# Semantic Anonymisation of Set-valued Data

Montserrat Batet[1], Arnau Erola[1], David Sánchez[1] and Jordi Castellà-Roca[1]
[1]*UNESCO Chair in Data Privacy,*
*Departament d'Enginyeria Informàtica i Matemàtiques,*
*Universitat Rovira i Virgili,*
*Av. Països Catalans, 26, 43007, Tarragona, Catalonia, Spain*
*{montserrat.batet,arnau.erola,david.sanchez,jordi.castella}@urv.cat*

Keywords: Data semantics, Set-valued data, Privacy, Microaggregation, Knowledge bases.

Abstract: It is quite common that companies and organisations require of releasing and exchanging information related to individuals. Due to the usual sensitive nature of these data, appropriate measures should be applied to reduce the risk of re-identification of individuals while keeping as much data utility as possible. Many anonymisation mechanisms have been developed up to present, even though most of them focus on structured/relational databases containing numerical or categorical data. However, the anonymisation of transactional data, also known as set-valued data, has received much less attention. The management and transformation of these data presents additional challenges due to their variable cardinality and their usually textual and unbounded nature. Current approaches focusing on set-valued data are based on the generalisation of original values; however, this suffers from a high information loss derived from the reduced granularity of the output values. To tackle this problem, in this paper we adapt a well-known microaggregation anonymisation mechanism so that it can be applied to textual set-valued data. Moreover, since the utility of textual data is closely related to their meaning, special care has been put in preserving data semantics. To do so, appropriate semantic similarity and aggregation functions are proposed. Experiments conducted on a real set-valued data set show that our proposal better preserves data utility in comparison with non-semantic approaches.

## 1 INTRODUCTION

It is quite usual to find databases in which records contain variable-length multi-valued attributes describing 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 & Naughton, 2009), or outcomes of a clinical record (He, *et al.*, 2008). Data sets with these characteristics are usually referred as *set-valued data* (Terrovitis, *et al.*, 2008). Due to their sensitive nature, the publication of this kind of data may compromise individuals' privacy, especially when adversaries have partial knowledge of individuals' actions.

To minimise the disclosure risk of published data, anonymisation/masking methods have been proposed (Domingo-Ferrer, 2008). These methods perform transformations over potentially identifying values thus reducing their level of specificity and/or creating groups of indistinguishable individuals. These transformations distort input data, making it less specific or detailed. Since the utility of anonymised data is closely related to the amount of information loss caused by the transformation, anonymisation methods should balance the trade-off between information loss and disclosure risk (Domingo-Ferrer, 2008).

Within the Statistical Disclosure Control (SDC) community, authors have proposed many techniques to anonymise structured/relational databases, consisting of records with several *univalued* attributes, each one corresponding to a *different* feature of the described entity (Domingo-Ferrer, 2008; Herranz, *et al.*, 2010; Jing, *et al.*, 2011; Martínez, *et al.*, 2012b; Matatov, *et al.*, 2010). In these methods, identifying attributes are removed, and quasi-identifier attributes (groups of attributes that result in unique combinations of values) are anonymised. Thus, the anonymisation process can

manage attribute values individually. Anonymisation of quasi-identifiers is usually done with *microaggregation* methods, which ensures that the masked database fulfils the *k*-anonymity property (Samarati & Sweeney, 1998; Sweeney, 2002) while achieving some compromise between data utility and disclosure risk (Domingo-Ferrer, 2008; Herranz, *et al.*, 2010).

However, the application of these methods to set-valued data faces additional problems. Unlike relational databases, set-valued data sets do not constitute well-defined sets of quasi-identifying attributes, because several subsets of values of an attribute could play the role of quasi-identifiers. Moreover, the set of values of each individual may have variable length and high dimensionality, compared to the relatively few attributes and values of relational records. Moreover, while attributes in relational databases are commonly either numerical or categorical, set-valued data sets are usually free text (e.g. query logs or other transactional data). The management of textual values add new challenges that are not considered by methods focused on numerical or categorical data.

In contrast to numerical data, which can be compared and transformed by means of mathematical operators, textual data require from comparison and aggregation operators that consider the meaning of words since, as acknowledged by several authors (Martínez, *et al.*, 2012b; Torra, 2011), the utility of textual data is closely related to the preservation of their semantics. Since semantics are an inherently human feature, the interpretation of textual data requires the exploitation of some sort of human-tailored machine-readable knowledge source, such as taxonomies, folksonomies and ontologies (Guarino, 1998). This allows mapping words to their conceptual abstractions, analysing the latter according to the semantic interrelations modelled in the knowledge source.

Considering the above challenges, specific anonymisation methods should be designed for textual set-valued data (Terrovitis, *et al.*, 2008). As it will be discussed in section 2, existing approaches on the anonymisation of set-valued data are based on generalising original values according to a hierarchical structure, so that the masked data set fulfils the *k*-anonymity property (He & Naughton, 2009; Terrovitis, *et al.*, 2008). These methods implicitly consider data semantics. However, they are affected by the large information loss resulting from the need of generalising concepts to a common abstraction, which causes a loss of granularity, and is seriously hampered by the presence of outlying values.

In order to not incur in such a high loss of granularity, in this paper we present an anonymisation method based on *microaggregation* (an approach originally designed for uni-valued attributes from relational databases), which can be applied to set-valued data in a natural way to preserve data semantics as much as possible. To do so, we propose a set of semantic operators to compare, sort and aggregate set-valued data from a semantic perspective, by using ontologies as the knowledge bases that guide the anonymisation process. The proposed method has been evaluated using a *real* set-valued data set consisting of search queries extracted from the AOL files and a widely used knowledge base.

The rest of the paper is organised as follows. Section 2 surveys and reviews anonymisation methods and approaches focusing on set-valued data. Section 3 introduces the basis of data anonymisation via microaggregation and details its adaptation to the anonymisation of set-valued data from a semantic perspective. Section 4 details the evaluation of our approach with regards to the preservation of data semantics. The final section contains the conclusions and depicts some lines of future research.

# 2 RELATED WORK

Anonymisation methods can be classified as *perturbative* and *non-perturbative* (Hundepool, *et al.*, 2012). The former distort original data while the latter reduce data detail or suppress them partially to fulfil the privacy criterion.

*Microaggregation* is a perturbative method that was originally defined for numerical data (Defays & Nanopoulos, 1993; Domingo-Ferrer & Mateo-Sanz, 2002). By using this approach, records are grouped and replaced by a prototypical record, so that they become indistinguishable, from at least, *k*-1 other records, thereby achieving the *k*-anonymity property (Samarati, 2001; Sweeney, 2002). There have been some attempts to extend microaggregation so as to be used with nominal attributes. However, most of them ignore data semantics. This is a drawback because, as mentioned by several authors (Martínez, *et al.*, 2012b; Torra, 2011), the lack of a semantically-coherent analysis compromises the utility of the anonymised results.

In (Torra, 2004), only ordinal categorical attributes are addressed and the median is proposed

as aggregation operator. In (Domingo-Ferrer & Torra, 2005), the equality/inequality predicate is used to compare nominal attributes and the modal value is proposed as an aggregation operator.

Non-perturbative methods based on attribute value *generalisations* implicitly consider data semantics (Li & Li, 2008; Samarati, 2001; Sweeney, 2002). These methods substitute attribute values by more general ones obtained from a hierarchical structure so that they also become indistinguishable (i.e. *k*-anonymous). The generalisations used to perform this substitution are selected in order to minimise the amount of information loss. These methods depend on the suitability of the hierarchical structure with regard to the input data and the granularity and level of detail of the taxonomy to minimise the loss of information resulting from value generalisations. For that reason, ad-hoc hierarchical structures, named Value Generalisation Hierarchies (VGH), are usually defined for the input data set. However, VGHs offer rough and overspecified knowledge sources in comparison with fine-grained and general or domain ontologies (Martínez, *et al.*, 2012b) and can be hardly defined for dynamic and unbounded domains such as query logs.

To the best of our knowledge, existing works on anonymisation of set-valued data follow the non-perturbative model based on value generalisations. In (Terrovitis, *et al.*, 2008) the authors anonymise textual set-valued data by proposing generalisations of input values according to ad hoc constructed VGHs, which iteratively generalise input values up to a common node until they become *k*-anonymous. He and Naughton (He & Naughton, 2009) adapted the previous method by starting from the most abstract generalisation and by specialising it progressively. The algorithm starts by generalising all items to the root of the hierarchy. Then, the algorithm recursively splits the current partition into sub-partitions until no further split is possible without violating *k*-anonymity.

Although non-perturbative methods based on generalisations take data semantics into account, they are affected by the large information loss resulting from concept generalisation, which necessarily cause loss of granularity. This is especially evident for heterogeneous data in which the need to generalise outliers results in abstract concepts (e.g. the root node of the ontology) and high information loss (Martínez, *et al.*, 2012c). On the contrary, perturbative methods based on microaggregation do not incur in a loss of granularity but scarcely consider data semantics.

To gain the benefits of both approaches and minimise their shortcomings, in this work, we propose an anonymisation method based on microaggregation, which can be applied to set-valued data sets while also considering the semantics of textual values by relying on available taxonomies/ontologies.

# 3 SEMANTIC PRESERVING ANONYMISATION OF SET-VALUED DATA

Microaggregation perturbs input data to generate *k*-anonymous data sets. To that end, input records, which are considered as standard records of relational databases, are clustered into groups of, at least, size *k* (*data partition*) and replaced by the cluster centroid (*data anonymisation*). In this manner, each record becomes indistinguishable from, at least, *k-1* other ones. To maximise the utility of anonymised data, similar records should be clustered together, so that the information loss resulting from the replacement by their centroid can be minimised.

Since optimal microaggregation is NP-hard (Oganian & Domingo-Ferrer, 2001), several heuristic algorithms have been proposed in the past. One of the most popular is the MDAV (*Maximum Distance Average Vector*) method (Domingo-Ferrer & Mateo-Sanz, 2002), which was specifically designed to minimise the information loss (Domingo-Ferrer, *et al.*, 2006; Martínez, *et al.*, 2012a). Consequently, we take the MDAV algorithm as the base to design our anonymisation method.

Algorithmically, MDAV performs the *data partition* by calculating the centroid of the whole data set and selecting the most distant record to it. Then, a cluster is constructed with the *k-1* least distant records. After that, the most distant record to the already clustered one is selected and a new cluster is constructed. The process is repeated until less than *2k* records remain ungrouped. The rest of records are grouped together in a last cluster. As a result, all clusters will have *k* records, except for the last one, which may have from *k* to *2k-1* records. *Data anonymisation* is performed by replacing each record of each cluster by the centroid of the cluster.

Figure 1 summarised this anonymisation process. The *m* records to be anonymised are partitioned in clusters of size *k* by the MDAV algorithm. Then, in the data anonymisation stage, records are replaced

by the centroid of the cluster to which they belong, thus obtaining a *k*-anonymous data set.

In the following section we present an adaptation of the MDAV microaggregation algorithm to support the anonymisation of textual set-valued data that puts special efforts in preserving the semantics of input data.

## 3.1 Set-valued Data Partition

To produce a *k*-anonymous partition of input data, MDAV relies on two basic functions that depend on the type of data to be processed: a *comparison* operator that measures the *distance* between records to add new ones in a cluster, and an *averaging* function to calculate the *centroid* used to guide the clustering process.

MDAV have been originally designed to deal with numerical data and structured databases with uni-valued attributes. Due to the characteristics of set-valued data and textual data, the adaptation of MDAV to this kind of data is not trivial. Contrary to numerical data that can be compared, averaged and transformed by means of mathematical functions, textual data require from operators that take their semantics into account (Martínez, *et al.*, 2012a). Moreover, as set-valued data have variable length, the coherent comparison/aggregation of records with different cardinalities is also challenging.

In this section, we propose a semantically-grounded comparison measure and we describe a set of averaging operators suitable for data with variable length.

### 3.1.1 Comparing Set-valued Data

To enable a semantic interpretation of textual values of items in the set, we first need to map them with their formal semantics. Since textual data may refer to one (e.g. a term, such as *"iPhone"*) or several concepts (e.g. a list of terms, such as *"AC charger for an iPhone"*), we first apply several morpho-syntactic analyses to the input data: *sentence detection*, *tokenisation*, *part-of-speech (POS) tagging* and *syntactic parsing*). As a result, noun phrases, which are the textual units which carry most of the semantics of the discourse are detected (e.g. *AC charger*, *iPhone*). Each noun phrase would refer to an individual concept (Sánchez, *et al.*, 2013). Thus, we map each noun phrase to a conceptual abstraction (e.g. *iPhone -> Smartphone*) by matching noun phrases and concept labels modelled in a knowledge base, such as an ontology. Notice that the core semantics of a noun phrase are carried by the noun most on the right, which can be qualified or specialised by adding new nous our adjectives to the left. Thus, in such cases in which the noun phrase is not found in the knowledge base we iteratively discard the words most on the left of the noun phrase until the result is found in the ontology (e.g. *a new iPhone -> new iPhone -> iPhone*).

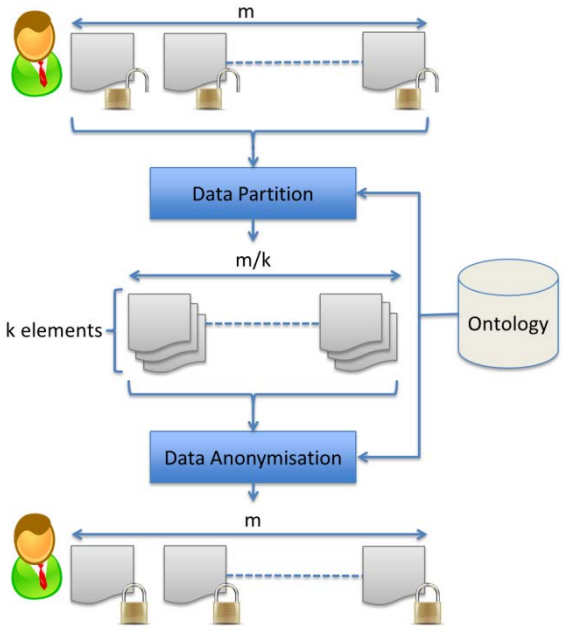



Figure 1: Semantic anonymisation process.

Formally, let $X=\{x_1,\ldots,x_m\}$ be the set of $m$ records represented by a unique multi-valued attribute (see figure 1), and let $x_r=\{q_1,\ldots,q_p\}$ be the items contained in the set-valued attribute of the record $x_r$. As a result of the conceptual mapping, each record is represented by a set of concepts $C_{xr}=\{c_1,\ldots,c_i,\ldots,c_n\}$ (e.g. $C_{xr}=\{iPhone, AC\ charger\}$), where each concept is taxonomically modelled in an ontology (e.g. $T(c_1)$= *iPhone -> Smartphone -> Handhelds -> Systems -> Computers*). This knowledge represents the basis that will enable a semantically-coherent comparison between records.

First, we propose a measure that computes the *semantic distance* between concepts by exploiting the knowledge modelled in their taxonomical trees. Several semantic measures can be found in the area of computational linguistics to estimate the distance between concepts modelled in a taxonomy. Notice that the availability of large and fine grained ontologies with a good coverage of analysed concepts will certainly improve the similarity accuracy. The most basic similarity measures

compute the length of the path that connects two concepts through their taxonomical specialisations/generalisations (Wu & Palmer, 1994). However, due to their simplicity, they omit much of the taxonomical knowledge explicitly modelled in the knowledge base, thus achieving a relatively low accuracy (Sánchez, *et al.*, 2012). More recent works (Batet, *et al.*, 2011; Sánchez, *et al.*, 2012) significantly improve these basic methods by evaluating *all* the taxonomical ancestors of the compared terms: they measure the distance between concepts as a function of the amount of their shared and non-shared taxonomical generalisations. These methods consider more information than path-based distances, which uses the path as an assessor of distance but that does not give clues on potential similarities of the compared concepts (given, for example, by their number of shared ancestors). In this work, we follow the same principles.

Given a pair of concepts $c_1$, $c_2$, we evaluate their distance $\delta_s(c_1, c_2)$ according to the amount of non-shared taxonomical generalisations in an ontology $O$. Moreover, we can also presume that concept pairs that have many generalisations in common are less distant than those sharing a small amount of generalisations. Hence, the semantic distance is computed as the ratio between the amount of non-shared concepts and the sum of shared and non-shared concepts (1):

$$\delta_s(c_1, c_2) = \frac{\left|T(c_1) \cup T(c_2)\right| - \left|T(c_1) \cap T(c_2)\right|}{\left|T(c_1) \cup T(c_2)\right|} \quad (1)$$

Where $T(c_i)=\{c_j \in O \mid c_j \text{ generalises } c_i\} \cup \{c_i\}$ represents the taxonomic generalisations of the concept $c_i$ in the ontology $O$, including $c_i$. Notice that, by including the compared concepts in $T$, we are able to distinguish different concepts that have all their generalisations in common from two identical concepts.

However, since data partition for microaggregation-based anonymisation of set-valued data requires comparing sets of concepts (instead of individual pairs), the above measure has to be extended. To do so, it is necessary to integrate distance values between sets of different cardinalities in a coherent manner. This can be done by means of aggregation functions, which, for example, are widely used in the area of bioinformatics to compare the functional similarity between gene products. In (Lord, *et al.*, 2003), authors used the average of all pairwise similarities; in (Sevilla, *et al.*, 2005), authors used the maximum of the pairwise similarity; in (Couto, *et al.*, 2007; Schlicker, *et al.*, 2006), authors used the composite average, where each term of the first set is paired only with the most similar term of the second set and vice-versa. In this paper, we will consider the following ones:

- *Minimum*: all concepts of $x_1$ are compared with all concepts of $x_2$, taking the minimum distance value as the result of this comparison (2):

$$D_{sMin}(C_{x_1}, C_{x_2}) = \underset{\forall c_i \in C_{x_1}}{Min}\left(\underset{\forall c_j \in C_{x_2}}{Min}\ \delta_s(c_i, c_j)\right) \quad (2)$$

- *Maximum*: provides the maximum distance value between all concept pairs (3):

$$D_{sMax}(C_{x_1}, C_{x_2}) = \underset{\forall c_i \in C_{x_1}}{Max}\left(\underset{\forall c_j \in C_{x_2}}{Max}\ \delta_s(c_i, c_j)\right) \quad (3)$$

- *Average*: computes the average distance between all concepts of $x_1$ and all concepts of $x_2$ (4):

$$D_{sAvg}(C_{x_1}, C_{x_2}) = \frac{\sum_{i=1}^{|C_{x_1}|}\sum_{j=1}^{|C_{x_2}|}\delta_s(c_i, c_j)}{|C_{x_1}| \times |C_{x_2}|} \quad (4)$$

- *Normalised sum of minimum*: given a concept $c_i$ of $x_1$, we compare it against all concepts of $x_2$, taking the minimum distance value as the result of this comparison. This states the *highest evidence of similarity* between records with respect to the feature $c_i$.

$$D_{sMinSum}(C_{x_1}, C_{x_2}) = \frac{\sum_{i=1}^{|C_{x_1}|}\underset{j=1}{^{|C_{x_2}|}}Min(\delta_s(c_i, c_j)) + \sum_{j=1}^{|C_{x_2}|}\underset{i=1}{^{|C_{x_1}|}}Min(\delta_s(c_i, c_j))}{|C_{x_1}| + |C_{x_2}|} \quad (5)$$

By repeating the process and adding the distance value between each $c_i$ of $x_1$ against $x_2$, we obtain the aggregated distance from $x_1$ to $x_2$. Note that this distance may be different when evaluating it from $x_2$ to $x_1$. Hence, the final distance between $x_1$ and $x_2$ will be the sum between the distances computed from $x_1$ to $x_2$ and from $x_2$ to $x_1$. Finally, since different records can be compared regardless of the cardinality of their sets of items, we divide it by the number of concepts of both records

($|C_{x1}|$ and $|C_{x2}|$) in order to obtain normalised distance values between sets of items.
As an example, let us suppose that given two records, $x_1$ and $x_2$, we have obtained the following concepts from their set of values using an ontology:

$C_{x_1}$ ={"Swimming" , "Mediterranean"},

$C_{x_2}$ ={ "Windsurfing", "Mediterranean"}.

And that their generalisations are the following:

*T*("Swimming")= {"Sports", "Water Sports", "Swimming" }

*T*("Windsurfing")={"Sports", "Water Sports", "Windsurfing"}

*T*("Mediterranean")={"Regional", "Europe", "Regions", "Mediterranean"}

Hence, the distance between records is computed as $D_{sMinSum}(C_{x_1}, C_{x_2})$ = (((1×0.5)+(1×0)+(1×0.5)+(1×0))/(2+2)) = 0.25.

Notice that, for example, the semantic distance between $c_i$ = "Swimming" and $c_j$="Windsurfing" (eq. 1) is computed as $\delta_s(c_i,c_j)$=((4-2)/4)=0.5.

- *Normalised sum of maximum*: the same as above but taking the maximum distance value instead of the minimum. This states the *highest evidence of dissimilarity* between records with respect to the feature $c_i$.

$$D_{sMaxSum}(C_{x_1}, C_{x_2}) = \frac{\sum_{i=1}^{|C_{x_1}|} \underset{j=1}{^{|C_{x_2}|}}{Max}(\delta_s(c_i,c_j)) + \sum_{j=1}^{|C_{x_2}|} \underset{i=1}{^{|C_{x_1}|}}{Max}(\delta_s(c_i,c_j))}{|C_{x_1}|+|C_{x_2}|} \quad (6)$$

In the next section, we generically refer to these aggregation distances (eq. 2 to 6) as $Ds(C_{xr}, C_{xi})$.

### 3.1.2 Record Aggregation

As stated at the beginning of the section, MDAV creates clusters by picking up the most distant record to the data set centroid. Moreover, centroids are also used at the data anonymisation stage since cluster elements are replaced by cluster centroids to become *k*-anonymous (see figure 1).

Numerically, the centroid of a group/data set is understood as the value (or the set of values in our case) that minimises the distance against all records in the data set. When dealing with continuous-scale numerical data, the centroid can be accurately computed by averaging numerical values. However, for textual data, the centroid must necessarily be discretised. In this case, some authors (Domingo-Ferrer & Torra, 2005) select the centroid of textual/categorical data sets by picking up those who appear the most (i.e. the mode). However, this approximation omits the semantics of data.

Given the aggregated distances presented above (eq. 2 to 6), we use them to discover the data set/cluster centroid, which is selected as the record that minimises the sum of distances to all other records in a given data set or cluster:

$$centroid\left(C_{x_1},...,C_{x_m}\right) = arg\ min_{\forall C_{x_r}} \left\{ \sum_{i=1}^{m} D_s\left(C_{x_r}, C_{x_i}\right) \right\} \quad (7)$$

Where $\{C_{x1},\ldots,C_{xm}\}$ corresponds to the set of concepts that represent the set-valued attribute of the records in the data set/cluster to evaluate, and $Ds(C_{xr}, C_{xi})$ is the same aggregated distance as the one used in the comparison of two sets of values (eq. 2 to 6).

## 3.2 Anonymising Set-valued Data

By using the above-proposed distance and centroid calculus on the MDAV algorithm, records will be grouped into $d=m/k$ clusters of, at least, *k* records (see figure 1). To fulfil the *k*-anonymity property, the last step requires replacing all records of each cluster by a representative, which usually corresponds to the centroid of the cluster. Since this centroid minimises the individual distances to all records in the cluster, the information loss resulting from this replacement will be minimised.

In a general microaggregation scenario, this centroid corresponds exactly to the "central" element of the cluster. However, in the textual set-valued anonymisation context, this may lead to undesirable consequences. Particularly, the fact that elements in a cluster are replaced by the *exact* centroid record may excessively expose her identity, especially if an attacker has partial knowledge (e.g. some items of her set of values are known (He & Naughton, 2009).

To palliate this problem, in some works (Terrovitis, *et al.*, 2008) the cluster representative is

synthetically built by replacing *terms* in clusters with *concepts* that generalise all/some of them according to a background taxonomy. Hence, anonymised records would be composed by sets of concepts rather than the original terms. This fact hampers the utility of the anonymised records in some environments in which original terms (instead of their conceptual abstraction) are needed, such as query formulation analysis (Bar-Ilan, 2007; Xiong & Agichtein, 2007).

In this work, we have chosen an intermediate solution that aims at retaining the semantic and syntactical utility of records while, at the same time, minimising the disclosure risk of the centroid record by creating a *synthetic* record. On the one hand, our cluster representative corresponds to the record that constitutes the centroid of the cluster. Notice that since we are working with sets of concepts representing a record, this corresponds to the $C_{xr}$ that minimises the semantic distance to all other sets of concepts (i.e. records) in the cluster, according to the centroid calculus (see eq. 7). Next, instead of recovering the concrete terms of the centroid record $x_r$, we replace concepts in $C_{xr}$ by suitable terms picked from the original data set. Specifically, each concept in $C_{xr}$ is replaced by a term taken *randomly* from those in the records from the whole input data set that corresponds to that concept (e.g. if the concept *Smartphone* appears in $C_{xr}$, then, we may retrieve suitable terms like *"iPhone 3"*, *"Samsung Galaxy S2"* or *"Nexus 4"* if those are in the input data set).

As a result of the above process, records of each cluster are anonymised by replacing each one with a *synthetic* record that semantically matches the cluster centroid (i.e. it maintains the semantics of the centroid) while protecting the privacy of the centroid record (see figure 1). On the one hand, since individual terms are picked randomly from different records of the input data set, we minimise the chance that cluster representatives contain *exact* subsequences of terms of individual records, a circumstance that may compromise its anonymity (He & Naughton, 2009). On the other hand, since selected terms match the concepts of the centroid record, we also retain semantics accurately, as the centroid record is the one that better represents the semantic features of the records in the cluster and, hence, the one that minimises the information loss resulting from the anonymisation process. Moreover, since concepts are randomly replaced by one of its corresponding terms using a uniform distribution, the distribution of individual terms found in the input data for each concept will be likely maintained, even though terms would not be unequivocally associated to the original records. Finally, since we are publishing *real* terms from the input data set, anonymised data can still be useful for tasks such as statistical analysis.

# 4 EVALUATION

In this section, we first introduce the measure used to quantify the degree of semantic preservation of the anonymised data set. Then, in section 4.2, we describe the evaluation data set and, finally, in section 4.3, the proposed method is evaluated from two perspectives: (1) the suitability of a semantically-grounded anonymisation, and (2) the influence of the different aggregation measures used to compute the distance between value sets.

## 4.1 Evaluation of Semantics Preservation

To evaluate up to which point an anonymisation method retains the utility of original data, that is, preserves their semantic content, we measure the *information loss* (L) between original and masked records from a semantic perspective. To measure information loss, we computed the well-known *Sum of Square Errors* (SSE) between original and masked records, which is the most usual measure employed by privacy-preserving methods based on microaggregation (Domingo-Ferrer & Mateo-Sanz, 2002; Domingo-Ferrer, *et al.*, 2006; Lin, *et al.*, 2010; Martínez, *et al.*, 2012a; Torra & Miyamoto, 2004). The SSE is defined as the sum of squares of attribute distances between original records and their versions in the anonymised data set. To measure the aggregated set of all semantic distances that SSE requires, we used the semantic distance $D_{sAvg}$ defined in section 3.1.1 (see (4)). Thus, the higher the SEE is, the higher the information loss will be. This is due to the replacement of values and the lower preservation of the data semantics.

Formally, given that $X$ is the set of original records and $X^A$ is the anonymised version, the information loss of masked data against its non-masked one is computed as (8):

$$L = SSE\left(X, X^A\right) = \left(\sum_{i=1}^{m} D_{sAvg}\left(C_{x_i}, C_{x_i}^A\right)\right)^2 \quad (8)$$

Notice that with a high information loss (i.e. a high SSE), a lot of data uses are severely damaged like, for example, subdomain analysis, that is, analysis restricted to parts of the data set.

## 4.2 Evaluation Data Set

The evaluation has been performed using *real* set-valued data which correspond to query logs extracted from the AOL log files released in 2006. From these, the query logs of 1,000 users have been randomly taken. They contain about 56,000 individual queries. Even though personal identifiers have been removed from published query logs, queries themselves may enable identity disclosure due to their specificity and personal nature (Barbaro & Zeller, 2006). This fact together with its textual nature, makes this data set suitable for testing our set-valued anonymisation method.

Different *k*-anonymity degrees between 2 and 5 have been tested. These *k*-values produced between 500 and 200 clusters as a result of the aggregation process.

As knowledge base, we use ODP (Open Directory Project, 2010). ODP is a multilingual open content directory of World Wide Web links. The purpose of ODP is to list and categorise web sites. It uses an ontology scheme to classify sites into different subjects. ODP offers more than 1 million taxonomically structured categories. Contrary to other knowledge bases, ODP covers recently minted terms and named entities, which are very usually referred in web queries (Sánchez, *et al.*, 2013). This helps to improve the recall of the query-category matching process. In fact, despite the sensitivity of named entities due to their high degree of concreteness, they are, in essence, noun phrases that can be associated to conceptual abstractions if they are covered in an ontology such as ODP. In fact, ODP has been extensively used in other works dealing with AOL queries, such as (Sánchez, *et al.*, 2013). Thus, no special treatment for named entities is necessary.

## 4.3 Discussion

In this section, we discuss the suitability of the semantically-grounded anonymisation of set-valued data, and the influence of the different aggregation measures presented in section 3.1.1.

In order to assess the improvement obtained by considering data semantics and to put into context the absolute information loss figures, we have also implemented a simplified version of the semantic aggregation distances defined in section 3.1.1, in which *no semantics* are considered at all, which is the usual scenario in related works on microaggregation (see section 2). In this case, neither the semantic distance between concepts $\delta_s$ (see eq.1) nor ODP are used. The terms of the set (i.e. queries) are treated as simple strings and compared according to their equality/inequality (Domingo-Ferrer & Torra, 2005). Hence, the non-semantic distance $\delta$ (see eq. 9) between concepts $c_1$ and $c_2$ is used instead of $\delta_s$:

$$\delta(c_1, c_2) = \begin{cases} 0 & if\ c_1 = c_2 \\ 1 & if\ c_1 \neq c_2 \end{cases} \quad (9)$$

We named the non-semantic versions of the aggregation distances as $D_{Min}$, $D_{Max}$ $D_{Avg}$, $D_{MinSum}$, $D_{MaxSum}$.

By analysing information loss (L) figures (see Table 1), we observe that all aggregation measures that do not consider the semantics of terms result in a higher information loss than their semantic version for all *k*-values. For example, the non-semantic version of the *normalised sum of minimum* distance, $D_{MinSum}$, obtains L values of 817, 859, 879 and 891 for k=2, k=3, k=4 and k=5 respectively, while its semantic version, $D_{sMinSum}$, obtains values of 683, 733, 759, 781. This represents around a 16% of improvement.

This is explained because, even though terminological resemblance is an evidence of semantic similarity, it poorly captures and evaluates the meaning of terms. This is especially evident in free text data in which the same terms may appear with different morphological forms or when synonymous words are used. Moreover, most terms in the data set are unique, so that, few evidences of similarity can be gathered to guide the partition and aggregation process.

Our approach exploits ODP to retrieve categories to which queries refer. Since ODP categories are conceptualisations of textual queries, they enable a semantically-coherent partition and aggregation of query logs. Hence, the obtained improvement is the result of considering the semantics of terms during the comparison between queries, the centroid selection, the cluster construction and the anonymisation stages. Note also that the morpho-syntactic analyses applied to identify noun phrases (see section 3.1.1) and to map them to ontological concepts (i.e. categories in ODP) also contribute to

improve the conceptual mapping recall and to provide a better interpretation of data semantics.

Table 1: Information loss (L) of the evaluated aggregation functions for different levels of *k*-anonymity with and without considering the data semantics.

| Aggregation distance | k=2 | k=3 | k=4 | k=5 |
|---|---|---|---|---|
| $D_{sMin}$ | 699 | 759 | 783 | 799 |
| $D_{Min}$ | 825 | 863 | 883 | 895 |
| $D_{sMax}$ | 724 | 778 | 806 | 820 |
| $D_{Max}$ | 824 | 860 | 884 | 895 |
| $D_{sAvg}$ | 639 | 692 | 724 | 745 |
| $D_{Avg}$ | 768 | 804 | 829 | 845 |
| $D_{sMinSum}$ | 683 | 733 | 759 | 781 |
| $D_{MinSum}$ | 817 | 859 | 879 | 891 |
| $D_{sMaxSum}$ | 716 | 780 | 819 | 838 |
| $D_{MaxSum}$ | 822 | 858 | 884 | 893 |

Even though semantics of anonymised data are better retained (i.e. information loss of anonymised data is minimised), the fact that the aggregation is made by randomly rearranging terms referring to the same concepts of different records for the concepts corresponding to the centroid record, contributes in reducing the chance that cluster representatives contain exact subsequences of terms of individual records, while preserving the distribution of the terms in the original data.

On the other hand, different information loss figures are obtained according to the semantic aggregation distance. The worst results are obtained by $D_{sMax}$ and $D_{sMaxSum}$. The problem of these aggregation distances is that they return the maximum dissimilarity between two sets of values. This goes against the notion of cluster cohesion (which is what SSE measures) when arranging records in clusters, and makes the process very sensitive to the presence of outlying values.

By contrast $D_{sMin}$ and $D_{sMinSum}$ distances state the highest evidence of similarity between records, and thereby, provide notably better results. However, $D_{sMin}$ is unable to assess the global distance between two sets of values because it detects if two sets share a value, but it is indifferent to the number of unrelated terms and to what extend they are different. This problem is clearly overcome by the $D_{sMinSum}$ distance, which also considers the distance between unrelated terms because, for each term of a set, it takes into account the most similar term in the other set.

The best results are however obtained by $D_{sAvg}$ because it accounts for similar and dissimilar terms. This fact benefits data sets such as query logs with a heterogeneous and unbounded nature. However, it may not be the best option for data sets composed by records with several shared or similar terms, such as the measurement of the functional similarity between gene products (Pesquita, *et al.*, 2009), because similarity can be distorted by few different terms. In that case, $D_{sMinSum}$ would likely obtain better results.

# 5 CONCLUSIONS

This paper presents an anonymisation method for set-valued data based on semantic microaggregation. While most of the research on privacy protection of set-valued data focuses on term generalisation, which produce a high information loss due to the loss of granularity of output values, our method has been especially designed to preserve the data semantics and, consequently, to improve data utility.

To achieve this goal, textual data are semantically interpreted by extracting their conceptualisations from an ontology. This enables to aggregate set-valued data from a semantic perspective by means of an adaptation of the MDAV algorithm. Moreover, suitable semantic operators to compare and average set-valued data have been proposed for that purpose. Finally, synthetic records that semantically match the cluster centroids are generated by randomly picking values from different records of the input data set. These records preserve the meaning of the record that better represents the semantics of the elements in the cluster.

The evaluation, carried out with a set of real query logs extracted from the AOL data set and a publicly available knowledge base (ODP), sustains the practical suitability of our method.

As future work, we plan to test the behaviour of the proposed method in other domains in which textual transactional data are available (such as electronic health-care records), exploiting domain-specific knowledge bases (such as biomedical terminologies like SNOMED-CT (Spackman, 2004)). Within scenarios with more restricted set-valued data (e.g. lists of diseases) which can be properly covered by available knowledge bases, we plan to compare our method against non-perturbative methods that extensively rely on those knowledge bases to propose generalisations. Moreover, we plan to combine multiple ontologies in order to improve the recall of the conceptual mapping of textual terms (Batet, *et al.*, 2013). Finally, we plan to use application-oriented metrics to measure the utility of the protected data in

specific tasks, such as query-log based profiling or query refinement accuracy.

## ACKNOWLEDGEMENTS

This work was partly supported by the European Commission under FP7 project Inter-Trust, by the Spanish Ministry of Science and Innovation (through projects eAEGIS TSI2007-65406-C03-01, ICWT TIN2012-32757, CO-PRIVACY TIN2011-27076-C03-01, ARES-CONSOLIDER INGENIO 2010 CSD2007-00004, Audit Transparency Voting Process IPT-430000-2010-31 and BallotNext IPT-2012-0603-430000), by the Spanish Ministry of Industry, Commerce and Tourism (through projects eVerification2 TSI-020100-2011-39 and SeCloud TSI-020302-2010-153) and by the Government of Catalonia (under grant 2009 SGR 1135). The authors are with the UNESCO Chair in Data Privacy, but they are solely responsible for the views expressed in this paper, which do not necessarily reflect the position of UNESCO nor commit that organisation.

## REFERENCES

Bar-Ilan, J. (2007). Access to query logs - an academic researcher's point of view. Proc. of the Proceedings of the Query Log Analysis: Social and Technological Challenges Workshop at the 16th World Wide Web Conference, WWW2007). Banff, Alberta, Canada.

Barbaro, M., & Zeller, T. (2006). A Face Is Exposed for AOL Searcher No. 4417749. The New York Times,

Batet, M., Sánchez, D., & Valls, A. (2011). An ontology-based measure to compute semantic similarity in biomedicine. Journal of Biomedical Informatics, 44, 118-125.

Batet, M., Sánchez, D., Valls, A., & Gibert, K. (2013). Semantic similarity estimation from multiple ontologies. Applied Intelligence, 38(1), 29-44.

Couto, F. M., Silva, M. J., & Coutinho, P. M. (2007). Measuring semantic similarity between Gene Ontology terms. Data & Knowledge Engineering, 61(1), 137-152.

Defays, D., & Nanopoulos, P. (1993). Panels of enterprises and confidentiality: the small aggregates method. Proc. of the 92 Symposium On Design and Analysis of Longitudinal Surveys (pp. 195-204). Otawwa, Canada.

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

Domingo-Ferrer, J., & Torra, V. (2005). Ordinal, Continuous and Heterogeneous k-Anonymity Through Microaggregation. Data Mining and Knowledge Discovery, 11(2), 195-212.

Domingo-Ferrer, J., Martínez-Ballesté, A., Mateo-Sanz, J. M., & Sebé, F. (2006). Efficient multivariate data-oriented microaggregation. The VLDB Journal, 15(4), 355-369.

Domingo-Ferrer, J. (2008). A survey of inference control methods for privacy preserving data mining Privacy preserving data mining: models and algorithms (Vol. 55-80).

Guarino, N. (1998). Formal Ontology in Information Systems. In N. Guarino (Ed.), Proc. of the 1st International Conference on Formal Ontology in Information Systems, FOIS-98 (pp. 3-15). Trento, Italy.

He, Y., & Naughton, J. F. (2009). Anonymization of SetValued Data via TopDown, Local Generalization. Proc. of the Thirtieth international conference on very large data bases (VLDB'09) (pp. 934-945). Lyon, France.

He, Z., Xu, X., & Deng, S. (2008). k-ANMI: A mutual information based clustering algorithm for categorical data. Information Fusion, 9(2), 223-233.

Herranz, J., Matwin, S., Nin, J., & Torra, V. (2010). Classifying data from protected statistical datasets. Computers & Security, 29(8), 875-890.

Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Nordholt, E. S., Spicer, K., & Wolf, P.-P. d. (2012). Statistical Disclosure Control: Wiley.

Jing, X., Zhang, N., & Das, G. (2011). ASAP: Eliminating algorithm-based disclosure in privacy-preserving data publishing. Information Systems, 36(5), 859-880.

Li, T., & Li, N. (2008). Towards optimal k-anonymization. Data & Knowledge Engineering, 65(1), 22-39.

Lin, J. L., Wen, T. H., Hsieh, J. C., & Chang, P. C. (2010). Density-based microaggregation for statistical disclosure contro. Expert Systems with Applications, 37, 3256-3263.

Lord, P., Stevens, R., Brass, A., & Goble, C. (2003). Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation. Bioinformatics, 19(10), 1275-1283.

Martínez, S., Sánchez, D., & Valls, A. (2012a). Semantic adaptive microaggregation of categorical microdata. Computers & Security, 31(5), 653-672.

Martínez, S., Sánchez, D., Valls, A., & Batet, M. (2012b). Privacy protection of textual attributes through a semantic-based masking method. Information Fusion, 13(4), 304-314.

Martínez, S., Valls, A., & Sánchez, D. (2012c). Semantically-grounded construction of centroids for datasets with textual attributes. Knowledge based systems., 35, 160-172.

Matatov, N., Rokach, L., & Maimon, O. (2010). Privacy-preserving data mining: A feature set partitioning approach. Information Sciences, 180(14), 2696–2720.

Oganian, A., & Domingo-Ferrer, J. (2001). On the complexity of optimal microaggregation for statistical disclosure control. Statistical Journal of the United Nations Economic Commision for Europe, 18, 345-353.

Open Directory Project (2010). http://www.dmoz.org,

Pesquita, C., Faria, D., Falcão, A. O., Lord, P., & Couto, F. M. (2009). Semantic Similarity in Biomedical Ontologies. PLoS Computational Biology, 5(7), 1-12.

Samarati, P., & Sweeney, L. (1998). Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression. Proc. of the Proceedings of the IEEE Symposium on Research in Security and Privacy, S&P). Oakland, CA.

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

Sánchez, D., Batet, M., Isern, D., & Valls, A. (2012). Ontology-based semantic similarity: a new feature-based approach. Expert Systems With Applications, 39, 7718-7728.

Sánchez, D., Castellà-Roca, J., & Viejo, A. (2013). Knowledge-Based Scheme to Create Privacy-Preserving but Semantically-Related Queries for Web Search Engines. Information Sciences, 218, 17-30.

Schlicker, A., Domingues, F. S., Rahnenführer, J., & Lengauer, T. (2006). A new measure for functional similarity of gene products based on Gene Ontology. BMC Bioinformatics, 7(3002).

Sevilla, J. L., Segura, V., Podhorski, A., Guruceaga, E., Mato, J. M., Martinez-Cruz, L. A., Corrales, F. J., & Rubio, A. (2005). Correlation between gene expression and GO semantic similarity. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2(4), 330-338.

Spackman, K. A. (2004). SNOMED CT milestones: endorsements are added to already-impressive standards credentials. Healthcare Informatics, 21(9), 54-56.

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

Terrovitis, M., Mamoulis, N., & Kalnis, P. (2008). Privacy-preserving anonymization of set-valued data. Proceedings of the VLDB Endowment, PVLDB, 1, 115-125.

Torra, V. (2004). Microaggregation for Categorical Variables: A Median Based Approach. Proc. of the Privacy in Statistical Databases (PSD 2004) (pp. 162-174).

Torra, V., & Miyamoto, S. (2004). Evaluating Fuzzy Clustering Algorithms for Microdata Protection. In J. Domingo-Ferrer & V. Torra (Eds.), Privacy in Statistical Databases (pp. 519-519).

Torra, V. (2011). Towards knowledge intensive data privacy Proceedings of the 5th international workshop on data privacy management, and 3rd international workshop on autonomous spontaneous security, DMP'10/SETOP'10. (Vol. LNCS 6514, pp. 1-7). Athens, Greece: Springer-Verlag.

Wu, Z., & Palmer, M. (1994). Verb semantics and lexical selection. Proc. of the 32nd annual Meeting of the Association for Computational Linguistics (pp. 133-138). Las Cruces, New Mexico.

Xiong, L., & Agichtein, E. (2007). Towards privacy-preserving query log publishing. Proc. of the Proceedings of the Query Log Analysis: Social and Technological Challenges Workshop at the 16th World Wide Web Conference, WWW2007). Banff, Alberta, Canada.