mode and equality predicate), and without exploiting background knowledge, has almost a negligible runtime. The approach by Abril et al. requires an average of 20 seconds due to the queries performed to the background ontology.

For Dataset 2, n=30,162 records, whereas p=388 indistinguishable values. In this case, since p represents only a 1.28% of n, the gain of $O(p^2)$ with respect to $O(n^2)$ results in the best performance

for our method, among 131 and 205 seconds. In comparison, the approach by Abril et al. that requires 7,873 seconds for k=2 and 1,839 seconds for k=128, whereas the simplest method by Domingo-Ferrer and Torra ranges between 318 seconds for k=2 to 242 seconds for k=128.

In conclusion, approaches based on fixed-sized clusters are severely penalised both by the fact that records are individually evaluated and by the restriction of grouping equal records in a k-sized cluster, which requires repeating the clustering process for the rest of equal records. The approach by Abril et al. is even more hampered, due to the need of querying the ontology at each aggregation step. On the contrary, our method is able to group all equal records in a single microaggregation step. This ensures its scalability with large datasets, because it neither depends on the k-anonymity level and the dataset size (n), only on its heterogeneity (p), which is commonly low when dealing with categorical data.

5. Conclusions

Microaggregation algorithms, and MDAV in particular, are some of the most commonly applied privacy-preserving methods for Statistical Disclosure Control in structured databases (Abril et al. , 2010; Domingo-Ferrer, 2008; Domingo-Ferrer and Mateo-Sanz, 2002; Domingo-Ferrer and Torra, 2005; Lin et al. , 2010; Torra, 2004). Even though most of these methods were designed for numerical data, in recent years, the interest in protecting categorical data has grown-up. As shown in Section 2, some authors have adapted the MDAV algorithm to this kind of data, proposing different ways of comparing and aggregating this kind of data. On the contrary to numbers, categorical data presents some special characteristics. On one hand, they take values from a discrete, finite and typically reduced set of modalities, words or noun phrases. Moreover, datasets are rarely uniform, and commonly follow a long tail distribution. On the other hand, categorical values (words) refer to concepts with an underlying meaning and, hence, a semantic analysis is needed to properly interpret them. The work presented in this paper aimed to carefully consider these characteristics during the microaggregation process. As a result, several modifications have been proposed to the MDAV algorithm.

The proposal relies on a hierarchical knowledge structure that represents the taxonomical relations of the values that appear in the dataset. Although other knowledge bases can be used, the exploitation of large and detailed ontologies enables an accurate interpretation of data semantics. Consequently, the proposed semantic-based techniques are able to retain the utility of masked data, while the adaptation of the microaggregation process to the distributional features of data is intended to produce more cohesive clusters and hence, to minimise the information loss.

The evaluation, performed over two different datasets with textual attributes, sustained the theoretical hypotheses. We analysed how each modification aided to minimise the information loss of the protected dataset. We have also proved that our method, 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. Finally, we illustrated the scalability of our method with large datasets, which basically depends on the dataset heterogeneity, typically low in categorical data rather on its size, as in related works.

As future work, we plan to further evaluate our method in other domains and ontologies. The medical field is especially interesting due to the importance and sensitivity of clinical data and due to the fact that standard domain ontologies have been developed. We will also study the possibility of combining several ontologies as background knowledge in order to complement knowledge modelled by each of them. In this manner, the evaluation could be extended to datasets with heterogeneous attributes. Finally, some of the modifications proposed for the MDAV algorithm could be applied to other privacy-preserving control methods, such as rank swapping (Nin et al. , 2008b) or *k*-ward (Domingo-Ferrer and Mateo-Sanz, 2002).

Acknowledgements

We would like to thank the Observatori de la Fundació d'Estudis Turístics Costa Daurada and the Delta de l'Ebre National Park for providing the data. This work was partly funded by the Spanish Government through the projects CONSOLIDER INGENIO 2010 CSD2007-0004 "ARES" and eAEGIS TSI2007-65406-C03-02,

and by the Government of Catalonia under grant 2009 SGR 1135 and 2009 SGR-01523. Sergio Martínez Lluís is supported by a research grant of the Ministerio de Educación y Ciencia.

References

Abril D, Navarro-Arribas G, Torra V. Towards semantic microaggregation of categorical data for confidential documents. Proceedings of the 7th international conference on Modeling decisions for artificial intelligence. Perpignan, France: Springer-Verlag; 2010. p. 266-76.

Batet M, Sanchez D, Valls A. An ontology-based measure to compute semantic similarity in biomedicine. J Biomed Inform. 2011;44:118-25.

Byun J-W, Kamra A, Bertino E, Li N. Efficient k-anonymization using clustering techniques. Proceedings of the 12th international conference on Database systems for advanced applications. Bangkok, Thailand: Springer-Verlag; 2007. p. 188-200.

Chakaravarthy V, Gupta H, Roy P, Mohania M. Efficient techniques for document sanitization. CIKM '08: Proceeding of the 17th ACM conference on Information and knowledge management. Napa Valley, California, USA: ACM; 2008. p. 843-52.

Chiu C-C, Tsai C-Y. A k-Anonymity Clustering Method for Effective Data Privacy Preservation. Proceedings of the 3rd international conference on Advanced Data Mining and Applications. Harbin, China: Springer-Verlag; 2007. p. 89-99.

Domingo-Ferrer J. A Survey of Inference Control Methods for Privacy-Preserving Data Mining. In: Aggarwal CC, Yu PS, editors. Privacy-Preserving Data Mining: Springer US; 2008. p. 53-80.

Domingo-Ferrer J, Martínez-Ballesté A, Mateo-Sanz J, Sebé F. Efficient multivariate data-oriented microaggregation. The VLDB Journal. 2006;15:355-69.

Domingo-Ferrer J, Mateo-Sanz JM. Practical Data-Oriented Microaggregation for Statistical Disclosure Control. IEEE Trans on Knowl and Data Eng. 2002;14:189-201.

Domingo-Ferrer J, Seb\ F, \#233, Solanas A. A polynomial-time approximation to optimal multivariate microaggregation. Comput Math Appl. 2008;55:714-32.

Domingo-Ferrer J, Torra V. A quantitative comparison of disclosure control methods for microdata. In: P. Doyle JL, J. Theeuwes and L. Zayatz, editor. Confidentiality, Disclosure and Data Access. Amsterdam: North-Holland2001. p. 111–33.

Domingo-Ferrer J, Torra V. Ordinal, Continuous and Heterogeneous k-Anonymity Through Microaggregation. Data Min Knowl Discov. 2005;11:195-212.

Erola A, Castella-Roca J, Navarro-Arribas G, Torra V. Semantic microaggregation for the anonymization of query logs. Proceedings of the 2010 international conference on Privacy in statistical databases. Corfu, Greece: Springer-Verlag; 2010. p. 127-37.

Fellbaum C. WordNet: An Electronic Lexical Database (Language, Speech, and Communication): The MIT Press; 1998.

Fung BCM, Wang K, Yu PS. Top-Down Specialization for Information and Privacy Preservation. Proceedings of the 21st International Conference on Data Engineering: IEEE Computer Society; 2005. p. 205-16.

He Y, Naughton J. Anonymization of Set-Valued Data via Top-Down, Local Generalization. VLDB '09: the Thirtieth international conference on Very large data bases. Lyon, France: VLDB Endowment; 2009.

He Z, Xu X, Deng S. k-ANMI: A mutual information based clustering algorithm for categorical data. Inf Fusion. 2008;9:223-33.

Herranz J, Matwin S, Nin J, Torra V. Classifying data from protected statistical datasets. Computers & Security. 2010;29:875-90.

Hettich S, Bay SD. The UCI KDD Archive. 1999. http://kdd.ics.uci.edu.

Huang KL, Kanhere SS, Hu W. Preserving privacy in participatory sensing systems. Computer Communications. 2010;33:1266-80.

Hundepool A, Wetering AVd, Ramaswamy R, Franconi L, Capobianchi A, DeWolf PP, et al. μ-ARGUS version 3.2 Software and User's Manual. Statistics Netherlands, Voorburg NL. hppt://neon.vb.cbs.nl/casc://neon.vb.cbs.nl/casc. 2003.

Iyengar VS. Transforming data to satisfy privacy constraints. KDD: ACM; 2002. p. 279-88.

Jin X, Zhang N, Das G. ASAP: Eliminating algorithm-based disclosure in privacy-preserving data publishing. Information Systems. 2011;36:859-80.

Laszlo M, Mukherjee S. Minimum Spanning Tree Partitioning Algorithm for Microaggregation. IEEE Trans on Knowl and Data Eng. 2005;17:902-11.

Lin J-L, Wei M-C. An efficient clustering method for k-anonymization. Proceedings of the 2008 international workshop on Privacy and anonymity in information society. Nantes, France: ACM; 2008. p. 46-50.

Lin J-L, Wen T-H, Hsieh J-C, Chang P-C. Density-based microaggregation for statistical disclosure control. Expert Syst Appl. 2010;37:3256-63.

Loukides G, Shao J. Capturing data usefulness and privacy protection in K-anonymisation. Proceedings of the 2007 ACM symposium on Applied computing. Seoul, Korea: ACM; 2007. p. 370-4.

Macqueen JB. Some Methods for classification and analysis of multivariate observations. Procedings of the Fifth Berkeley Symposium on Math, Statistics, and Probability: University of California Press; 1967. p. 281-97.

Martinez S, Sanchez D, Valls A, Batet M. Privacy protection of textual attributes through a semantic-based masking method. Inf Fusion. 2011;DOI: 10.1016/j.inffus.2011.03.004.

Martínez S, Sánchez D, Valls A, Batet M. The Role of Ontologies in the Anonymization of Textual Variables. Proceeding of the 2010 conference on Artificial Intelligence Research and Development: Proceedings of the 13th International Conference of the Catalan Association for Artificial Intelligence: IOS Press; 2010. p. 153-62.

Martínez S, Sánchez D, Valls A, Batet M. Privacy protection of textual attributes through a semantic-based masking method. Information Fusion. 2011;In Press, Accepted Manuscript.

Meystre S, Friedlin J, South B, Shen S, Samore M. Automatic de-identification of textual documents in the electronic health record: a review of recent research. BMC medical research methodology. 2010;10:70.

Nin J, Herranz J, Torra V. On the disclosure risk of multivariate microaggregation. Knowl Data Eng. 2008a;67:399-412.

Nin J, Herranz J, Torra V. Rethinking rank swapping to decrease disclosure risk. Knowl Data Eng. 2008b;64:346-64.

Oganian A, Domingo-Ferrer J. On the complexity of optimal microaggregation for statistical disclosure control. Statistical Journal of the United Nations Economic Commission for Europe. 2001;18:345-53.

Oliveira SRM, Zaïane OR. A privacy-preserving clustering approach toward secure and effective data analysis for business collaboration. Computers & Security. 2007;26:81-93.

Rada R, Mili H, Bicknell E, Blettner M. Development and application of a metric on semantic nets. IEEE Trans Syst Man Cybern. 1989;19:17-30.

Sánchez D, Batet M. Semantic similarity estimation in the biomedical domain: An ontology-based information-theoretic perspective. Journal of Biomedical Informatics.In Press, Corrected Proof.

Sanchez D, Batet M, Isern D. Ontology-based information content computation. Know-Based Syst. 2011;24:297-303.

Sanchez D, Batet M, Valls A, Gibert K. Ontology-driven web-based semantic similarity. J Intell Inf Syst. 2010;35:383-413.

Shin H, Vaidya J, Atluri V. Anonymization models for directional location based service environments. Computers & Security. 2010;29:59-73.

Sweeney L. Achieving *k*-anonymity privacy protection using generalization and suppression. Int J Uncertain Fuzziness Knowl-Based Syst. 2002a;10:571-88.

Sweeney L. k-anonymity: a model for protecting privacy. Int J Uncertain Fuzziness Knowl-Based Syst. 2002b;10:557-70.

Terrovitis M, Mamoulis N, Kalnis P. Privacy-preserving anonymization of set-valued data. Proc VLDB Endow. 2008;1:115-25.

Torra V. Microaggregation for Categorical Variables: A Median Based Approach. In: Domingo-Ferrer J, Torra V, editors. Privacy in Statistical Databases: Springer Berlin / Heidelberg; 2004. p. 518-.

Torra V. Towards knowledge intensive data privacy. Proceedings of the 5th international Workshop on data privacy management, and 3rd international conference on Autonomous spontaneous security. Athens, Greece: Springer-Verlag; 2011. p. 1-7.

Torra V, Domingo-Ferrer J. Record Linkage methods for multidatabase data mining. In: Torra V, editor. Information Fusion in Data Mining: Springer; 2003.

Torra V, Miyamoto S. Evaluating Fuzzy Clustering Algorithms for Microdata Protection. In: Domingo-Ferrer J, Torra V, editors. Privacy in Statistical Databases: Springer Berlin / Heidelberg; 2004. p. 519-.

Wei J, Mummoorthy M, Chris C, Luo S. t-Plausibility: Semantic Preserving Text Sanitization. 2009. p. 68-75.

Willenborg L, Waal Td. Elements of Statistical Diclosure Control: Springer; 2001.

Wu Z, Palmer M. Verbs semantics and lexical selection. the 32nd annual meeting on Association for Computational Linguistics. Las Cruces, New Mexico: Association for Computational Linguistics; 1994. p. 133-8.

Yancey W, Winkler W, Creecy R. Disclosure Risk Assessment in Perturbative Microdata Protection. In: Domingo-Ferrer J, editor. Inference Control in Statistical Databases: Springer Berlin / Heidelberg; 2002. p. 49-60.