ONTOLOGY BASED SEMANTIC ANONYMISATION OF MICRODATA

Sergio Martínez Lluís

Dipòsit Legal: T.451-2013

Sergio Martínez Lluís

ONTOLOGY BASED SEMANTIC ANONYMISATION OF MICRODATA

Ph. D. Thesis

Supervised by
Dr. Aida Valls and Dr. David Sánchez

Department of
Computer Science and Mathematics

UNIVERSITAT ROVIRA I VIRGILI

December, 2012

UNIVERSITAT ROVIRA I VIRGILI
ONTOLOGY BASED SEMANTIC ANONYMISATION OF MICRODATA
Sergio Martínez Lluís
Dipòsit Legal: T.451-2013

Acknowledgements

This work has been partially supported by the CONSOLIDER- INGENIO-2010 research project ARES (Advanced Research in Information Security and Privacy), funded by the Spanish ministry. In particular, the author has worked, as a member of the ITAKA research group, in the team called IF-PAD (Information Fusion for Privacy and Decision) that also includes people from Institut d' Investigació en Intel·ligència Artificial (IIIA-CSIC).

This work has also been partially supported by the Spanish Ministry of Science and Innovation (DAMASK project, *Data mining algorithms with semantic knowledge*, TIN2009-11005).

The author has been supported by the Universitat Rovira i Virgili predoctoral research and by a research grant of the Ministerio de Educación y Ciencia (Spain).

The author, finally, thanks the "Observatori de la Fundació d'Estudis Turístics Costa Daurada" and "Parc Nacional del Delta de l'Ebre (Departament de Medi Ambient i Habitatge, Generalitat de Catalunya)" for the data provided.

I would like to thank ITAKA group members.

To my family.

To my wife and my son.

Abstract

The exploitation of microdata compiled by statistical agencies is of great interest for the data mining community. However, such data often include sensitive information that can be directly or indirectly related to individuals. Hence, an appropriate anonymisation process is needed to minimise the risk of disclosing identities and/or confidential data. In the past, many anonymisation methods have been developed to deal with numerical data, but approaches tackling the anonymisation of non-numerical values (e.g. categorical, textual) are scarce and shallow. Since the utility of this kind of information is closely related to the preservation of its meaning, in this work, the notion of semantic similarity is used to enable a semantically coherent interpretation. Ontologies are the basic pillar to propose a semantic framework that enables the management and transformation of categorical attributes, defining different operators that take into account the underlying semantics of the data values. The application of the operators defined in this semantic framework to the data anonymisation task allows the development of three anonymisation methods especially tailored to categorical attributes: Semantic Recoding, Semantic and Adaptive Microaggregation and Semantic Resampling. In addition a new Semantic Record linkage method is proposed, which considers data semantics in order to more accurately evaluate the disclosure risk of anonymised non-numerical data. The proposed methods have been extensively evaluated with real datasets with encouraging results. Experimental results show that a semantic-based treatment of categorical attributes significantly improves the semantic interpretability and utility of the anonymised data.

Contents

List of figures

List of tables

Chapter 1
Introduction

The protection of individuals' privacy is a fundamental social right. To guarantee the privacy of people participating in surveys (e.g. questionnaires) whose data is made available for secondary use (e.g. medical research from electronic health-care records), an appropriate data anonymisation should be carried out prior publication. This is especially relevant in recent years with the enormous growth of the Information Society. In this context, collections of data associated to individuals are far more extended than some years ago. As a positive aspect, this kind of data is of outmost importance from the data analysis perspective. For example, the publication of census data is useful for market research by companies. Public medical records are valuable resources for clinical research and to improving medical treatments. The incorporation of genomic data into medical records is a significant healthcare advancement. On the negative side, this data usually contains sensitive information that puts individuals' privacy at a risk.

Even though Statistical Offices compiling and releasing *microdata* (a dataset organized as a matrix where each row corresponds to an individual and each column to an attribute) do not publish identifying attributes (e.g. ID-cards or person names), identity disclosure may still appear through statistical inference of combinations of published attributes. For example, in small towns, the publication of the birthplace, birth year and occupation of an individual, may unequivocally re-identify a person, due to the uniqueness of attribute value combinations in such a limited dataset. If other sensitive attributes (e.g. incomes, healthcare diagnosis, etc.) are also published and associated to attributes that may enable the re-identification, the individual's privacy will be compromised.

The *Statistical Disclosure Control* discipline aims at protecting statistical microdata in a way that it can be released and exploited without publishing any private information that could be linked with or identify a concrete individual. This is achieved by means of a masking algorithm that creates a new anonymised version of the original dataset.

Microdata may include numerical (as age or zip code) and non-numerical attributes (as occupation or country). In the past, many masking methods have been designed to deal with numerical data (Domingo-Ferrer 2008). Numbers are easy to manage, compare and transform. Moreover, the quality and utility of the anonymised dataset can be optimized by retaining a set of statistical features (Domingo-Ferrer 2008). However, the extension of these methods to non-

numerical attributes is not straightforward, because of the limitations on defining appropriate comparison and aggregation operators on symbols, which have a restricted set of categories. Non-numerical attributes have been treated as flat categorical variables, defining methods based on equality/inequality operators at a string level, or considering some kind of ordering between the words. In those methods, the quality of masked data is retained by solely preserving data distribution.

In addition to the classical categorical attributes that consider a limited set of categories, microdata compiled from polls or questionnaires may include attributes with large universes of values, such as responses to questions with free answers, like “main hobby” or “preferred type of food”. Even though these kind of non-numerical attributes provide a detailed and precise description of individuals, their utility has not been yet exploited due to the lack of proper anonymisation tools. Moreover, at the same time, the privacy of individuals is more critical when publishing this kind of attributes, due to the uniqueness of the answers.

The anonymisation and, at the same time, utility-preservation of this kind of attributes is challenging due to the lack of appropriate comparison and aggregation operators. In this case, word semantics plays a crucial role in the proper interpretation of these data, a dimension which is commonly ignored in the literature that considers categorical values. In fact, retaining the meaning of categorical values is of outmost importance to retain the utility of anonymised values (Torra 2011).

This thesis studies how to integrate Artificial Intelligence techniques focused on knowledge representation within the context of microdata anonymisation and SDC methods, an area that has been focused on mathematical statistics. Its originality consists on the management and transformation of categorical attributes from a semantic point of view rather than from a symbolic way.

The semantic interpretation of categorical attributes for masking purposes requires the exploitation of some sort of structured knowledge sources, which allow a mapping between words and their conceptual abstractions. To do so, in this work we rely in the well-known ontological paradigm. Ontologies are rigorous and exhaustive organizations of knowledge domains, modelling concepts and their interrelations (Guarino 1998). Works in other fields (Batet 2011) demonstrates that, exploiting the knowledge offered by ontologies, we are able to better interpret this kind of categorical or textual data.

The main objective of this thesis is the design of masking methods well suited for the anonymisation of categorical attributes from a semantic point of view. That is, we aim at obtaining an anonymised dataset that is as semantically similar as possible with respect to the original one. In this manner, the utility of categorical values, which is a function of their semantics, can be better preserved. To do so, in this thesis, we rely on the notion of ontology-based semantic similarity, which enables a semantically-coherent management of textual data. This is used as the basic pillar to propose a framework that defines semantically-grounded comparison and transformation operators. The framework is then applied to

propose or adapt SDC anonymisation and disclosure evaluation methods initially designed for numerical data, so that they can manage categorical attributes from a semantic perspective.

1.1 Framework of this Ph.D. thesis

This work is part of a CONSOLIDER- INGENIO research project funded by the Spanish ministry, called ARES (Advanced Research in Information Security and Privacy). ARES gathers around 60 people from six of the most dynamic Spanish research groups in the area of information security and virtually all existing Spanish groups in the area of information privacy. In particular, I have worked, as a member of the ITAKA research group, in the team called IF-PAD (Information Fusion for Privacy and Decision) that also includes people from *Institut d' Investigació en Intel·ligència Artificial* (IIIA-CSIC). The work of this thesis belongs to workpackage 4 of the ARES project, devoted to Data Privacy Technologies. I have been supported by a research grant of the Universitat Rovira I Virgili and a research grant of the Ministerio de Educación y Ciencia (Spain).

1.2 Objectives

The main objectives of the thesis can be summarized as follows:

1. To study works on Statistical Disclosure Control, focusing on those dealing with categorical data on metrics aimed to optimize the utility of anonymised categorical data.
2. To study the possibilities offered by ontologies and the notion of semantic similarity to guide the anonymisation from a semantic perspective.
3. To design a framework consisting on a set of semantic operators that enable the comparison and transformation of categorical attributes form a semantic point of view.
4. To apply the framework to design and adapt SDC methods so that they are especially suited to deal with categorical attributes. Concretely, data recording, data microaggregation and data resampling methods are considered.
5. To design a new method to better evaluate the disclosure risk of anonymised categorical attributes according to their semantics.
6. To implement the proposed methods and to evaluate and compare them against related works with a variety of datasets under the dimensions of information loss (as a function of data utility) and re-identification risk.

1.3 Contributions

The main contributions of this Ph.D. thesis to the field of ontology-based anonymisation of microdata can be summarised as follows:

A semantic framework for categorical data. Three semantically-grounded operators suitable to manage categorical data are presented: comparison, aggregation and sorting. The framework captures the semantics of data by relying on the measurement of the semantic similarity between terms, assessed from ontologies, while taking into also account the distribution of data. It has been published in the following journals:

1J. Martínez, S., Valls, A., Sánchez, D.: Semantically-grounded construction of centroids for datasets with textual attributes. International journal: Knowledge-Based Systems. 35(0), 160-172 (2012). *Impact Factor: 2.422*

2J. Martínez, S., Valls, A., Sánchez, D.: A semantic framework to protect the privacy of Electronic Health Records with non-numerical attributes. Journal of Biomedical Informatics. Accepted manuscript. *Impact Factor: 1.792*

A new recoding method for categorical data. We presented a new recoding method preserving the semantic content of categorical data. It has been published in the following journal:

3J. Martinez, S., Sanchez, D., Valls, A., Batet, M.: Privacy protection of textual attributes through a semantic-based masking method. International Journal: Information Fusion 13(4), 304-314 (2011). *Impact Factor: 1.467*

Preliminar work on this topic has been presented in the follow international conference papers:

1C. Martinez, S., Sanchez, D., Valls, A.: Ontology-Based Anonymization of Categorical Values. In: Torra, V., Narukawa, Y., Daumas, M. (eds.) Modeling Decisions for Artificial Intelligence, vol. 6408. Lecture Notes in Computer Science, pp. 243-254. Springer Berlin / Heidelberg, (2010). *CORE B*

2C. Martinez, S., Sanchez, D., Valls, A., Batet, M.: The Role of Ontologies in the Anonymization of Textual Variables. In: the 13th International Conference of the Catalan Association for Artificial Intelligence 2010, pp. 153-162.

3C. Martínez, S., Valls, A., Sánchez, D.: Anonymizing Categorical Data with a Recoding Method Based on Semantic Similarity. In: Hüllermeier, E., Kruse, R., Hoffmann, F. (eds.) Information Processing and Management of Uncertainty in Knowledge-Based Systems. Applications, vol. 81. Communications in Computer and Information Science, pp. 602-611. Springer Berlin Heidelberg, (2010). *CORE C*

A new microaggregation method for categorical data. We extended a well-known microaggregation method to enable a semantic anonymisation of categorical data. This method has been published in the following journal:

4J. Martínez, S., Sánchez, D., Valls, A.: Semantic Adaptive Microaggregation of Categorical Microdata. International Journal: Computers & Security 31(5), 653-672 (2012). *Impact Factor: 0.868*

A new resampling method for categorical data. We presented a new method based on the classic data resampling with improved privacy guarantees and utility preservation from a semantic perspective. The method has been presented in the following conference:

4C. Martínez, S., Sánchez, D., Valls, A.: Towards k-anonymous non-numerical data via Semantic Resampling. In: Information Processing and Management of Uncertainty in Knowledge-Based Systems, Catania, Italy 2012, pp. 519-528. S. Greco et al. (Eds.). *CORE C*

A new semantic record linkage method. We proposed a new record linkage method specially tailored to accurately evaluate the disclosure risk of anonymised categorical data. The method has been presented in the follow journal:

5J. Martínez, S., Sánchez, D., valls, A.: Evaluation of the Disclosure Risk of Masking methods Dealing with Textual Attributes. International Journal of Innovative Computing, Information & Control 8(7(A)), 4869-4882 (2012).

Previous work about this topic have been published in the following conference paper:

5C. Martínez, S., Valls, A., Sánchez, D.: An ontology-based record linkage method for textual microdata. In, vol. 232. Frontiers in Artificial Intelligence and Applications, pp. 130-139. (2011).

1.4 Document structure

The present document is divided into the following chapters:

- Chapter 2 details the basic privacy notions and describes related works in SDC, focusing on those dealing with categorical attributes.
- Chapter 3 discusses the notion of knowledge representation and ontologies and studies different semantic similarity measures and quality metrics for the treatment of datasets containing categorical attributes.
- Chapter 4 proposes a new framework consisting on semantically-grounded operators for categorical attributes.
- Chapter 5 presents and evaluates a new recoding method to anonymise categorical attributes.
- Chapter 6 details and evaluates the adaptation of a well-known microaggregation method to anonymise categorical attributes.
- Chapter 7 describes and evaluates a resampling method to anonymise categorical attributes with improve privacy guarantees.
- Chapter 8 presents and evaluates a new semantically-grounded disclosure risk evaluation method.
- Chapter 9 evaluates and compares the proposed methods under a common perspective, using a real medical dataset as case study.
- Chapter 10 depicts the conclusions and presents some lines of future research.

Chapter 2
SDC: Concepts and methods

Inference control in statistical databases or Statistical Disclosure Control (SDC) aims to disseminate statistical data while preserving confidentiality. SDC is focused mainly on the preservation of privacy in structured databases (Willenborg,Waal 2001; Domingo-Ferrer 2008; Jin et al. 2011; Herranz et al. 2010a; Oliveira,Zaïane 2007; Shin et al. 2010). Statistical Disclosure Control methods transform the original database into a new database, taking into account that the protected data satisfies simultaneously utility and security conditions. The dataset will be useful if it is representative of the original dataset and it will be secure if it does not allows the re-identification of the original data.

2.1 Statistical Disclosure Control

There are several areas of application of SDC techniques, which include but are not limited to the following:

Official statistics. Most countries have legislation to guarantee statistical confidentiality when they release data collected from citizens or companies. This justifies the research on SDC (e.g. ESSnet project (European project IST-2000-25069 ESSproject 2009) on the European Union). For example, census data is publicly available in the UCI repository[1] for research purposes.

Health information. This is an important area regarding privacy. E.g., the Privacy Rule of the Health Insurance Portability and Accountability Act in the U. S., (HIPAA (HIPAA 2010)) requires the strict regulation of protected health information for use in medical research. In most others countries, the situation is similar. For example, the California Office of Statewide Health Planning and

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

Development (OSHPD) publish data containing impatient information collected from licensed hospitals in California[2].

E-commerce. The extensive use of electronic commerce generates automatically collection of large amounts of consumer data. This wealth of information is very useful to companies, which are often interested in sharing it with their subsidiaries or partners. Such consumer information transfer should not result in public profiling of individuals and is subject to strict regulation; e.g. in (European_Comission 2012) we can consult regulations in the European Union.

The information confidentiality is guaranteed when it is minimized its disclosure risk. The concepts of confidentiality and disclosure are defined in (F.C.Statistical-Methodology 2005) as follows:

Confidentiality: it assures that the dissemination of data in a manner that would allow public identification of the respondent or would in any way be harmful to him is prohibited, so that the data are immune from legal processes. Confidentiality differs from privacy because it applies to business as well as individuals. Privacy is an individual right whereas confidentiality often applies to data on organizations and firms.

Disclosure: relates to an inappropriate attribution of information to a data subject, whether an individual or an organization. Disclosure occurs when a data subject is identified from a released file (identity disclosure), sensitive information about a data subject is revealed through the released file (attribute disclosure), or the released data make it possible to determine the value of some characteristic of an individual more accurately than otherwise would have been possible (inferential disclosure).

The challenge for SDC is to modify data in such a way that sufficient protection is provided while keeping at a minimum the information loss. The protection provided by SDC techniques normally entails some degree of data modification, which is an intermediate option between no modification (maximum utility, but no disclosure protection) and data encryption (maximum protection but no utility for the user without clearance).

Databases considered in Statistical Disclosure Control can be divided into the following formats (Domingo-Ferrer 2008):

- *Tabular data*: have been the traditional outputs of national statistical offices. The goal here is to publish static aggregate information, i.e. tables, in such a way that no confidential information on specific individuals among those to which the table refers can be inferred.
- *Dynamic databases*: The scenario here is a database to which the user can submit statistical queries (sums, averages, etc.). The aggregate

[2] http://www.oshpd.ca.gov/HID/Products/PatDischargeData/PublicDataSet/index.html

information obtained by a user as a result of successive queries should not allow him to infer information on specific individuals.

- *Microdata*: files where each register corresponds to information of a subject (person or company). It is only recently that data collectors (statistical agencies and the like) have been persuaded to publish microdata. Therefore, microdata protection is the youngest subdiscipline of Statistical Disclosure Control.

In this work it will refer exclusively to databases of microdata and their concrete protection and masking methods, because their disclosure risk is higher than first two. Tabular data publish aggregated information and its aim is not to contain confidential information that can be inferred. Dynamic databases should also ensure that the successive queries do not allow inferring specific information. But midrodata implies a higher risk of disclosure because as it refers to individual information. Due to this reason, microdata is also the most common data used for data mining, implying that the published information must be also analytically useful.

2.1.1 Statistical disclosure control in microdata

To understand the methods used for controlling the risk of disclosure of an individual's identity, one must know the sources of risk in microdata files. Two main cases are distinguished:

- Existence of attributes with high risk:
  - Some registers of a file can represent subjects with unique features that identify them definitely, for example, uncommon works (actor, judge) very high incomes and others.
  - Many registers of a file can be known to belong to the same cluster, for example, family or college.
  - A data dimension is published by a detail level too fine, for example, the publication of the zipcode.
- Possibility of coincidence of a microdata file with extern files: there are some persons or firms that have a unique combination of their attributes. Intruders could use extern files with the same attributes and identifiers to link the unique subjects with their file registers of original microdata.

Otherwise, there are various circumstances that positively affect the disclosure prevention:

- Age of data of the microdata file: The individual and firms features may change significantly over time. The age of the extern files with which one tries link the original file may not match with the original.

- Noise in the information of the microdata file and extern files.
- Different definition of variables of microdata file and extern files.
- Other factors: time, effort and economic resources.

Since the purpose of SDC is to prevent confidential information from being linked to specific respondents, we will assume in what follows that original microdata sets to be protected have been pre-processed to remove from them all identifiers.

Goal of microdata disclosure control:

Given an original microdata set D with n records (corresponding n individuals) and m values in each record (corresponding to m attributes that are not identifiers), the goal is to release a protected microdata set D^A (with also n records an m attributes) in such a way that:

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

Notice, that the posterior use of the data plays an important role in the anonymisation process because the masked version must permit to extract the same knowledge than the original one. With respect to Artificial intelligence techniques, this is important specially if data mining analysis must be done in this data, such as clustering, rules induction, profiling, or prediction, among others. In fact, privacy preserving data mining is a new research field that attempts to develop tools to study in an integrated way how to deal with privacy issues while performing data analysis (Aggarwal,Yu 2008).

Microdata protection methods can generate the protected microdata set D^A:

- Either by masking original data, i.e. generating D^A a modified version of the original microdata set D;
- Or by generating synthetic data D^A that preserve some statistical properties of the original data D.

In this thesis we focus on the first case: **masking methods**.

A masking method converts an original database in a publishable database through a masking process. Fig. 1 show an illustrative example of the masking process, the dataset of the example corresponds to two individuals represented by five attributes. The DNI attribute is identifier and it is removed during the masking process. The attributes birth place, birth year and occupation are not identifiers individually but rare combination of them could identify an individual. For example the combination of judge born in Salou in the year 1968 could

correspond really to a single person. Thus, the masking process should also protect these attributes. In the example, the publishable dataset has been protect changing the original values for more general ones, "Valls", " Salou" for "Tarragona", "1962", "1968" for "1960s" and "lawyer", "judge" for "justice" preventing, in this manner, unique values in the dataset. Finally the income attribute is not changed because it is object of study. Note that the masking process involves some information loss due to the lower specificity of the anonymised values respect to the original ones.

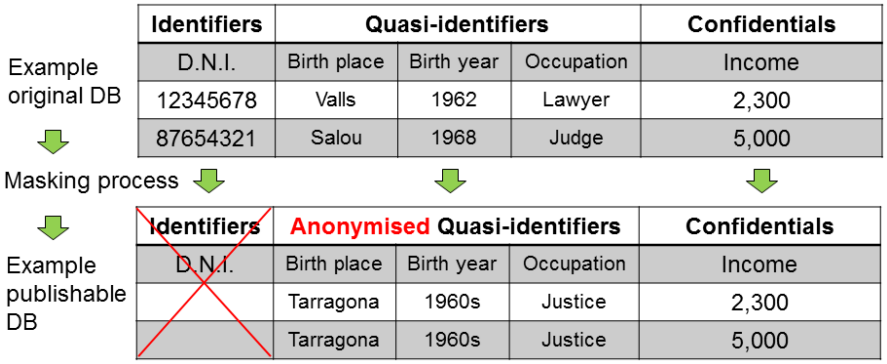

Fig. 1. Example of masking process

Masking methods try to ensure that statistics computed on the anonymised dataset do not differ significantly from the statistics that would be obtained on the original dataset. They can be divided in two categories depending on their effect on the original data (Willenborg,Waal 2001):

- *Perturbative*: data is distorted before publication. The microdata set is distorted before publication. Thus, unique combinations of scores in the original dataset may disappear and new unique combinations may appear in the perturbed dataset; such confusion is beneficial for preserving statistical confidentiality.
- *Non-perturbative*: data values are not altered but generalized or eliminated (Willenborg,Waal 2001; Xu et al. 2006b). The goal is to reduce the detail given by the original data. This can be achieved with the local suppression of certain values or with the publication of a sample of the original data which preserves the anonymity. Recoding by generalization is also another approach, where several categories are combined to form a new and less specific value.

If we consider the type of data on which they can be used, the traditional division distinguishes two types:

- *Numerical*. An attribute is considered numerical if arithmetic operations can be performed with it. Examples are income and age. Note that a

numerical attribute does not necessarily have an infinite range, as is the case for age. When designing methods to protect continuous data, one has the advantage that arithmetic operations are possible, and the drawback that every combination of numerical values in the original dataset is likely to be unique, which leads to disclosure if no action is taken.

- *Categorical*. An attribute is considered categorical when it takes values over a finite set and standard arithmetic operations do not make sense. Ordinal and nominal scales can be distinguished among categorical attributes. In ordinal scales the order between values is relevant, whereas in nominal scales it is not. In the former case, max and min operations are meaningful while in the latter case only pair wise comparison is possible. The instruction level is an example of ordinal attribute, whereas eye colour is an example of nominal attribute. In fact, all sensitive values in a microdata set are normally categorical nominal. When designing methods to protect categorical data, the inability to perform arithmetic operations is certainly inconvenient. Allowing unbounded categorical attributes (e.g. a free answer), the same drawback of the continuous attributes is present i.e. the privacy of the individuals is critical, as the disclosure risk increases due to the uniqueness of the answers. This work is focused on the privacy protection of this type of attributes, specifically on the categorical unbounded attributes.

2.2 *K*-Anonymity

One important type of privacy attacks is re-identifying individuals by joining known values of different attributes taken from multiple public data sources, e.g. according to (Sweeney 2002b), around the 87% of the population of the United States can be uniquely identified using their zipcode, gender and date of birth. Anonymisation methods must mask data in a way that disclosure risk is ensured at an enough level while minimising the loss of accuracy of the data, (i.e. the information loss). A common way to achieve a certain level of privacy is to fulfil the *k*-anonymity property proposed by (Sweeney 2002b; Samarati,Sweeney 1998).

To define the *k*-anonymity concept, previously it is necessary to know the classification of types of attributes that can appear in a dataset, (Domingo-Ferrer 2006) enumerates the various (non-disjoint) possible types of attributes:

- *Identifiers*: the attributes that unambiguously identify the individual, such as the social security number, full name or passport number. To preserve the confidential information, we assume that those attributes must be previously removed or encrypted.

- *Quasi-identifiers*: the attributes that may identify some of the respondents, especially if they are combined with the information provided by other attributes. Unlike identifiers, quasi-identifiers cannot be removed from the dataset because any attribute can potentially be a quasi-identifier.
- *Confidential* outcome attributes: the attributes that contain sensitive information. For example: salary, religion, political affiliation, etc.
- *Non-confidential outcome attributes*: the rest of attributes.

The *k*-anonymity property tries to keep the balance between the information loss and disclosure risk. Once identified the different types of attributes that can appear in a dataset, we can define the ***k*-anonymity property** as (Domingo-Ferrer 2006):

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

Once a value for *k* is fixed (considering a value that keeps the re-identification risk low enough), the goal of the masking method is to find an anonymisation that minimises the information loss.

In order to fulfil the *k*-anonymity property, masking methods have been designed aiming to build groups of *k* indistinguishable records by substituting the original values with a prototype. Obviously, this process results in a loss of information which may compromise the utility of the anonymised data for a further exploitation with data mining techniques. Ideally, the masking method should minimize this loss and maximize data utility according to a certain metric. We can distinguish between global anonymisation methods in which all identifier or quasi identifier attributes are considered and anonymised at the same time (i.e. records will fulfil *k*-anonymity) and local ones in which each attribute is anonymised independently (i.e. each attribute will fulfil *k*-anonymity individually). In the latter case, the information loss of the whole dataset is not optimized because the transformations only have a local view of the problem.

2.3 Perturbative masking methods

These types of masking methods are based on the microdata set are distorted before publication. The anonymisation process may include new data, delete and/or modify the existing data, benefiting the statistic confidentiality.

The main perturbative masking methods are:

- *Additive noise*: add noise with the same correlation structure as the original data. Appropriated method for numerical data. The main noise additions algorithms in the literature are:
    - *Masking with uncorrelated noise addition*: A Register r_i of the original dataset is replaced by a vector $z_i = r_i + \epsilon_i$ where ϵ_i is a vector of normally distributed errors drawn from a random variable $\epsilon_i \sim N(0, \sigma^2_{\epsilon_i})$, such that $Cov(\epsilon_t, \epsilon_l) = 0 \ \forall \ t \neq l$. This does not preserve variances nor correlations.
    - *Masking by correlated noise addition*: preserves means and additionally allows preservation of correlation coefficients. The covariance matrix of the errors is now proportional to the covariance matrix of the original data.
    - *Masking by noise addition and linear transformation*: This method ensures by additional transformations that the sample covariance matrix of the masked attributes is an unbiased estimator for the covariance matrix of the original attributes.
    - *Masking by noise addition and nonlinear transformation*: An algorithm combining simple additive noise and nonlinear transformation. The advantages of this proposal are that it can be applied to discrete attributes and that univariate distributions are preserved. By contrast, the application of this method is very time-consuming and requires expert knowledge on the data set and the algorithm.

  Additive noise is not suitable to protect categorical data. On the other hand, it is well suited for continuous data.
- *Data distortion by probability distribution*: distortion the data with estimated series in function of density of the variables.
- *Microaggregation*: Creates small microclusters, these groups are formed using a criterion of maximal similarity. The size of groups (clusters) must be equal o higher than a variable k to guarantee the confidentiality. For each attribute, the average value over each group is computed and is used to replace each of the original averaged values. Once the procedure has been completed, the resulting (modified) dataset can be published.
- *Re-sampling*: Originally proposed for protecting tabular data, re-sampling can be used for microdata. Take t independent samples $X_1 \ldots X_t$ of the values of an original attributeV_i. Sort the data of each sample. Calculate the average of the first values of each sample. Replace those values by the calculate average. Repeat the process with the $n - 1$values of the next positions.
- *Lossy compression*: consider the dataset as a image and apply compression algorithms (e.g. JPEG)

- *Multiple imputation*: generates a new version of the simulated data created from multiples techniques of imputation from the original data. For example, an imputation method consists on making regressions with a random distribution of the error, to impute "unknown" values to a continuous variable.

- *Camouflage*: camouflage the original information in a range (finite set).it is an appropriate method for numerical data, but causes a high information loss.

- *PRAM (Post-Randomization Method)* (Gouweleeuw et al. 1997): is a perturbative method for privacy protection of categorical attributes in microdata files. In the masked files the original values have been replaced by different information according to a probabilistic mechanism named Markov matrix. The Markov approach makes PRAM very general, because it merges noise addition, data suppression and data recoding. PRAM information loss and disclosure risk depend on the choice of the Markov matrix. The PRAM matrix contains a row for each possible value of each attribute to be protected. This rule excludes this method from being applicable on continuous data.

- *MASSC (Micro Agglomeration, Substitution, Subsampling and Calibration)* (Singh et al. 2003) is a masking method that has four steps:

  1. Micro agglomeration is applied to divide the original dataset into groups of records which are at a similar risk of disclosure. These groups are formed using the key attributes, i.e. the quasi-identifiers in the records. The idea is that those records with rarer combinations of key attributes are at a higher risk.
  2. Optimal probabilistic substitution is then used to perturb the original data.
  3. Optimal probabilistic subsampling is used to suppress some attributes or even entire records.
  4. Optimal sampling weight calibration is used to preserve estimates for outcome attributes in the treated database whose accuracy is critical for the intended data use.

  The method is interesting because it is the first attempt to design a perturbative masking method where disclosure risk can be quantified. In practice MASSC is a method only suited when continuous attributes are not present.

- *Data swapping and rank swapping*: The basic idea is to transform a database by exchanging values of confidential attributes among individual records. Data swapping is originally presented as a SDC method for datasets that contains only categorical data, in (Reiss 1984) data swapping was introduced to protect continuous and categorical microdata. Another variant of data swapping is rank swapping, although

originally described for ordinal attributes, can also be used for numerical attributes.

In the rank swapping method, values of an attribute V_i are ranked in ascending order, then each ranked value of V_i is swapped with another ranked value randomly chosen within a restricted range (e.g. the rank of two swapped values cannot differ by more than $p\%$ of the total number of records, where p is an input parameter).

It is reasonable to expect that multivariate statistics computed from data swapped with this algorithm will be less distorted than those computed after an unconstrained swap

- *Rounding*: replace original values of attributes with rounded values, choosing values that belong to a predefined rounding set, often the multiples of a base value. The rounding method is suitable for numerical data. In a multivariate original dataset, usually, rounding is performed one attribute at time, however, multivariate rounding is also possible.

2.4 Non-perturbative masking methods

This type of masking techniques do not alter the data of the original set but produces partial suppressions or reductions of detail in the original dataset. Some of the methods are suitable for both continuous and categorical data, but others are only usable for categorical data.

The main non-perturbative methods are:

- *Sampling*: (Willenborg,Waal 2001) publish a sample of the original set of records. This methodology is suitable for categorical microdata, for continuous microdata would be necessary combine with others masking methods, otherwise, the disclosure risk is high.
- *Global recoding*: also known as generalization (Samarati,Sweeney 1998). The methodology combines several categories to form new (less specific) categories. For continuous attributes, global recoding means replacing an attribute by its discretized version, but the discretization leads very often to an unaffordable loss of information. This technique is more suitable for categorical attributes, some of these techniques rely on hierarchies of terms covering the categorical values observed in the sample, in order to replace a value by another more general one.
- *Top and bottom coding*: is a special case of global recoding which can be used if the attribute can be ranked, thus, continuous or ordinal. The method determines a threshold for top and bottom values and form new categories with these extreme values.

- *Local suppression*: removes certain values with the aim of increase the set of records agreeing on a combination of key values. In (Samarati,Sweeney 1998) proposes ways to combine local suppression and global recoding. Local suppression is rather oriented to categorical attributes. Local suppression is not always allowed as anonymisation methodology because sometimes the anonymised dataset must have the same number of records as the original dataset.

The following table summarizes and compare all the presented masking methods with respect to the different data type on can be applied:

Table 1. Masking method vs. data types

Method	Type	Continuous data	Categorical data
Additive noise	P	X	
Data distortion by probability distribution	P	X	X
Microaggregation	P	X	X
Re-sampling	P	X	
Lossy compression	P	X	
Multiple imputation	P	X	
Camouflage	P	X	
PRAM	P		X
MASSC	P		X
Data swapping	P	X	X
Rounding	P	X	
Sampling	NP		X
Global recoding	NP	X	X
Top and bottom coding	NP	X	X
Local suppression	NP		X

N: Perturbative NP: Non-Perturbative X: denotes applicable

As shown in Table 1, on categorical data some of the techniques cannot be applied due to the lack of proper operators that permit to quantify those values in some sense.

Notice that the methods for categorical data mainly consider the values an enumerated set of terms, for which only Boolean word matching operations can be performed. On one hand, we can find methods based on data swapping (which exchange values of two different records) and methods that add of some kind of noise (such as the replacement of values according to some probability distribution done in PRAM (Gouweleeuw et al. 1997; Guo,Wu 2009)). On the other hand, other authors (Samarati,Sweeney 1998; Truta,Vinay 2006) perform local suppressions of certain values or select a sample of the original data aimed to fulfil *k*-anonymity property (see section 2.2) while maintaining the information distribution of input data. We can see that the approaches do not make use of

intelligent techniques for dealing with linguistic or textual information, making neither a use of background knowledge to support the anonymisation task.

Even though those methods are effective in achieving a certain degree of privacy in an easy and efficient manner, they fail to preserve the meaning of the original dataset, due to their complete lack of semantic analysis. Some exceptions exist in the set of recoding methods, as it is explained in the next section.

2.5 Categorical data anonymisation

In this section, we review related works on the anonymisation of categorical 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 thesis. We analyse in more detail the methods that incorporate some semantics in the anonymisation process.

2.5.1 Anonymisation of unstructured categorical data

Methods dealing with unstructured textual data usually aim at finding and hiding personal identifiable information in narrative text as part of 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, some sanitisation methods (Wei et al. 2009) based on generalising the sensitive words have been proposed. 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 others that are more general. All the possible generalisations are generated and a suitable combination is selected based on a pruning strategy. 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. An extension of *k*-anonymity is used to evaluate the disclosure risk, denoted *t*-Plausibility. 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 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, 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.

2.5.2 Anonymisation of set-valued data

Another type of textual data can be found collecting 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,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, values that form unique combination 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,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 of *quasi-identifiers* that can lead to the identification of individuals. Attributes are usually single-valued and it is usually assumed independent between them.

2.5.3 Anonymisation of structured databases

As said before in section 2.1, privacy preservation in structured databases is framed in the Statistical Disclosure Control discipline (Willenborg,Waal 2001; Domingo-Ferrer 2008; Jin et al. 2011; Herranz et al. 2010a; Oliveira,Zaïane 2007; Shin et al. 2010). In the literature, the consideration of some kind of semantics on the values of categorical attributes can only be found in *recoding* methods. This section reviews in detail those methods, which constitute a first attempt to incorporate some kind of knowledge background to the masking process as a means to anonymise categorical data.

Most recoding methods (also known as generalisation) rely on hierarchies of terms covering the categorical values observed in the sample, in order to replace a value by a more general one. This replacing mechanism uses the semantics given by those hierarchies of terms to determine which value will be used to make the masking. Therefore, these recoding methods are the most similar ones to our proposal using ontologies.

In some recoding methods, the set of values of each categorical attribute of the input records in the dataset are structured by means of Value Generalization Hierarchies (VGHs). Those are ad-hoc and manually constructed tree-like structures defined according to a given input dataset, where categorical labels of an attribute represent leafs of the hierarchy and they are recursively subsumed by common generalizations. Fig. 2 shows an example of a VGH for an attribute containing information on type of work. In this type of representation, leafs of the tree correspond to the different possible values that the attribute has in the dataset. Note that in this example we have quite few values in the domain of the categorical attribute. This is the usual case of methods applying VGHs.

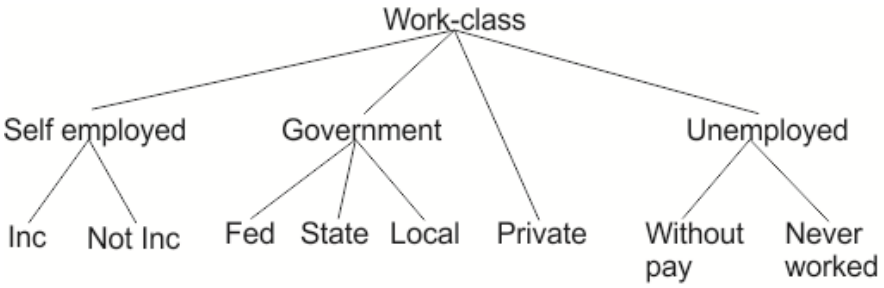

Fig. 2. Example of a VGH for the attribute “type of work”

The recoding masking process consists on, for each attribute, substituting several original values by a more general one, obtained from the hierarchical structure associated to that attribute. This generalization process decreases the number of distinct tuples in the dataset and, in consequence, increases the level of k-anonymity. In general, for each value, different generalizations are possible according to the depth of the tree. The concrete substitution is selected according

to a metric that measures the information loss of each substitution with regards to the original data.

The following summarizes the related work that uses recoding masking methods for categorical data.

- (Samarati,Sweeney 1998; Bayardo,Agrawal 2005; LeFevre et al. 2005) propose a global hierarchical scheme in which all values of each attribute are generalized to the same level of the VGH. The number of valid generalizations for each attribute is the height of the VGH for that attribute. For each attribute, the method picks the minimal generalization which is common to all the record values for that attribute. In this case, the level of generalization is used as a measure of information loss.
- (Iyengar 2002) presented a more flexible scheme that also uses a VGH, where a value of each attribute can be generalized to a different level of the hierarchy in different steps. This scheme allows a much larger space of possible generalizations. Again, for all values and attributes, all the possible generalizations fulfilling the k-anonymity are generated. Then, a genetic algorithm finds the optimum one according to a set of information loss metrics measuring the distributional differences with regards to the original dataset.
- T. Li and N. Li (Li,Li 2008) propose three global generalization schemes:
  - The Set Partitioning Scheme (SPS) represents an unsupervised approach in which each possible partition of the attribute values represents a generalization. This supposes the most flexible generalization scheme but the size of the solution space grows enormously, meanwhile the benefits of a semantically coherent VGH are not exploited.
  - The Guided Set Partitioning Scheme (GSPS) uses a VGH per attribute to restrict the partitions of the corresponding attribute and uses the height of the lowest common ancestor of two values as a metric of semantic distance.
  - The Guided Oriented Partition Scheme (GOPS) adds ordering restrictions to the generalized groups of values to restrict even more the set of possible generalizations.

  Notice that in the three cases all the possible generalizations allowed by the proposed scheme for all attributes are constructed, selecting the one that minimizes the information loss (evaluated by means of a discernibility metric (Bayardo,Agrawal 2005)).
- (He,Naughton 2009) propose a local partitioning algorithm in which generalizations are created for an attribute individually in a Top-Down fashion (recursively). The best combination, according to quality metric (Normalized Certainty Penalty (Terrovitis et al. 2008)), is recursively refined.

- (Xu et al. 2006b) also proposes a local generalization algorithm based on individual attribute utilities. In this case, the method defines different "utility" functions for each attribute, according to their importance. Being local methods, each attribute is anonymised independently, resulting in a more constrained space of generalizations (i.e. it is not necessary to evaluate generalization combinations of all attributes at the same time). However, the optimization of information loss for each attribute independently does not imply that the result obtained is optimum when the whole record is considered. As stated in the introduction, non-necessary generalizations would be typically done in a local method as each attribute should fulfil k-anonymity independently.
- (Bayardo,Agrawal 2005) propose an alternative to the use of VGHs. Their scheme is based on the definition of a total order over all the values of each attribute. According to this order, partitions are created to define different levels of generalization. As a result, the solution space is exponentially large. The problem here is that the definition of a semantically coherent total order for categorical attributes is very difficult and nearly impossible for unbounded textual data. Moreover, the definition of a total order unnecessarily imposes constraints on the space of valid generalizations.

In order to compare the related work, Table 2 summarizes their main characteristics, regarding to the type of anonymisation, the use of different knowledge structures to guide the process of masking, the global vs. local approach, the metric used to measure the quality of the result (more details about these measures are given in the next section) and, finally, the algorithmic search scheme.

Table 2. Related work comparison

Work	Anonymisation method	Background knowledge	Global/ Local	Quality metric	Algorithm type
Samarati & Sweeney (Samarati,Sweeney 1998)	generalization & suppression	small ad-hoc VGH	global	no	Exhaustive
Bayardo & Agrawal (Bayardo,Agrawal 2005)	generalization & suppression	small ad-hoc VGH	global	DM	Heuristic
Lefreve, DeWitt & Ramakrishnan (LeFevre et al. 2006)	generalization	small ad-hoc VGH	global	DM	Exhaustive
Iyengar (Iyengar 2002)	generalization	small ad-hoc VGH	global	LM	Genetic algorithm
Li & Li SPS (Li,Li 2008)	generalization	Partition	global	DM	Heuristics
Li & Li GSPS (Li,Li 2008)	generalization	small ad-hoc VGH	global	DM	Heuristics
Li & Li GOPS (Li,Li 2008)	generalization	small ad-hoc VGH	global	DM	Heuristics
He and Naughton (He,Naughton 2009)	generalization	small ad-hoc VGH	local	NCP	Recursive
Xu et al. (Xu et al. 2006b)	generalization	small ad-hoc VGH	local	NCP & DM	Heuristics

All the approaches relying on VGHs present three main drawbacks.

1. VGHs are manually constructed from each attribute value set of the input data (i.e. categorical values directly correspond to leaves in the hierarchy). So, human intervention is needed in order to provide the adequate semantic background in which those algorithms rely. If input data values change, VGHs should be modified accordingly. Even though this fact may be assumable when dealing with reduced sets of categories (e.g. in (Li,Li 2008) a dozen of different values per attribute are considered in average), this hampers the scalability and applicability of the approaches, especially when dealing with unbounded categorical data (with potentially hundreds or thousands of individual answers).
2. On the other hand, the fact that VGHs are constructed from input data (which represents a limited sample of the underlying domain of knowledge), produces ad-hoc and small hierarchies with a much reduced

taxonomical detail. It is common to observe VGHs with three or four levels of hierarchical depth whereas a detailed taxonomy (such as WordNet) models up to 16 levels (Fellbaum 1998) (see section 3.1.1). From a semantic point of view, VGHs offer a rough and biased knowledge model compared to fine grained and widely accepted ontologies. As a result, the space for valid generalizations that a VGH offers would be much smaller than when exploiting an ontology.

3. Due to the coarse granularity of VGHs, it is likely to suffer from high information loss due to generalizations. As stated above, some authors try to overcome this problem by trying all the possible generalizations exhaustively, but this introduces a considerable computational burden and lacks of a proper semantic background. Therefore, the quality of the results heavily depends on the structure of VGHs that, due to their limited scope, offer a partial and biased view of each attribute domain.

2.6 Quality metrics for anonymised categorical databases

A widely used way to anonymise a dataset and achieving a certain level of privacy is to fulfil the *k*-anonymity property, once a value *k* that keeps the re-identification risk low enough is selected, the main objective is to *k*-anonymise with the least information loss as possible. When anonymising categorical data, particularly using a recoding masking method, anonymisation incurs information loss when a detailed item is generalized to its more generic super-category. The goal of anonymisation in general is to find a transformation of the original data that satisfies a privacy model while minimizing the information loss and maximizing the utility of the anonymised data. Thus a metric is necessary to measure the quality of the resulting data. The difference between the original dataset and the anonymised dataset measures the quality of the anonymisation. In the literature we can find several ways to measure the quality of the dataset anonymised by a masking method.

The main quality metrics can be grouped as:

- Distributional models: the quality measure only evaluates the distribution of the results groups. These models do not use VGH in the measurement and, therefore, do not incorporate semantic knowledge.
- VGH-based models: the quality measure evaluates the information loss as a function of distance calculated on the VGH. These models incorporate additional semantic knowledge in the measure.

On the distributional model, the quality of masked non-numerical data is only considered by preserving the distribution of the input data. On the other hand, the VGH-based model incorporates a poor semantic knowledge base for the

calculation of the quality, and they are not extensible to larges ontologies. Considering their dimensions, the use of ontologies instead of VGHs as semantic background for data anonymisation would result in a generalization space which size would be several orders of magnitude bigger. In fact, as most of the related works make generalizations in an exhaustive fashion, the generalization space is exponentially large according to the depth of the hierarchy, the branching factor, the values and the number of attributes to consider. So, those approaches are computationally too expensive and hardly applicable in a big ontology. Some solutions will be provided in this thesis.

Even though data distribution is a dimension of data utility, we argue, as it has been stated by other authors (Torra 2011) that retaining the semantics of the dataset plays a more important role when one aims to extract conclusions by means of intelligent data analysis. For this reason in chapter Chapter 3 we will propose a new way of measuring the quality of the anonymisation, using a knowledge-based approach.

2.7 Summary

Statistical Disclosure Control aims to preserve the anonymity of individuals while publishing statistical data. The anonymisation of the data deals with two *a priori* conflicting aspects of information: on the one hand, the minimisation of the disclosure risk and, on the other hand, the maximisation of data utility in order to properly exploit the data.

A common and widely accepted way to achieve certain level of privacy is to fulfil the *k*-anonymity property. Many masking methods have been designed aiming to build groups of *k* indistinguishable records, most of them dealing with numerical attributes. However, applying these methods to categorical attributes is not straightforward because of the limitations on defining appropriate operators to manage this kind of categorical values. Some methods exist in the literature designed to treat categorical data do not make use of intelligent techniques for dealing with linguistic information and they fail to preserve the meaning of the original dataset, due to their complete lack of semantic analysis.

In the literature, the consideration of some kind of semantics on the values of categorical attributes can be found in *recoding* (*generalisation*) methods. Most related works aggregate data using a generalisation approach that relies on tailor-made hierarchical structures by means of Value Generalization Hierarchies (VGHs) that present three main drawbacks:

1. VGHs are manually constructed from each attribute in function of the input data (categorical values correspond to leaves). Human intervention is required and the VGH is only valid for a concrete dataset. This fact may not be assumable when dealing with large sets of categories.

2. VGHs produce ad-hoc and small hierarchies with a much reduced taxonomical detail offering a rough and biased knowledge model.
3. Generalisations based on VGHs involve suffering a high information loss due to their coarse granularity. Moreover, the quality of the results heavily depends on the structure of VGH.

To find a transformation of the original data that satisfies a privacy model it is necessary a quality metric to measure the utility of the anonymised data. The difference between the original dataset and the anonymised dataset measures the quality of the anonymisation.

As it has been stated by other authors (Torra 2011) and we will discuss in the following chapters of this thesis, the meaning of data is an important dimension when analysing the anonymised results to extract useful knowledge since it is required in data mining, decision making and recommendation processes. The complete lack of semantic analysis by the categorical treatment and the scarce and shallow semantic interpretation of the VGHs approaches denote the necessity of defining appropriate semantic operators to adapt well-defined masking algorithms using intelligent techniques to manage and evaluate categorical data from a semantic point of view using a knowledge-based approach, stating the starting point for our research.

Chapter 3
Semantic interpretation of categorical data

Semantic interpretation of categorical attribute values for masking purposes requires the exploitation of some sort of structured knowledge sources which allow a mapping between words and semantically interrelated concepts. As it explained in Section 2.5, some privacy approaches have incorporated some sort of background knowledge during the masking process. However, the lightweight and ad-hoc nature of that knowledge and the shallow semantic processing of data hamper their applicability as a general-purpose solution. Hypothesis of this thesis is that the use of well-defined general purpose semantic structures, as ontologies, will allow a better interpretation of data (Little,Rogova 2009; Kokar et al. 2009). Ontologies are formal and machine readable structures of shared conceptualisations of knowledge domains, expressed by means of semantic relationships. Thanks to initiatives such as the Semantic Web (Ding et al. 2004), many ontologies have been created in the last years, such as general purpose ones or specific domain or task ontologies. In this section we present a review of the most important concepts related to ontologies, as well as semantic similarity measures.

3.1 Ontologies

Ontology, in Information Science, can be defined as a rigorous and exhaustive organization of some knowledge domain that is usually hierarchical and contains all the relevant entities and their relations. It is used to reason about the properties of that domain, and may be used to describe the domain. In this section, the ontological paradigm is formalized, and the knowledge representation possibilities of modern ontological languages are analysed.

In (Guarino 1998) an ontology (O) has been defined as:

$$O = (C, \leq_C, R, \sigma_R, \leq_R, A, \sigma_A, T)$$

,where

- *C, R, A* and *T* represent disjoint sets of concepts, relations, attributes and data types. Concepts (or classes) are sets of real world entities with common features (such as different types of diseases, treatments, actors,

etc.). Relations are binary associations between concepts. There exist inter-concept relations, which are common to any domain (such as hyponymy, meronymy, etc.) and domain-dependant associations (e.g., an Actor performs an Action). Attributes represent quantitative and qualitative features of particular concepts (e.g., the medical code of a Disease), which take values in a given scale defined by the data type (e.g., string, integer, etc.).

- $\leq C$ represents a concept hierarchy or taxonomy for the set C. In this taxonomy, a concept c_1 is a subclass, specialization or subsumed concept of another concept c_2 if and only if every instance of c_1 is also an instance of c_2 (which represent its superclass, generalization or subsumer). Concepts are linked by means of transitive is-a relationships (e.g., if respiratory disease is-a disorder and bronchitis is-a respiratory disease, then it can be inferred that bronchitis is-a disorder). Multiple inheritances (i.e., the fact that a concept may have several hierarchical subsumers) are also supported (for example, dog may be both a subclass of canine and pet).

- $\leq R$ which represents a hierarchy of relations (e.g., has primary cause may be a specialization of the relation has cause, which indicates the origination of a Disorder).

- $\sigma R: R \rightarrow C+$ refers to the signatures of the relations, defining which concepts are involved in one specific relation of the set R. It is worth noting that some of the concepts in $C+$ correspond to the domain (the origin of the relation) and the rest to the range (the destination of the relation). Those relationships may fulfil axioms such as functionality, symmetry, transitivity or being the inverse to another one. Relations between concepts are also called object properties.

- $\sigma A: A \rightarrow C \times T$ represents the signature describing an attribute of a certain concept C, which takes values of a certain data type T (e.g., the number of leukocytes attribute of the concept Blood Analysis, which must be an integer value). Attributes are also called data type properties.

Additionally, an ontology can be populated by instantiating concepts with real world entities (e.g., St. Eligius is an instance of the concept Hospital). Those are called instances.

By default, concepts may represent overlapping sets of real entities (i.e., an individual may be an instance of several concepts, for example a concrete disease may be both a Disorder and a Cause of another pathology). If necessary, ontology languages permit specifying that two or more concepts are disjoint (i.e., individuals cannot be instances of more than one of those concepts).

Ontologies are machine-interpretable so that it can be queried. In this sense, some standard languages have been designed to codify ontologies. They are usually declarative languages based on either first-order logic or on description

logic. Some examples are KIF, RDF (Resource Description Framework), KL-ONE, DAML+OIL and OWL (Web Ontology Language) (Gomez-Perez et al. 2004). The most used are OWL (Bechhofer et al. 2009) and RDF (Lassila et al. 2000).

Ontologies can be classified in several forms. A classification was proposed by Guarino (Guarino 1998), who classified types of ontologies according to their level of dependence on a particular task or point of view:

- *Top-level ontologies*: describe general concepts like space, time, event, which are independent of a particular problem or domain.
- *Domain-ontologies*: describe the vocabulary related to a generic domain by specialising the concepts introduced in the top-level ontology. There are a lot of examples of this type of ontologies in e-commerce UNSPSC, NAICS, I biomedicine SNOMED CT, MESH, etc.
- *Task ontologies*: describe the vocabulary related to a generic task or activity by specialising the top-level ontologies.
- *Application ontologies*: they are the most specific ones. Concepts often correspond to roles played by domain entities. They have a limited reusability as they depend on the particular scope and requirements of a specific application. Those ontologies are typically developed ad-hoc by the application designers (Batet et al. 2010; Valls et al. 2010).

Domain ontologies, on one hand, are general enough to be required for achieving consensus between a wide community of users or domain experts and, on the other hand, they are concrete enough to present an enormous diversity with many different and dynamic domains of knowledge and millions of possible concepts to model. Being machine readable, they represent a very reliable and structured knowledge source.

Thanks to initiatives such as the Semantic Web, which brought the creation of thousands of domain ontologies (Ding et al. 2004), ontologies have been extensively exploited to compute semantic likeness. In the following, well-known ontologies that have been widely used for research are introduced.

3.1.1 WordNet

Nowadays, there exist massive and general purpose ontologies like WordNet (Fellbaum 1998). WordNet is a general purpose semantic electronic repository for the English language. It is the most commonly used online lexical and semantic database. In more detail it offers a lexicon, a thesaurus and semantic linkage between the major part of English terms. WordNet distinguishes between nouns, verbs, adjectives and adverbs because they follow different grammatical rules. It seeks to classify words into many categories and to interrelate the meanings of those words. It groups English words into sets of synonyms called synsets,

provides short, general definitions. A synset is a set of words that are interchangeable in some context, because they share a commonly-agreed upon meaning with little or without variation. Each word in WordNet has a pointer to at least one synset. Each synset, in turn, must point to at least one word. It is useful to think of synsets as nodes in a graph. A semantic pointer is simply a directed edge in the graph whose nodes are synsets.

These relations between synsets vary based on the type of word, In the case of nouns, the main relation types includes:

- Hyponym: X is a hyponym of Y if X is a (kind of) Y (dog is a hyponym of canine).
- Hypernym: X is a hypernym of Y if Y is a (kind of) X (canine is a hypernym of dog).
- Holonym: X is a holonym of Y if Y is a part of X (building is a holonym of window).
- Meronym: X is a meronym of Y if X is a part of Y (window is a meronym of building).
- Coordinate terms: Y is a coordinate term of X if X and Y share a hypernym (wolf is a coordinate term of dog, and dog is a coordinate term of wolf)

Each synset also contains a description of its meaning that is expressed in natural language as a gloss. Some example sentences of typical usage of that synset are also given.

The Table 3 summarizes the WordNet 2.1 database statistics (number of words, synsets and senses

Table 3. WordNet 2.1 database statistics

POS	Unique Strings	Synsets	Total Word-Sense Pairs
Noun	117.097	81.426	145.104
Verb	11.488	13.650	24.890
Adjective	22.141	18.877	31.302
Adverb	4.601	3.644	5.720
Totals	155.327	117.597	207.016

The result is a network of meaningfully related words, where the graph model can be exploited to interpret concept's semantics. Hypernymy is, by far, the most common relation, representing more than an 80% of all the modelled semantic links. The maximum depth of the noun hierarchy is 16. Polysemous words present an average of 2.77 synsets (i.e. they belong to almost three different hierarchies).

3.1.2 SNOMED CT

An example of domain ontology that describes the vocabulary focused in a concrete domain of knowledge is the SNOMED CT ontology developed for medical purposes.

SNOMED CT (Systematized Nomenclature of Medicine, Clinical Terms) (Spackman et al. 1997) is the largest structured lexicon of those distributed in the UMLS repository used for indexing electronic medical records, ICU monitoring, clinical decision support, medical research studies, clinical trials, computerised physician order entry, disease surveillance, image indexing and consumer health information services. It contains more than 311,000 concepts with unique meanings and formal logic-based definitions organised into 18 overlapping hierarchies: clinical findings, procedures, observable entities, body structures, organisms, substances, pharmaceutical products, specimens, physical forces, physical objects, events, geographical environments, social contexts, linkage concepts, qualifier values, special concepts, record artifacts and staging and scales. Each concept may belong to one or more of these hierarchies by multiple inheritance (e.g., euthanasia is an event and a procedure), or it may inherit from multiple concepts within one of these hierarchies. Concepts are linked with approximately 1.36 million relationships. Its size and fine-grained taxonomical detail make it especially suitable to assist semantic similarity assessments (Batet et al. 2011; Sanchez,Batet 2011; Pedersen et al. 2007).

3.2 Ontology-based semantic similarity

Most masking process requires the evaluation of data attributes in order to detect the degree of alikeness between values. On the contrary to numerical data, which can be directly and easily manipulated and compared by means of classical mathematical operators, the processing of categorical data is more difficult. Words are labels referring to concepts, which define their semantics. Semantic similarity is precisely the science that aims to estimate the alikeness between words or concepts by discovering, evaluating and exploiting their semantics. Due to semantics is an inherently human feature, methods to automatically calculate semantic similarity relies on evidences retrieved from one or several manually constructed knowledge sources (for example, from ontologies). The goal is to mimic human judgments of similarity by exploiting implicit or explicit semantic evidences. Works in other fields use semantic similarity theory to interpret better the meaning of concepts (Sanchez,Moreno 2008a, b; Sánchez 2010).

In general, the assessment of concept's similarity is based on the estimation of semantic evidence observed in a knowledge resource (Sánchez et al. 2012b; Sánchez,Isern 2011). So, background knowledge is needed in order to measure the degree of similarity between concepts. From the similarity point of view,

taxonomies and, more generally, ontologies, provide a graph model in which semantic interrelations are modelled as links between concepts. Many approaches have been developed to exploit this geometrical model, computing concept similarity as inter-link distance.

In order to guide the anonymisation process towards the transformation that would result in the minimum information loss, a similarity measure that evaluates the semantic difference between the original data and the data resulting from each transformation is needed. To determine the most appropriate measure to guide the masking process, it is necessary the study of the different semantic similarity measures.

In the literature, we can distinguish several different approaches to compute semantic similarity according to the techniques employed and the knowledge exploited to perform the assessment. The most classical approaches exploit structured representations of knowledge as the base to compute similarities.

3.2.1 Edge counting-based measures

By mapping input terms to ontological concepts by means of their textual labels, a straightforward method to calculate the similarity between terms is to evaluate the Path Length connecting their corresponding ontological nodes via is-a links (Rada et al. 1989). As the longest the path, the more semantically far the terms appear to be, this defines a semantic distance measure. The most basic edge counting-based measures are:

- *Path Length* (Rada et al. 1989): in an is-a hierarchy, is the simplest way to estimate the distance between two concepts c_1 and c_2. Consist of calculating the shortest Path Length (i.e. the minimum number of links) connecting c_1 and c_2 concepts (Eq. 3.1).

$$dis_{pL}(c_1, c_2) = \min \#\ of\ is-a\ edges\ connecting\ c_1\ and\ c_2 \qquad (3.1)$$

- *Leacock and Chodorow* (Leacock,Chodorow 1998) also proposed a measure in order to normalize this distance dividing the path length between two concepts (N_p) by the maximum depth of the taxonomy (D) in a non-linear fashion (Eq. 3.2). The function is inverted to measure similarity.

$$sim_{L\&C}(c_1, c_2) = -\log(N_p/2D) \qquad (3.2)$$

- *Wu and Palmer* (Wu,Palmer 1994): However, those measures omit the fact that equally distant concept pairs belonging to an upper level of the taxonomy should be considered as less similar than those belonging to a

lower level, as they present different degrees of generality. Based on this premise Wu and Palmer's measure also takes into account the depth of the concepts in the hierarchy (Eq. 3.3).

$$sim_{W\&P}(c_1, c_2) = \frac{2 \times N_3}{N_1 + N_2 + 2 \times N_3} \quad (3.3)$$

where N_1 and N_2 are the number of is-a links from c_1 and c_2 respectively to their Least Common Subsumer (LCS), and N_3 is the number of is-a links from the LCS to the root of the ontology. It ranges from 1 (for identical concepts) to 0.

The main advantage of the presented measures is their simplicity. They only rely on the geometrical model of an input ontology whose evaluation requires a low computational cost. However, several limitations hamper their performance.

In general, any ontology-based measure would depend on the degree of completeness, homogeneity and coverage of the semantic links represented in the ontology. So, they require rich and consistent ontologies like WordNet to work properly (Pirro,Seco 2008).

A problem of path-based measures typically acknowledged (Bollegala et al. 2007) is that they rely on the notion that all links in the taxonomy represent a uniform distance. Wide ontologies with a relatively homogenous distribution of semantic links and good domain coverage minimize these problems (Jiang,Conrath 1997).

3.2.2 Feature-based measures

On the contrary to edge-counting measures which, as stated above, are based on the notion of path distance (considered in a uniform manner), feature-based approaches assess similarity between concepts as a function of their properties.

By features, authors exploit the information provided by the input ontology. For WordNet, concept synonyms (i.e. synsets, which are sets of linguistically equivalent words), definitions (i.e. glosses, containing textual descriptions of word senses) and different kinds of semantic relationships can be exploited.

- Similarity, in *Tversky* (Tversky 1977) concepts and their neighbours (according to semantic pointers) are represented by synsets. The similarity is computed as (Eq. 3.4):

$$sim_{Tve}(c_1, c_2) = \frac{|A \cap B|}{|A \cap B| + \gamma(c_1, c_2)|A \backslash B| + \big(1 - \gamma(c_1, c_2)\big)|B \backslash A|} \quad (3.4)$$

Where *A, B* are the synsets for concepts corresponding to c_1 and c_2, $A \backslash B$ is the set of terms in *A* but not in *B* and $B \backslash A$ the set of terms in *B* but not in *A*. $\gamma(c_1, c_2)$ is computed a function of the depth of c_1 and c_2 in the taxonomy (Eq. 3.5):

$$\gamma(c_1, c_2) = \begin{cases} \dfrac{depth(c_1)}{depth(c_1) + depth(c_2)}, depth(c_1) \leq depth(c_2) \\ 1 - \dfrac{depth(c_1)}{depth(c_1) + depth(c_2)}, depth(c_1) > depth(c_2) \end{cases} \quad (3.5)$$

- *Rodriguez* (Rodriguez,Egenhofer 2003), the similarity is computed as the weighted sum of similarities between synsets, features and neighbour concepts of evaluated terms (Eq. 3.6):

$$sim_{rod}(c_1, c_2) = w \cdot S_{synsets}(c_1, c_2) + u \cdot S_{features}(c_1, c_2) + v \cdot S_{neighborhoods}(c_1, c_2) \quad (3.6)$$

- *Petrakis* (Petrakis et al. 2006) a feature-based function called *X-similarity* relies on matching between synsets and concept's glosses extracted from WordNet (i.e. words extracted by parsing term definitions). They consider that two terms are similar if the synsets of their concepts and the synsets of concepts in their neighbourhood (following is-a and part-of links) and their glosses are lexically similar. The similarity function is expressed as follows (Eq. 3.7):

$$sim_{X-Similarity}(c_1, c_2) = \begin{cases} 1, if\ S_{synsets}(c_1, c_2) > 0 \\ max\left(S_{neighborhods}(c_1, c_2), S_{descriptions}(c_1, c_2)\right), if\ S_{synsets}(c_1, c_2) = 0 \end{cases} \quad (3.7)$$

where $S_{neighborhods}$ is computed as (Eq. 3.8):

$$S_{neighborhods}(c_1, c_2) = \max \frac{|A_i \cap B_i|}{|A_i \cup B_i|} \quad (3.8)$$

where *A* and *B* denote synsets or description sets for term *a* and *b*.

- Recently, in (Sánchez et al. 2012a) it is proposed a new measure that, relying on similar principles as ontology-based approaches, aims to improve edge-counting measures by evaluating additional taxonomic knowledge modelled in ontologies. Instead of basing the assessment only on the length of the minimum path, authors evaluate, in a non-linear way, the number of non-common subsumers between the compared concepts

as an indication of distance. This value is normalised by the complete set of subsumers of both concepts (Eq. 3.9).

$$dis_{logsc}(c_1, c_2) = log_2\left(1 + \frac{|T(c_1) \cup T(c_2)| - |T(c_1) \cap T(c_2)|}{|T(c_1) \cup T(c_2)|}\right) \quad (3.9)$$

where $T(a)$ is the set of taxonomic subsumers of the concept a, including itself. The advantage of this measure is that it implicitly considers all taxonomic paths between concept pairs (which appear due to multiple taxonomic inheritance), while retaining the efficiency and scalability of path-based measures.

Feature-based measures exploit more semantic evidences than edge-counting approaches, evaluating both commonalties and differences of compared concepts. However, by relying on features like glosses or synsets (in addition to taxonomic and non-taxonomic relationships), those measures limit their applicability to ontologies in which this information is available. Another problem is their dependency on weighting parameters that balance the contribution of each feature.

3.2.3 Information Content-based measures

Resnik (Resnik 1995)] proposed measure the quality of an anonymised set calculating and comparing the information content of both original and result sets. Information content (IC) of a concept c is the inverse to its probability of occurrence. IC computation is based on the probability $p(c)$ of encountering a concept c in a given corpus (Eq. 3.10). In this way, infrequent words obtain a higher IC.

$$IC(c) = -\log p(c) \quad (3.10)$$

The main IC-based similarity measures are:

- *Resnik* (Resnik 1995) introduced the idea of computing the similarity between a pair of concepts (c_1, c_2) as the IC of their Least Common Subsumer (LCS), which is the most concrete taxonomical ancestor common c_1 and c_2 in a given ontology (Eq. 3.11). This gives an indication of the amount of information that the two concepts share in common. The more specific he subsumer is (higher IC), the more similar the terms are.

$$sim_{res}(c_1, c_2) = IC\big(LCS(c_1, c_2)\big) \quad (3.11)$$

- *Lin similarity* (Lin 1998) is an extension of Resnik's measure. This measure depends on the relation between the information content of the LCS of two concepts and the sum of the information content of the individual concepts (c_1, c_2), (Eq. 3.12).

$$sim_{lin}(c_1, c_2) = \frac{2 \times sim_{res}(c_1, c_2)}{\big(IC(c_1) + IC(c_2)\big)} \tag{3.12}$$

- *Jiang and Conrath* presented in (Jiang,Conrath 1997) another extension of Resnik's measure that subtract the information content of the LCS from the sum of the information content of the individual content (Eq. 3.13).

$$dis_{jcn}(c_1, c_2) = \big(IC(c_1) + IC(c_2)\big) - 2 \times sim_{res}(c_1, c_2) \tag{3.13}$$

  Note that this function is a dissimilarity measure because the more different the terms are, the higher the difference from their IC to the IC of their LCS will be.

Finally, other approaches aiming to compute semantic alikeness exploiting the notion of concept's Context Vector. They are based on the premise that words are similar if their contexts are similar. In this case, vectors are constructed from the context of words extracted of the text. Then, the semantic relatedness of two concepts c_1 and c_2 is computed as the cosine of the angle between their context vectors (Patwardhan,Pedersen 2006) (Eq. 3.14).

$$rel_{vector}(c_1, c_2) = \frac{\overrightarrow{v_1} \cdot \overrightarrow{v_2}}{|v_1| \cdot |v_2|} \tag{3.14}$$

Where $\overrightarrow{v_1}$ and $\overrightarrow{v_2}$ are the context vectors corresponding to c_1 and c_2 respectively.

Using the information offered by WordNet and the measures seen above, it is possible to compute the similarity between concepts. There have been some initiatives for computing some standard measures that have been widely used by several authors, such as the software WordNet::Similarity (Pedersen et al. 2004).

WordNet and SNOMED CT ontologies are particularly well suited for similarity measures, since they organise nouns into is-a hierarchies and, therefore, they can be adequate to evaluate taxonomics relationships.

3.2.4 Evaluation of semantic similarity measures

After the study of the different semantic similarity measures, it is necessary to evaluate and compare between them, in order to determine and select the most appropriate one to guide the masking process of our method.

An objective evaluation of the accuracy of a semantic similarity function is difficult because the notion of similarity is subjective (Bollegala et al. 2007). In order to enable fair comparisons, several authors created evaluation benchmarks consisting on word pairs whose similarity were assessed by a set of humans. Rubenstein and Goodenough (Rubenstein,Goodenough 1965) defined the first experiment in 1965 in which a group of 51 students, all native English speakers, assessed the similarity of 65 word pairs selected from ordinary English nouns on a scale from 0 (semantically unrelated) to 4 (highly synonymous). Miller and Charles (Miller,Charles 1991) re-created the experiment in 1991 by taking a subset of 30 noun pairs which similarity was reassessed by 38 undergraduate students. The correlation obtained with respect to Rubenstein and Goodenough experiment was 0.97. Resnik (Resnik 1995) replicated again the same experiment in 1995, in this case, requesting 10 computer science graduate students and post-doc researchers to assess similarity. The correlation with respect to Miller and Charles results was 0.96. Finally, Pirró (Pirro,Seco 2008) replicated and compared the three above experiments in 2008, involving 101 human subjects, both English and non-English native speakers. He obtained an average correlation of 0.97. It is interesting to see the high correlation obtained between the experiments even though being performed in a period of more than 40 years and a heterogeneous set of human subjects. This means that similarity between the selected words is stable over the years, making them a reliable source for comparing similarity measures.

In fact, Rubenstein and Goodenough and Miller and Charles benchmarks have become de facto standard tests to evaluate and compare the accuracy of similarity measures. As a result, correlation values obtained against those benchmarks can be used to numerically quantify the closeness of two ratings sets (i.e. the human judgments and the results of the computerized assessment). If the two rating sets are exactly the same, correlations coefficient is 1 whereas 0 means that there is no relation. Correlations coefficients have been commonly used in the literature; both are equivalent is ratings sets are ordered (which is the case). They are also invariant to linear transformations which may be performed over results such as change between distance and similarity (for the corresponding functions) or normalizing values in a range. This enables a fair and objective comparison against different approaches.

So, we have taken the correlation values originally reported by related works for Rubenstein and Goodenough and Miller and Charles benchmarks (when available) and summarized in Table 4Table 4. In case in which a concrete measure depends on certain parameters (such as weights or corpora selection/processing) the best correlation value reported by the authors was compiled. It is important to note that, even though some of them rely on different knowledge sources (such as

tagged corpora or the Web), all ontology-based ones use WordNet. WordNet 2 is the most common version used in related works. In cases in which original authors used an older version (WordNet 2 was released in July 2003), we took a replication of the measure evaluation performed by another author in order to enable a fair comparison. As a result, we picked up results reported by authors in papers published from 2004 to 2009.

Table 4. Correlation values for each measure. From left to right: measure authors, family type, correlation reported for Miller and Charles benchmark, correlation reported for Rubenstein and Goodenough benchmark.

Measure	**Type**	**M&C**	**R&G**	**Evaluated in**
Path (Rada)	Edge	0.59	N/A	(Petrakis et al. 2006)
Wu & Palmer	Edge	0.74	N/A	(Petrakis et al. 2006)
Leacock & Chodorow	Edge	0.74	0.77	(Patwardhan,Pedersen 2006)
Rodriguez	Feature	0.71	N/A	(Petrakis et al. 2006)
Tversky	Feature	0.73	N/A	(Petrakis et al. 2006)
Petrakis	Feature	0.74	N/A	(Petrakis et al. 2006)
LogSC	Feature	0.85	0.86	(Batet et al. 2011)
Resnik	IC	0.72	0.72	(Patwardhan,Pedersen 2006)
Lin	IC	0.7	0.72	(Patwardhan,Pedersen 2006)
Jiang & Conrath	IC	0.73	0.75	(Patwardhan,Pedersen 2006)

Correlation values indicate that measure accuracies are very similar through the different families. However, the applicability and generality of each measure type depend on the principle they exploit.

On one hand, with respect to feature-based type measures, the main problem is their dependency on weighting parameters that balance the contribution of each feature. In all cases, those parameters should be tuned according the nature of the ontology and even to the evaluated terms. This hampers their applicability as a general purpose solution. Only Petrakis(Petrakis et al. 2006) does not depend on weighting parameters, as the maximum similarity provided by each feature alone is taken. Even though this adapts the behaviour of the measure to the characteristics of the ontology and the knowledge modelling, by taking only the maximum value at each time the contribution of other features is omitted.

On the other hand, the I.C. type measures need an accurate computation of concept probabilities that requires a proper disambiguation and annotation of each noun found in the corpus. If either the taxonomy or the corpus changes, re-computations are needed to be recursively executed for the affected concepts. So, it is necessary to perform a manual and time-consuming analysis of corpora and resulting probabilities would depend on the size and nature of input corpora. Moreover, the background taxonomy must be as complete as possible (i.e. it should include most of the specializations of a specific concept) in order to provide reliable results. Partial taxonomies with a limited scope may not be suitable for this purpose. All those aspects limit the scalability and applicability of those approaches.

In this thesis, we have chosen the edge counting-based and featured-based types family as the similarity measure type for testing purposes because, as they neither depend on corpora nor tuning parameters they present a low computational cost and lack of constraints. This ensures their applicability and generality especially when dealing with large sets of data, which is common when anonymising data.

3.3 Summary

Semantic interpretation of categorical data requires knowledge sources to map concepts aiming to extract their semantic information. The main hypothesis of this thesis is that the use of domain ontologies will allow a better interpretation of categorical data to guide the masking process. WordNet and SNOMED CT are examples of domain ontologies widely used in the literature.

Masking process requires the evaluation of data. Numerical data can be easily manipulated and compared with mathematical operators. On the contrary, the processing of categorical data is not simple because the limitations on defining appropriate operators for categorical values. Semantic similarity science aims to estimate the alikeness between concepts. In this chapter, we have provided a review of the semantic similarity measures that can be used to estimate the resemblance between concepts.

Through the exploitation of knowledge offer by ontologies and the semantic measures reviewed in this chapter, it is possible to compute the similarity/distance between concepts in order to develop appropriate operators to guide the anonymisation process of categorical attributes.

Chapter 4
A semantic framework for categorical data

As discussed in chapter 2 (section 2.5 and summary), classical SDC methods omit the semantic component of categorical data, focusing solely on their distributional properties. In this section, we present a general framework that integrates both features of non-numerical data. The framework captures the semantics of data by relying on the measurement of the *semantic similarity* between terms, assessed from medical knowledge bases, while taking into account the frequency of data, as well. In the following, we formalise the problem to solve; then, three semantically-grounded operators to manage categorical data are presented: *comparison*, *aggregation* and *sorting*.

4.1 Problem formalisation

A structured database consists of n records corresponding to individuals, each one containing m attribute values.

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

Example 1. Given the univariate dataset V stating patient diagnosis, {*asbestosis*, *degenerative disorder*, *amyotrophia*, *myofibrosis*, *asbestosis*, *allergy*, *myofibrosis*, *allergy*, *squint*, *amyotrophia*, *degenerative disorder*, *allergy*}, we represent it as:

V = {<*asbestosis*, 2>, <*degenerative disorder*, 2>, <*amyotrophia*, 2>, <*myofibrosis*, 2>, <*allergy*, 3>, <*squint*, 1>}

Note that, in this example, m=1, n=12 and p=6.

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

Example 2. Given the multivariate dataset *MV* with two attributes describing conditions and treatments of a set of patients, {{*lumbago*, *rehabilitation*}, {*colic*, *antibiotic*}, {*lumbago*, *rehabilitation*}, {*migraine*, *aspirin*}, {*lumbago*, *rehabilitation*}, {*lumbago*, *codeine*}, {*colic*, *hospitalisation*}, {*lumbago*, *codeine*}}, we represent it as:

MV = {<{*lumbago*, *rehabilitation*}, 3>, <{*colic*, *antibiotic*}, 1>, <{*migraine*, *aspirin*}, 1>, <{*lumbago*, *codeine*}, 2>, <{*colic*, *hospitalisation*}, 1>}

In this example, $m=2$, $n=8$ and $p=5$.

Considering that the goal of many SDC methods (like those introduced in section 2.5) is to fulfil the *k*-anonymity property, if ω_i is equal or greater than a given value of *k,* the corresponding records in this tuple are already *k*-anonymous since they fulfil desired level of privacy. Hence, the goal of the anonymisation process consists of generating a dataset where $\omega_i \geq k, \ \forall i$.

Treating input data as formalised above, different value tuples with frequencies of appearance, results in an improvement of computation cost. This is due to, by definition, $p \leq n$ and, considering that categorical data is characterised by a limited and usually reduced set of modalities, is very common that $p << n$.

4.2 Comparison operator

As introduced in section 2.2, to achieve *k*-anonymous datasets, records should be put together (i.e., grouping) and replaced by an average value, that is, aggregation, so that they become indistinguishable. To minimise the information loss produced by the aggregation step, the *most similar* (or *least distant*) records should be put together to obtain cohesive groups and averages.

To group similar records, a *comparison* operator is needed. Groups are built around a base value *b,* whose selection depends on the anonymisation method, to which the most similar records of the dataset are joined. Hence, the comparison operator should be able to rank the set of records in the dataset according to their *distance* with the base value *b*.

To consider the term semantics during this comparison, we rely on the notion of *semantic similarity*, which quantifies the taxonomical resemblance of compared terms based on semantic evidences extracted from a knowledge base (Batet et al. 2011; Sanchez,Batet 2011; Pedersen et al. 2007). As introduced in section 3.2, to evaluate terms, we use a structured terminology like WordNet or SNOMED CT as

the knowledge base. They offer a taxonomic structure in which subsumption relations are modelled as links between terms. Based on the results of the comparison of correlation values shown in section 3.2.4 we used the dis_{LogSC} measure (Sánchez et al. 2012a; Batet et al. 2011), (see Eq. 3.9) which quantifies the *semantic distance* between term pairs $sd(v_1, v_2)$.

An advantage of this measure is that it evaluates *all* the taxonomical ancestors of the evaluated terms, considering also multiple taxonomical inheritance, which are very common in medical taxonomies (Batet et al. 2011). As a result, it showed improved accuracy over related works in several medical (Batet et al. 2011) and general purpose (Sánchez et al. 2012a) benchmarks.

We denote $sd(v_1, v_2)$ to any *semantic distance* (see section 3.2) for the pair of concepts (v_1, v_2) measured using the ontology O.

Example 3. As an illustrative example of the semantic distance measure, let us consider a univariate dataset where the attribute refers to diseases: V_1 = {*asbestosis*, *amyotrophia*, *myofibrosis*, *allergy*, *degenerative disorder*, *squint*}. Fig. 3 shows an extract of the taxonomy modelling these diseases in SNOMED CT. Applying Eq. 4.1 to all the possible pairs of terms we obtain the semantic distance values shown in Table 5. We can see, for example, that "sibling" terms like *amyotrophia* and *myofibrosis* are less distant than *allergy* and *squint*, because the former are more specific (i.e. they are located lower in the taxonomy) and, hence, they share more subsumers.

Any of the semantic similarity functions described in section 3.2 may be suitable to measure the distance between terms using an ontology. In this example we consider SNOMED CT as ontology and the *semantic distance* between term pairs $sd(v_1, v_2)$ as:

$$sd(v_1, v_2) = dis_{LogSC}(v_1, v_2) \tag{4.1}$$

where dis_{LogSC} is the measure described in Eq. 3.9.

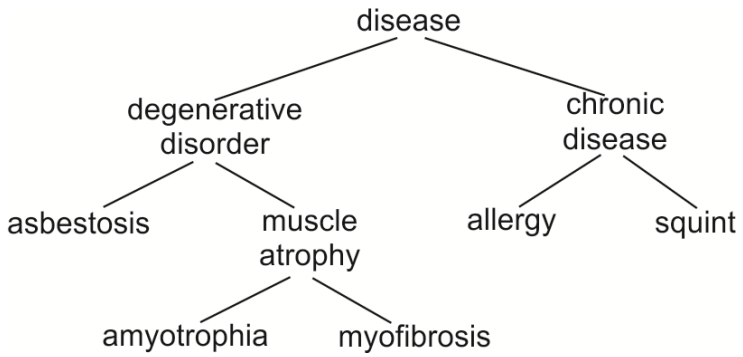

Fig. 3. The subsumer hierarchy for the set V_1, extracted from SNOMED CT.

Table 5. Semantic distance between term pairs of Example 2, according to the SNOMED CT taxonomy extract shown in Fig. 3.

Semantic distance	asbestosis	amyotrophia	myofibrosis	allergy	squint	degenerative disorder
asbestosis	0	0.68	0.68	0.85	0.85	0.42
amyotrophia	0.68	0	0.49	0.87	0.87	0.58
myofibrosis	0.68	0.49	0	0.87	0.87	0.58
allergy	0.85	0.87	0.87	0	0.58	0.81
squint	0.85	0.87	0.87	0.58	0	0.81
degenerative disorder	0.42	0.58	0.58	0.81	0.81	0

Applying this measure, we can semantically compare, rank and group the most similar (i.e., lest distant) record values v_i in a dataset with respect to a base value b from which each group is built, using $sd(b, v_i)$.

To consider also the distribution of data during the comparison of attribute values, their frequency of appearance is also taken into account. Since each distinct value v_i appears ω_i times in the dataset (as formalised in section 2.1), we propose to count the semantic distance between a given value v_i and the base value b as many times as indicated by its frequency of appearance ω_i. Since this is equivalent to multiplying the semantic distance by the frequency, ω_i acts as a weighting factor. In that way, the accumulated distances resulting from grouping together b and all the records with the value v_i can be minimised. Moreover, since all repetitions ω_i of v_i are treated as a unit, sets of identical records will be grouped together, obtaining more cohesive groups.

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

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

$$wsd(b, < v_i, \omega_i >) = \omega \cdot sd(b, v_i) \tag{4.2}$$

This measure can be generalised to multivariate data as follows:

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

$$wsd(\{b_1, \ldots, b_m\}, < \{v_{i1}, \ldots v_{im}\}, \omega_i >) = \sum_{j=1}^{m} \frac{wsd(b_j, < v_{ij}, \omega_i >)}{m} \tag{4.3}$$

4.3 Aggregation operator

Aggregation refers to the process of combining several values into a single one, so that the final result is a value that summarises all individual values. The result of the aggregation process is understood as the prototype or *centroid* of a set of values.

Data protection is another relevant field (closely related to database analysis) in which centroids are useful (Domingo-Ferrer 2008). In order to preserve the privacy of individuals and institutions, databases must guarantee a certain level of privacy before being made accessible to third parties (Martinez et al. 2011; Martinez et al. 2010a; Martinez et al. 2010b). Privacy-preserving data mining algorithms use masking methods applying data transformations to hide the original data values (Domingo-Ferrer 2008). In particular, microaggregation methods have proven to be very effective in keeping low the information loss inherent to data transformations while maintaining an acceptable disclosure risk (Domingo-Ferrer,Mateo-Sanz 2002). In this approach, individual records are grouped into small clusters prior to publication. Then, a centroid for each cluster is calculated and published instead of the original values. Ideally, the group partition and the centroid should be constructed so that the information loss of the masked data with respect to the original dataset is minimised. Moreover, to prevent the disclosure of individual information, it is required that each cluster contains at least *k* records (fulfilling the *k*-anonymity property (Samarati,Sweeney 1998)). Notice that this method is quite similar to clustering algorithms, with an additional requirement regarding the minimum size of the clusters. In fact, some variations of data mining clustering methods are commonly used, such as the *k-Ward*

microaggregation (Domingo-Ferrer,Mateo-Sanz 2002) or the *MDAV* method (which is similar to *k-means* clustering) (Domingo-Ferrer,Torra 2005).

Many clustering algorithms rely on a *central* value (called centroid or prototype) that best *represents* a dataset or cluster. In general, a centroid is defined as a typical example, basis, or standard for other elements of the same group (Han 2005). In practice, a centroid of a dataset is a vector that encodes, for each attribute, the most representative value of the objects found in the dataset. Centroids are needed in methods such as the *Ward*'s unsupervised agglomerative clustering, in which the centroid is used to measure the inertia intra and inter clusters (Ward 1963) (inertia is a function of the distance between the objects and the centroids). Similarly, the widely-used clustering algorithm *k-means* (based on iteratively partitioning a dataset), uses centroids to create new clusters, which are adjusted at each iteration step (Everitt et al. 2011). This centroid-based partitioning approach is especially interesting for data mining applications due to their efficiency in processing large data sets (Huang 1998).

In addition to clustering methods, centroids are also required in other tasks such as: classification of new objects into already defined clusters (Huang et al. 2010b), detection of outliers (Shin et al. 2006), parameter selection (Maniruzzaman et al. 2002), text clustering (Zhang et al. 2010) or data classification and processing (Keikha et al. 2009; Wong,Hamouda 2003; Yihui 2012).

Formally, the centroid of a dataset or cluster is a representative value (or a tuple of values in multivariate datasets) that *minimises the distance* to all other elements in the set (Han 2005). In accordance with this definition, geometrically, the centroid of a group is an object located at the *centre* of the group. To compute the centroid, a simple approach consists of selecting one of the objects of the cluster so that it minimises the distance with respect to all the members of the cluster. This kind of centroid is called Medoid (Park,Jun 2009). More flexible approaches are based on constructing a new synthetic centroid using an averaging function, which is applied to all the attribute values of the objects in the dataset. In both cases, distance functions are needed.

The case of centroid construction for numerical databases has been widely studied, being the *mean* the most common averaging operator (e.g., Arithmetic, Geometric, Harmonic mean or other types described in (Torra,Narukawa 2007)). As distance measure, the Euclidean is the most common one (Domingo-Ferrer,Mateo-Sanz 2002; Domingo-Ferrer,Torra 2005).

Centroid calculus for numerical data relies on standard averaging operators (e.g. arithmetic mean)(Domingo-Ferrer 2008). However, the accurate centroid calculus for categorical data is challenging due to the lack of semantic aggregation operators and the necessity of considering a discrete set of centroid candidates, that is, it should be a value taken from a finite set of modalities, instead of a continuous scale numerical value. Related works propose methods to compute centroids for categorical data either rely on the distributional features of data, where the centroid is the modal value (Torra 2004), or on background semantic, where the centroid is the term that generalises all aggregated values in a

background taxonomy (Abril et al. 2010). Since only one dimension of data (distribution or semantics) is considered, both approaches result in suboptimal results (Martínez et al. 2012b).

In this section, we propose a centroid calculation method for multivariate categorical data that considers, in an integrated manner, both semantics and distribution of data. In addition, On the contrary to most related works (discussed in section 3.3.4), which bound the centroid to values already contained in the input dataset (Domingo-Ferrer,Torra 2005; Guzman-Arenas et al. 2011; Torra 2004), we have put special efforts in maximising the flexibility of the centroid construction. In our approach, we exploit, as much as possible, the knowledge provided by the background ontology, resulting in finer grained and more accurate centroids. The background knowledge base is exploited not only to semantically compare terms, as proposed in the previous section, but also to retrieve the centroid candidates.

In a nutshell, first, we present a method to semantically construct centroids for univariate data, which provides an optimal minimisation of the semantic distance between the input dataset and the centroid, according to the background data. Then, the method is generalised to multivariate data, ensuring the scalability when dealing with large and high dimensional multivariate datasets. Our method has been evaluated and compared against related works (being semantically-grounded or not) using real data containing categorical attributes. The dataset has been evaluated as a whole and also in a clustered fashion. Results show that our method better retains the semantic of data, minimising the semantic distance between the obtained centroids and the input data.

4.3.1 Approaches on centroid calculus for categorical values

When dealing with non-numerical data, we can distinguish two types of methods to construct centroids: those analysing data in a categorical fashion, which compare textual values according to their labels, and knowledge-based ones, which exploit a knowledge source to interpret data semantics.

Works considering textual data as categorical terms are based on Boolean (equality/inequality) operations applied to textual labels. In (Varde et al. 2006), it is proposed an approach called DesCond to extract a centroid for clusters of scientific input conditions. The centroid is selected from each cluster as a single object (in this case, this refers to all input conditions in a given experiment) such that it is the nearest neighbour to all other objects in the cluster. For this, the centroid is such value in the cluster that the sum of its distances to the rest of values of the cluster is minimal. Because textual attributes are considered as categorical, the distance is defined as 0 if the attribute values are identical and 1 otherwise (Cao et al. 2012; Bai et al. 2011; Domingo-Ferrer,Torra 2005; Torra 2004). In (Torra 2004; Domingo-Ferrer,Torra 2005) authors propose a method for categorical microaggregation of confidential data (i.e., records with values linked

to a particular individual) in order to ensure the privacy of individuals before its publication. The microaggregated groups of records are substituted at the end of the algorithm by the centroid of the group. The centroid of textual attributes is selected as the value that most frequently occurs in the group (i.e., mode). In (Erola et al. 2010) authors also use a microaggregation-based masking method to protect query logs. To group and mask similar queries, it is proposed a clustering algorithm based on finding similarities between queries by exploiting a taxonomy of topics. Then, for each cluster, a centroid consisting of a set of queries replaces all queries in the cluster. Queries in the centroid are selected as those more frequently appearing in the cluster (i.e., mode). In (Greenacre,Hastie 2010), authors use a similar strategy, classifying documents according to the most frequently appearing words. In (Zhang et al. 2010) authors propose a new method for document clustering. To make document clusters comprehensible, they assign the most frequent items in a cluster as the topic of the grouped documents. In (Bai et al. 2011), a new method is proposed to find the initial clusters centres for grouping algorithms dealing with categorical data. Authors select the most frequent attribute value (mode) as the cluster representative. In (Cao et al. 2012) it is proposed a dissimilarity measure for clustering categorical objects. Again, the mode is used as the criterion to select cluster representatives. In (Huang et al. 2010b) authors proposed a supervised classification algorithm based on labelled training labels and local cluster centres. In order to avoid the interference of mislabelled data, authors select cluster centres so that they reflect the distribution of data (i.e. most frequent labels).

Due to the categorical treatment of textual data, the above-mentioned approaches can only evaluate data values as identical (i.e. maximum similarity) or not (i.e., minimum similarity). This naïve criterion can be improved by semantically analysing the data and, hence, being able to quantify more accurately the similarity between values. As stated in the introduction, this requires exploiting background knowledge to interpret semantics.

In recent years, some authors started using knowledge sources to assist the construction of centroids. In (Abril et al. 2010) authors use the WordNet structured thesaurus (Pedersen et al. 2004) as ontology to assist the classification and masking of confidential textual documents. WordNet models and semantically interlinks more than 100,000 concepts referred by means of English textual labels. Authors exploit WordNet both to assist the classification process, in which relevant words are extracted from text and those are grouped according to the similarity of their meaning, and to select a centroid for each obtained cluster, which is used to mask confidential text. The Wu and Palmer's similarity measure (Wu,Palmer 1994) is used to estimate the semantic alikeness between words by mapping them to WordNet concepts and computing the number of semantic links separating them. As a result, terms are clusterised according to their semantic similarity. The centroid of the resulting clusters, however, is selected as the *Least Common Subsumer (LCS)*, which is the most concrete taxonomical ancestor found in WordNet for the terms found in the cluster. As a result, the centroid represents the semantic content that all the concepts referred in the cluster have in common. Even though term semantics are considered, the use of the LCS as centroid has

some drawbacks. First, the presence of outliers (i.e., terms referring to concepts which are semantically far to the major part of the other elements in the cluster) will cause that the LCS becomes a very general concept, for example, in the worst case, the root of the taxonomy. The substitution of cluster terms by such as general concept (e.g., entity, thing, abstraction, etc.) implies a high loss of semantic content. Moreover, the number of term repetitions is not considered during the centroid selection and hence, a scarce term will be considered as important as common ones, biasing results. Those issues imply that the use of the LCS as centroid does not minimise the semantic distance to all elements in the cluster (incoherently to the centroid definition), resulting in a sub-optimal semantic loss.

In (Guzman-Arenas et al. 2011; Guzman-Arenas,Jimenez-Contreras 2010) authors propose a generic way to calculate the centroid, called *consensus*, of a set of textual values. Their proposal exploits the knowledge modelled in ad-hoc hierarchies that taxonomically link input values to measure their *confusion* (i.e., a measure of semantic alikeness). Confusion of two values *Conf (a,b)* refers to the confusion derived of using *a* instead of *b*. To measure this confusion, they count the number of descending links in the hierarchy from *a* to *b*, with the particularity that the confusion will be 0 if *a* is a descendent of *b* (i.e., the use of taxonomical subsumers is not penalised). The centroid or consensus of the set of textual values is selected as the term in the set that minimises the total confusion against the rest of elements of the set. This work incorporates better the notion of semantic centroid as a representative that minimises the semantic distance to all the other elements. However, as discussed above, the semantic distance derived from the substitution of a term by its subsumer, for example, the root node, in the worst case, should also be quantified because it implies a noticeable loss of semantic content. Moreover, authors' approach is focused on very simple and overspecified taxonomies that must be constructed ad-hoc for each dataset because they only incorporate the values that appear in the input dataset. Hence, the quality of the results (i.e. the suitability of the selected centroid and the minimisation of the semantic distance) closely depends on the homogeneity, completeness and granularity of input values from the taxonomical point of view.

4.3.2 Ontology-based centroid construction

This thesis proposes a definition of centroid of a categorical set of values as the term or tuple of terms (in case of multivariate data) that minimises the distance against all the elements in a dataset/group. Formally, given a distance function *d* on the space of data, the centroid of a set of values $\{x_1,x_2,...,x_n\}$ is defined as:

$$centroid\,(x_1,x_2,...,x_n) = arg\ min_c \left\{ \sum_{i=1}^{n} d(c,x_i) \right\} \tag{4.4}$$

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

As discussed in chapter 2, most related works dealing with categorical attributes do not incorporate the notion of distance when constructing the centroid. Instead, some heuristics based on frequency analysis or on the notion of concept subsumption are used, assuming that those aid to minimise the distance. Considering the semantic nature of categorical attribute values, we propose computing the distance relying on the estimation of their *semantic similarity* (see section 3.2).

A second relevant difference of our approach compared to related works concerns the search space explored to select the centroid. In general, knowledge-based approaches have a potential advantage with respect to methods dealing with data in a categorical fashion. When selecting the centroid according to the frequency of repetitions (i.e., mode) the search space of possible centroids is limited to the set of different values found in the cluster. As a result, the centroid value is discretised according to input values. In ontology/taxonomy-based approaches, on the contrary, the search space can be potentially extended to all concepts modelled in the ontology (e.g., taxonomical subsumers of terms contained in the cluster) and, hence, the centroid can be synthetically constructed from a finer grained set of candidates. This advantage, however, is either slightly considered in related works (Abril et al. 2010), restricting the centroid to the Least Common Subsumer (LCS), or even not exploited at all (Guzman-Arenas et al. 2011; Guzman-Arenas,Jimenez-Contreras 2010), in cases in which the background taxonomy only incorporates terms found in the input dataset.

In this thesis, we propose:

- To expand the search space and obtain fine-grained centroids, we propose relying on detailed ontologies like WordNet (Fellbaum 1998). We map input labels to concepts in WordNet so that the hierarchical trees where each one is located can be explored to retrieve possible centroid candidates (e.g. complete sets of taxonomical ancestors).
- To use semantic similarity by applying the comparison operator between terms proposed in section 4.2

This strategy will help to minimise even more the semantic distance and propose more accurate centroids.

4.3.3 Construction the centroid for univariate data

In this section, the case of univariate datasets is studied first, to facilitate the definition of concepts and notation involved in our proposal. After that, in section 4.3.4, the method will be generalised to multivariate data.

Let us take an input dataset with a single categorical attribute that can be represented as $V=\{<v_1,\omega_1>,...,<v_p,\omega_p>\}$, where v_i is each distinct categorical label and ω_i is its number of occurrences in the dataset. 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 semantic relationships. In this work, we focus the analysis on taxonomical relationships because those are the most structure-building ones and, hence, the most usually available in ontologies. In fact, an investigation of the structure of existing ontologies via the Swoogle ontology search engine (Ding et al. 2004) has shown that available ontologies very occasionally model non-taxonomic knowledge. By analysing taxonomic relationships we retrieve new concepts (i.e. taxonomical ancestors) that become centroid candidates for the values in V. Following a similar premise as in (Abril et al. 2010; Guzman-Arenas et al. 2011), 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 Subsumer (LCS) of all values. We omitted taxonomical ancestors of the LCS because those will always be more general than the LCS (i.e., more semantically distant) and, hence, worse centroid candidates. Similarly, taxonomical specialisations below the bound defined by the minimum hierarchy are also omitted because they will be also more distant to the other values. Formally:

Definition 3. The *Least Common Subsumer* (LCS) of a set of textual values V in an ontology O ($LCS_O(V)$) is the deepest (i.e., specific) taxonomical ancestor that all terms in V have in common for the ontology O.

Definition 4. 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. Note that in 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 5. The minimum subsumer hierarchy ($H_O(V)$) extracted from the ontology O conrresponding to all values in V is defined as the union of all the concepts in ($S_{LCSO}(v_i)$) for all v_i.

$$H_O(V) = \bigcup_{i=1}^{p} \{S_{LCS_O}(v_i)\} \tag{4.5}$$

where n is the cardinality of V.

Example 4. As illustrative example, consider a univariate dataset of patients, where each record corresponds to an individual and the unique attribute describes his/her condition, so that the following set of textual values can be extracted: V_1 = *{<colic, 1>, <lumbago, 3>, <migraine, 2>, <pain, 1>, <appendicitis, 1>, <gastritis ,1>}*. By mapping these values to concepts found in the background ontology O (WordNet –WN-, for this example) and applying Definition 5, we are able to extract the minimum hierarchy H_{WN}, shown in the Fig. 4.

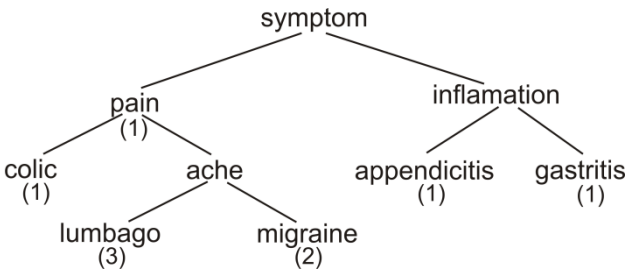

Fig. 4. Minimum subsumer hierarchy H_{WN} for the set V_1, extracted from WordNet. Numbers in parenthesis represent the number of repetitions of each value.

Applying Definition 3, the LCS between a pair of values such as *lumbago* and *migraine* is $LCS_{WN}(\{lumbago, migraine\})=ache$; whereas the LCS of the entire set V_1 is $LCS_{WN}(V_1)=symptom$. Applying Definition 4, the ancestors of a value such as *migraine* are $S_{LCSWN}(migraine)=\{migraine, ache, pain, symptom\}$.

Definition 6. All the concepts c in H_O are the *centroid candidates* for V.

Following Example 4, the centroid candidates of V_1 are those in H_{WN}: *{colic, lumbago, migraine, appendicitis, gastritis, pain, ache, inflammation, symptom}*.

Note that, according to the values in V and the structure of the background ontology, centroid candidates may be taxonomical ancestors of v_i, but also specialisations or siblings are found. In Example 4, the candidate *ache* is a specialisation of *pain* and a sibling of *colic*. This configures a larger search space of centroid candidates and a more flexible approach to construct the centroid than related works that limit to the values found in V or to their LCS (Guzman-Arenas et al. 2011; Abril et al. 2010).

From the set of centroid candidates, and following the centroid definition (Eq. 4.4), the term c in H_O that minimises the semantic distance to all the v_i in V, considering also their number of repetitions in the dataset (ω_i), will be selected as the final centroid. It is important to consider the number of repetitions of individual values so that the centroid is adapted to the data distribution (i.e., values with a high number of repetitions should weigh more during the centroid selection than scarcer ones). Relying on comparison operator proposed in 4.2 over the same background ontology, we evaluate each centroid candidate according to its aggregated distance to all input values.

Finally, we can define the semantic centroid of a set of textual attributes as follows.

Definition 7. The *semantic centroid (SC)* of a set of non-numerical values v_i in V is defined as the term c_j belonging to $H_O(V)$ that minimises the sum of weighted semantic distance (*wsd*, section 4.2, Eq. 4.2) with respect to all the values v_i in the space V.

$$centroid_O(V) = \{\arg\min_{\forall c_j \in H}(\sum_{i=1}^{p} wsd(c_j, < v_i, \omega_i >))\} \qquad (4.6)$$

where p is the cardinality of V.

When 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 for obtaining the centroid, let us take Example 4 and select as semantic distance $sd(c_1, c_2)$ the path length distance measure for its simplicity (i.e. the minimum number of links connecting c_1 and c_2 concepts (Eq. 3.1)). Taking the values in V_1, Table 6 shows the sum of weighted semantic distance (*wsd*, comparison operator for univariate data, Eq. 4.2) for each centroid candidate (first column) in H_{WN}. In this case, the candidate *ache* is the concept that minimises the sum of weighted semantic distances with respect to all values in V_1. Hence, applying Definition 7, $centroid_{WN}(V_1)=\{ache\}$.

Table 6. Sum of weighted semantic distance (*sum wsd*, last column) obtained for each centroid candidate (first column) in Example 4. Inner columns show the path length distance between each centroid candidate and the values belonging to V_1 (numbers in parenthesis are the amount of repetitions, ω).

Centroid candidates	colic (1)	lumbago (3)	migraine (2)	appendicitis (1)	gastritis (1)	pain (1)	*sum wsd*
colic	0	3	3	4	4	1	24
lumbago	3	0	2	5	5	2	19
migraine	3	2	0	5	5	2	21
appendicitis	4	5	5	0	2	3	34
gastritis	4	5	5	2	0	3	34
pain	1	2	2	3	3	0	17
ache	**2**	**1**	**1**	**4**	**4**	**1**	**16**
inflammation	3	4	4	1	1	2	27
symptom	2	3	3	2	2	1	22

The fact that all centroid candidates are evaluated in order 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 4, neither the LCS of V (*symptom* in Example 4) (Abril et al. 2010) nor the most frequently appearing value in V (*lumbago* with 3 appearances in Example 4) (Torra 2004; Domingo-Ferrer,Torra 2005) 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 ($sd_{WN}(symptom, V_1) = 22$). For Example 4, the optimal centroid (*ache*) is the result of commonly appearing term (*lumbago*, with 3 appearances) and an unbalanced distribution of terms within the hierarchy, that is, the *pain* branch includes more concepts than the *inflammation* one.

To show the influence of data value repetitions during the centroid construction, let us consider the following example:

Example 5. Given the set V_2 = *{<colic,1>, <lumbago,1>, <migraine,1>, <appendicitis,1>, <gastritis,1>}* with uniformly distributed values and being $H_{WN}(V_2)$ = *{colic, lumbago, migraine, appendicitis, gastritis, pain, ache, inflammation, symptom}*,(see Fig. 5) the sum of weighted semantic distance for each candidate is shown in Table 7.

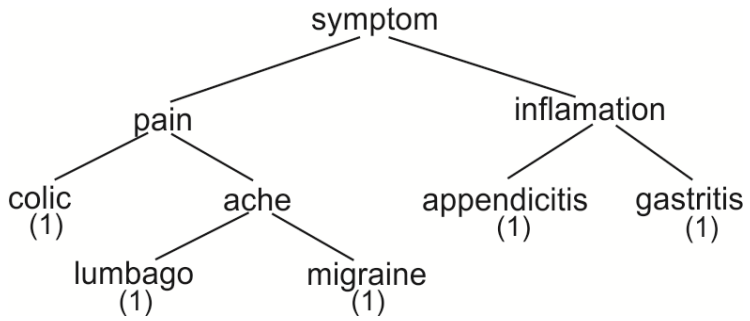

Fig. 5. Minimum hierarchy H_{WN} for V_2, extracted from WordNet, with the number of occurrences of each value.

Table 7. Sum of weighted semantic distance (*sum wsd*, last column) obtained for each centroid candidate (first column) in Example 5. Inner columns show the path length distance between each centroid candidate and the values belonging to V_2 (numbers in parenthesis are the amount of repetitions, ω).

Centroid candidates	Colic (1)	Lumbago (1)	Migraine (1)	Appendicitis (1)	Gastritis (1)	*sum wsd*
colic	0	3	3	4	4	14
lumbago	3	0	2	5	5	15
migraine	3	2	0	5	5	15
appendicitis	4	5	5	0	2	16
gastritis	4	5	5	2	0	16
pain	**1**	**2**	**2**	**3**	**3**	**11**
ache	2	1	1	4	4	12
inflammation	3	4	4	1	1	13
symptom	2	3	3	2	2	12

In this last case, the centroid is only influenced by the different terms found at each taxonomical branch and not its frequency. Because the hierarchy is unbalanced, the LCS "*symptom*" does not minimise the semantic distance. In fact, the candidate "*pain*" is the concept that minimises the sum of weighted semantic distance with respect to all values in V_2 being, hence, the centroid of the dataset.

4.3.4 Constructing the centroid for multivariate data

Real-world datasets are usually multivariate, with several attributes describing each object. In this section, we generalise the previous method to deal with multivariate categorical datasets.

We consider object attributes as independent variables because they potentially belong to different domains of knowledge (i.e., they could define a different set of modalities or even they could be mapped to different ontologies). Consequently, for each attribute, an individual centroid should be constructed so that is minimises the distance to all values in that attribute. The final centroid for a multivariate dataset is a *tuple* with the same cardinality as the number of attributes in the dataset. Note that because multivariate data is considered as a set of univariate attributes the search space of final centroid has a polynomial size ($|H_1|+...+|H_m|$).

Formally, let us take a multivariate dataset with n objects, m categorical attributes and p distinct tuples, the set MV of value tuples can be extracted: $MV=\{<\{v_{11},...,v_{1m}\},\omega_1>,...,<\{v_{p1},...,v_{pm}\},\omega_p>\}$, where $v_i=\{v_{i1},...,v_{im}\}$ is an object with m categorical attributes and ω_i is its number of occurrences in the dataset.

Considering them as independent attributes, let A_j be the set of distinct values in the jth attribute in MV so that $A_j=\{<a_1,\psi_1>,...,<a_q,\ \psi_q>\}$, where q is the cardinality of A_j, a_i is each distinct textual value for the jth attribute and ψ_i is its number of occurrences. Then, the centroid construction method presented in section 3.3.5.1 is applied to each univariate dataset A_j obtaining ($centroid_O(A_j)$. Concretely, applying Definition 5 over each A_j, a *minimum subsumer hierarchy* for each A_j will be extracted (H^j_O, covering all values in the jth attribute) configuring the set of centroid candidates for each attribute. The final centroid is the tuple of all attribute centroids, as follows.

Definition 8. The *semantic centroid (SC)* of a set of multivariate data MV in ontology O is defined as the tuple in which the jth value corresponds to the $centroid_O(A_j)$, being A_j the set of distinct values for the jth attribute in MV:

$$centroid_O(MV) = \{centroid_O(A_1), centroid_O(A_2), \dots, centroid_O(A_m)\} \quad (4.7)$$

where, m is the number of attributes in MV.

Note that this multivariate centroid corresponds to the optimal solution with respect to the minimisation of the global semantic distance to all objects in the dataset. By global distance we refer to the sum of weighted semantic distances for multivariate data as follow:

$$sum\_wsd(centroid(MV), MV) = \sum_{i=1}^{p} wsd(centroid(MV), v_i)\ \forall v_i\ in\ MV \quad (4.8)$$

where p is the cardinality of MV and wsd is the comparison operator defined in Definition 2, Eq. 4.3.

Hence, owing that each component of the centroid tuple ($centroid_O(A_j)$) minimises the distance of an attribute, the tuple also minimises the aggregated distance to all attributes.

Example 6. As illustrative example, consider a dataset with two attributes describing the conditions and treatments of a set of patients so that the following set of textual values can be extracted: MV_1 = *{<{colic, antibiotic}, 1>, <{lumbago, rehabilitation}, 3>, <{migraine, aspirin}, 2>, <{appendicitis, hospitalisation}, 1>, <{gastritis, codeine}, 1>, <{lumbago, codeine}, 2>, <{colic, hospitalisation}, 1>}*. Then, two sets of distinct textual values with their local number of occurrences are extracted: A_1 = *{<colic,2>, <lumbago,5>, <migraine,2>, <appendictits,1>, <gastritis,1>}*, for the first attribute and A_2 = *{<antibiotic,1>, <rehabilitation,3>, <aspirin,2>, <hospitalisation,2>, <codeine,3>}* for the second one.

The *minimum subsumer hierarchies* H^1_{WN} and H^2_{WN} extracted from WordNet (and labelled with ψ_1) are shown in Fig. 6 and Fig. 7 respectively. The centroid candidates for the first attribute are those in H^1_{WN} = *{colic, lumbago, migraine, appendicitis, gastritis, pain, ache, inflammation, symptom}*, whereas for the second one are those in H^2_{WN} =*{rehabilitation, antibiotic, aspirin, codeine, hospitalisation, therapy, medication, analgesic, medical_care}*. Note that in this case, 9+9=18 centroid candidates will be evaluated.

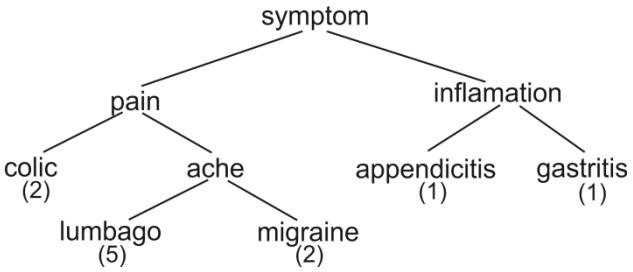

Fig. 6. Minimum hierarchy H^1_{WN} for A_1, extracted from WordNet, with the number of occurrences of each value.

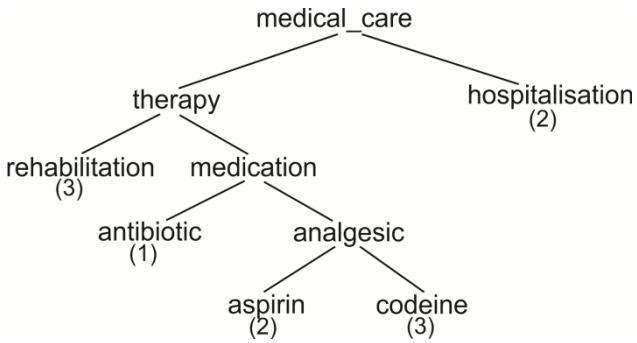

Fig. 7. Minimum hierarchy H^2_{WN} for A_2, extracted from WordNet, with the number of occurrences of each value.

Applying Definition 6, for each set H^j_O of centroid candidates, the centroid of the jth attribute ($centroid_O(A_j)$) is constructed. Using the path length-based measure (Eq. 3.1) as semantic similarity, the results for A_1 and A_2 are shown in Table 8 and Table 9, respectively.

Table 8. Sum of weighted semantic distance (sum *wsd*, last column) obtained for each centroid candidate (first column) in A_1 of Example 4. Inner columns show the path length distance between each centroid candidate and the values belonging to A_1 (numbers in parenthesis are the amount of repetitions, ψ).

Centroid candidates	colic (2)	lumbago (5)	migraine (2)	appendicitis (1)	gastritis (1)	*sum wsd*
colic	0	3	3	4	4	29
lumbago	3	0	2	5	5	20
migraine	3	2	0	5	5	26
appendicitis	4	5	5	0	2	45
gastritis	4	5	5	2	0	45
pain	1	2	2	3	3	22
ache	**2**	**1**	**1**	**4**	**4**	**19**
inflammation	3	4	4	1	1	36
symptom	2	3	3	2	2	29

Table 9. Weighted semantic distance (sd_o, last column) obtained for each centroid candidate (first column) in A_2 of Example 3. Inner columns show the path length distance between each centroid candidate and the values belonging to A_2 (numbers in parenthesis are the amount of repetitions, ψ).

Centroid candidates	Rehabilitation (3)	Antibiotic (1)	Aspirin (2)	Codeine (3)	Hospitalisation (2)	*sum wsd*
rehabilitation	0	3	4	4	3	29
antibiotic	3	0	3	3	4	32
aspirin	4	3	0	2	5	31
codeine	4	3	2	0	5	29
hospitalisation	3	4	5	5	0	38
therapy	1	2	3	3	2	24
medication	**2**	**1**	**2**	**2**	**3**	**23**
analgesic	3	2	1	1	4	24
medical_care	2	3	4	4	1	31

As a result, the centroid of A_1 is $centroid_{WN}(A_1)$={*ache*} and the centroid of A_2 is $centroid_{WN}(A_2)$={*medication*}. Hence, applying Definition 8, the centroid of MV_1 taking WordNet as knowledge base is $centroid_{WN}(MV_1)$ = {$centroid_{WN}(A_1)$, $centroid_{WN}(A_2)$} = {*ache*, *medication*}.

Note that the sum of global distance for our centroid is *sum wsd*({*ache*, *medication*}, MV_1) = 21, whereas other criteria, such as the LCS (for which the centroid is *sum wsd* ({*symptom*, *medical_care*}, MV_1) = 30) and the mode (for which the centroid is *sum wsd* ({*lumbago*, *rehabilitation*}, MV_1) = 24.5) would give worse results.

4.3.5 Evaluation of aggregation operator

In this section, we evaluate the centroid construction strategy proposed in section 4.3.4. Our method is compared against those proposed by related works (being or not semantically-grounded). In order to numerically quantify the quality of the centroid proposed by each method, we compute the sum of semantic distance between the centroid and each element in the input dataset as in Eq. 4.8. The semantic distance (*sd*) function has been assessed with the measure dis_{LogSC} presented in Eq. 3.9, because it better quantifies the semantic differences between terms (Sánchez et al. 2012a). In this manner, we are able to objectively quantify the loss of semantic content resulting from the substitution of a dataset or cluster by its centroid and, hence, the quality of the centroid. This is an important aspect from the point of view of data exploitation as it represents a measure of up to which level the semantics of the original values are preserved in the centroid.

4.3.5.1 Evaluation data

The first dataset used for evaluating the methods proposed was provided by *Observatori de la Fundació d'Estudis Turístics Costa Daurada* at the *Delta de l'Ebre Catalan National Park*. As evaluation dataset we used a set of answers given by visitors of the *Delta de l'Ebre Catalan National Park*. Each object in the dataset corresponds to an individual and consists of a tuple (i.e. record) with two textual answers expressed as a noun phrase, corresponding to the two main reasons to visit the park. They correspond to answers to open questions posed to visitors. The textual values for these two attributes are widely used concepts mainly related to the visitor's hobbies, interests and activities (i.e. sports, beach, fishing, ecology, etc.) which can be directly found in WordNet.

The dataset comprises 975 records corresponding to the individual answers. Analysing the dataset distribution, we observe a total of 211 different response combinations, being 118 of them unique. The tuple with the highest amount of repetitions appears 115 times. Fig. 8 details the distribution of data according to the number of repetitions (i.e., frequency of appearance).

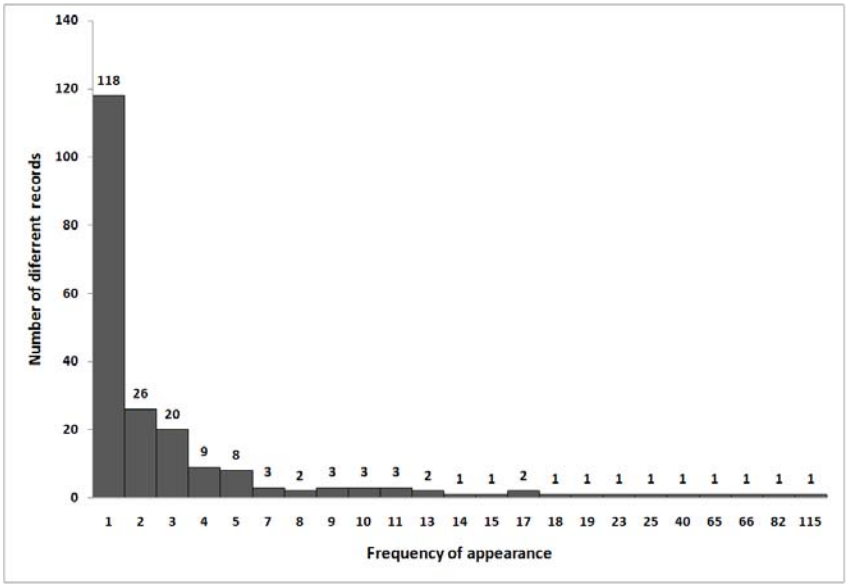

Fig. 8. Dataset value tuple distribution.

As stated before, the centroid is typically needed either when selecting a "central" value for all the input data from which data classification algorithms, for example, K-means or MDAV can be executed, or when extracting a "representative" for each of the clusters resulting from the classification process. Even though the centroid can be obtained in both cases in the same manner, the characteristics of the data are different in each case. Due to the goal of a

classification algorithm (which is to minimise the distance between the elements in each cluster), one can expect that when considering the whole input set, data distribution would be more spread, with less homogenous and more distant values than when evaluating clusters individually. As a result, in the former case, the centroid calculation will be more affected by the presence of outliers and heterogeneously distributed values, a circumstance that will negatively influence some methods more than others.

In order to consider these two well-differentiated cases during the evaluation and to analyse the behaviour of each method in each circumstance, we configured two different scenarios. In the first one (*Scenario1*), the whole dataset described above will be analysed, constructing the centroid for all the 975 records. In the second one (*Scenario2*), an ontology-based hierarchical clustering algorithm (Batet et al. 2008), aimed to build a partition of the individuals according to their preferences, has been applied to the dataset. In this clustering method, ontologies are exploited to guide the comparison of categorical features, assessing their resemblance using semantic similarity measures. According to these similarities, an iterative aggregation of objects is performed based on Ward's clustering method (Ward 1963). As a result, a hierarchical classification of non-overlapping sets of objects is obtained, that is, a dendogram. The height of the internal nodes in the resulting dendogram is related to the variance of the clusters. Then, a cutting level in the dendogram is selected so that the cohesion of the resulting groups and the distinguishability among clusters is maximised. In this case, as a result of the clustering process, seven clusters of records have been obtained. Their size and data distribution are shown in Fig. 9.

When comparing the data distribution of each individual cluster against the whole dataset (Fig. 8 vs. Fig. 9), it is easy to realise about the higher homogeneity of the data in the clusters. It is also interesting to observe the differences between clusters, which range from highly homogenous ones, for example, *Cluster1*, with a dominant value tuple of 109 repetitions to less compact ones, as for example, *Cluster6* or *Cluster2*, with more spread value tuples. These last ones tend to incorporate a higher number of outliers that can hardly fit in other clusters due to their higher cohesion.

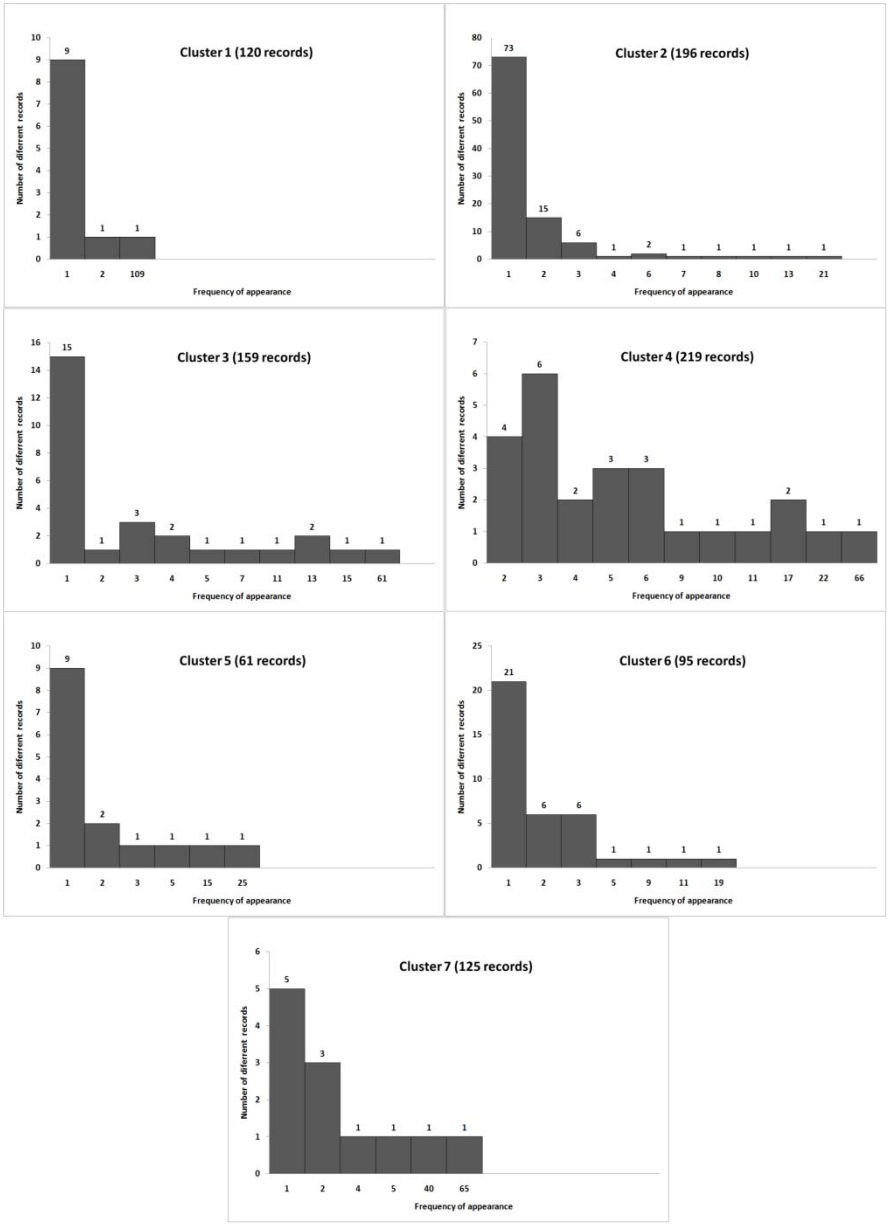

Fig. 9. Value tuple distribution for each cluster.

4.3.5.2 Implemented methods

The Semantic Centroid (SC) proposal has been tested using three edge counting-based semantic measures introduced in section 3.2.1 (path (Rada et al. 1989), W&P (Wu,Palmer 1994) and LogSC (Sánchez et al. 2012a)) have been applied to Eq. 4.8, that is, during the centroid construction as goal function.

Furthermore, for evaluation purposes, the following strategies applied by related works have been also implemented:

- *Random centroid* (RC): a random record of the input dataset is selected as the centroid. This naïve strategy is used as baseline to compare against other methods.
- *Most frequently appearing centroid* (MFC): the record in the dataset with the highest amount of repetitions (e.g., 115 when considering the whole dataset) is considered as the centroid (i.e., mode). This is a common strategy employed by non-semantic approaches like (Torra 2004; Domingo-Ferrer,Torra 2005; Erola et al. 2010; Cao et al. 2012; Bai et al. 2011; Zhang et al. 2010) in which textual data are evaluated in a categorical fashion.
- *Least Common Subsumer centroid* (LCSC): this method relies on a background taxonomy/ontology and selects, as centroid, the concept that subsumes all the different values found in the dataset. When dealing with multivariate data, the subsumer is a tuple so that each value individually subsumes all the values of an attribute. According to the input data and the background ontology, the resulting centroid may or may not belong to input data. This is also a common strategy used by semantically-grounded methods like (Abril et al. 2010).
- *Consensus-based centroid* (CC): reproducing the method proposed in (Guzman-Arenas et al. 2011; Guzman-Arenas,Jimenez-Contreras 2010), the centroid is selected as the record in the dataset that minimises the average *confusion* to all other records. A background taxonomy/ontology is also used to map objects' values to concepts so that the confusion can be computed as the number of descending links in the hierarchy from *a* to *b* when substituting the former with the latter. Note that, in the original method, the background ontology is constructed according to the individual values in the input dataset. This hampers the method when input values cannot be taxonomically organised among them, for example, when most values correspond to leaves in the taxonomy. To enable a more favourable situation against our methods, we used the same background ontology to compute the confusion between values.

In all cases in which an ontology is required as background knowledge, we used WordNet to retrieve candidate centroids and/or compute semantic similarity measures.

4.3.5.3 Evaluation with scenario 1

The first scenario considers the whole dataset, that is, 975 records. We quantify the centroid's accuracy in preserving data semantics by computing the proposed comparison operator for semantic categorical attributes (using Eq. 4.8 and the dis_{logSC} measure Eq. 3.9 (Sánchez et al. 2012a)) between the centroid obtained with each of the methods listed above and all the objects in the dataset. Results are presented in Fig. 10.

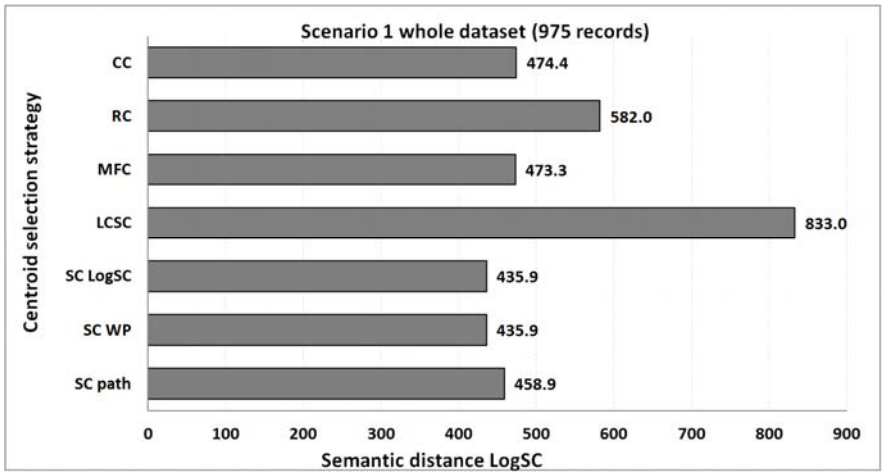

Fig. 10. Comparison of centroid construction strategies.

Analysing the results, we observe that our method (SC) is able to improve related works in all cases. Regarding the different semantic measures used to measure the distance used during the construction of the centroid (Eq. 4.7), LogSC and W&P measures give identical results, whereas the path-based measure provided slightly worse results. As stated in section 3.2.1, measures relying only on the taxonomical path length provide less accurate results than other strategies because the path omits semantic evidences explicitly available in the ontology that are relevant in similarity estimation, such as the relative depth of concepts in the taxonomy or the presence of multiple inheritance relations.

Compared to related works, we observe that the worst method was the one using the LCS of the whole data as the centroid (LCSC). This method does not quantify the semantic distance between values and does not consider their distribution (i.e., number repetitions). It only assumes that the most specific common ancestor to all the values will serve as a good representative for all the objects. As discussed in section 2, however, this method is hampered by the presence of value outliers that are semantically far to most other values in the dataset, forcing the generalisation of the LCS to very abstract and, hence, semantically distant concepts. Even though values in the evaluated dataset are relatively homogenous (see Fig. 10), the variability of responses, with 118

answers unique answers, forced the generalisation to the root concept of the ontology (i.e., entity) as the LCS. Hence, the centroid is semantically distant to most values in the dataset, resulting in a high loss of semantics.

The other semantically grounded related work, based on measuring the *confusion* between input values (CC), significantly improved the above method by numerically quantifying the cost of substituting a value with a specialisation. However, the fact that this quantification is based only on the path, that is, number of descending levels in the taxonomy), that only input values are considered as centroid candidates, and that the cost of replacing a value with a subsumer is not considered, provided worse results than our methods. Note that in this case, the method is favoured by the fact that a large ontology as WordNet is used to compute confusion values, rather than a constrained taxonomy constructed ad-hoc from input values.

Finally, methods based on data distribution performed reasonably well thanks to the characteristics of data. On the one hand, the MFC method, based on selecting the most frequently appearing value tuple, is favoured by the fact that a tuple with a significant 115 repetitions is found. Hence, by selecting this tuple as centroid it is achieved a zero semantic distance to 115 records of the dataset. Moreover, this tuple corresponds to nature-related reasons that are similar to other frequently appearing tuples, resulting in a low semantic distance. As a result, the aggregated semantic distance is implicitly minimised. A similar situation is observed for the naïve strategy (RC) that, thanks to the presence of numerous sets of frequently appearing tuples, it increases the chance to randomly select a tuple that results in a low semantic distance. These behaviours, however, depend very closely on the characteristics of the input data, in particular, on the fact that predominant values exist. When dealing with more scattered and less cohesive datasets, results are likely to worsen.

Regarding the computational cost of the different centroid calculation strategies, in an univariate setting, we observe that our method (SC) scales as $O(|H|\cdot p)$, where $|H|$ is the number of concepts in the minimum hierarchy modelling values in the input dataset (i.e., the number of centroid candidates), and p is the number of *distinct* values in the dataset, to which, for each centroid candidate, the weighted sum of distances should be computed. Note that, when dealing with categorical data, p is usually much lower than the size of the dataset. This is because, as shown in the data distribution analysis, categorical data usually present a limited set of modalities that tend to repeat. Since multivariate datasets (with m attributes) are considered as sets of univariate attributes, the cost is maintained at a polynomial size $O(|H_1|\cdot n_1+...+|H_m|\cdot n_m)$. In comparison, the *consensus-based centroid* (CC) selects the record in the dataset that minimises *confusion* against all other records. Confusion corresponds to the number of descending links between records in the background ontology, which in the worst case would be the maximum taxonomical depth of H. For univariate data, this scales $O(h\cdot r)$, where n is the number of records in the dataset and h is the maximum taxonomical depth of H. For multivariable datasets, this results in $O(h_1\cdot n +...+ h_m\cdot n)$, where h_i is the maximum depth of the H_i for the i^{th} attribute.

Even though $h_i<|H_i|$, as stated above, typically $n>p_i$. As a result, the final cost is similar to that of our method. The *Least Common Subsumer centroid* (LCSC) simply picks the LCS generalising all the values of each attribute, resulting in a cost of $O(|H|)$ for univariate data and $O(|H_1|+...+|H_m|)$ for multivariate datasets. Hence, the expected runtime is lower than that of our method but at the cost of sensibly worse results. Finally, the non-semantic approach (*Most frequently appearing centroid* (MFC)) offers the lowest cost ($O(p)$, being p the number of distinct values/tuples in univariate/multivariate datasets), at the cost of lacking both distance and semantic optimisations.

4.3.5.4 Evaluation with scenario 2

The second scenario evaluates, individually, each of the seven clusters in which the dataset has been classified, as described in section 4.3.5.1. The semantic distance of the obtained centroids for each cluster using different strategies to build centroids is shown in Fig. 11. In addition, we have calculated the accumulated sum of distances for all the clusters and each strategy.

The consideration of clustered data allows analysing the differences observed between different centroid construction strategies according to the characteristics of each cluster. On the one hand, when dealing with highly homogenous clusters (like *Cluster1*), all approaches, except the LCSC, tend to provide the same –optimum- results. In fact, analysing the frequency distribution of data in Cluster1 (Fig. 8), we observe a clear predominant value, with 109 repetitions in a cluster with 120 elements. As a result, this value is selected as the centroid of the cluster by all strategies except the LCSC, which selects a more abstract taxonomical ancestor that also subsumes scarce values. This is the most favourable scenario for methods based on data distribution because the most frequent tuple is the optimum representative. *Cluster3* also presents a similar behaviour, even though the RC approach has not selected the most frequent tuple and, hence, the distance is higher. In this case, the chance to randomly select the optimal centroid is lower than with *Cluster1* due to the lower predominance of the most frequent tuple (the probability of selecting randomly the mode is p=0.98 for the *cluster1* and p=0.38 for the *cluster3*). On the other hand, when dealing with less uniform clusters with no predominant values, like *Cluster5*, *Cluster6* or *Cluster7*, only our proposal is able to obtain optimal centroids. In this case, the selection of a taxonomical subsumer results in a lower semantic distance than the most frequent value tuple. This shows that distributional methods are not ideal when dealing with less cohesive datasets, because centroids are selected only from those values in the input dataset, a characteristic also shared with CC. Hence, their accuracy heavily depends on the distribution of objects. Only our methods and the LCSC expand search space to other concepts extracted from the background ontology. However, as stated above, the LCSC strategy is hampered by the presence of outliers and the lack of frequency analysis, providing the worst results from the bunch.

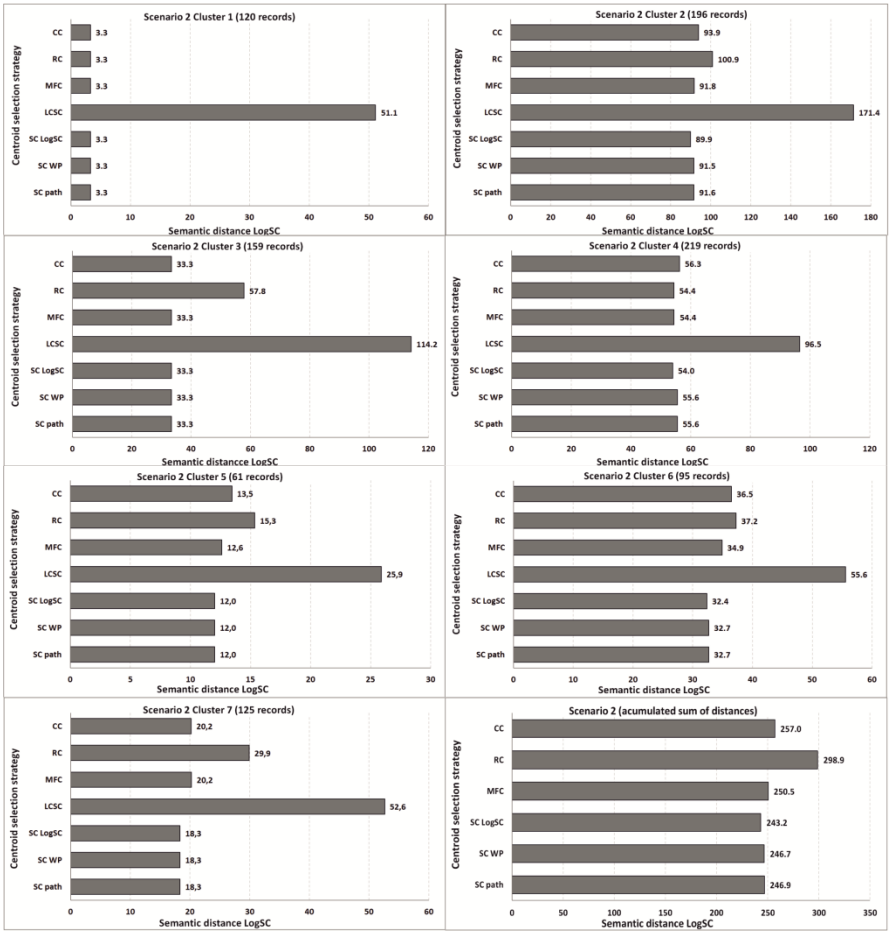

Fig. 11. Semantic distances between each cluster and its centroid.

In addition, we have calculated the accumulated sum of distances for all 7 clusters (with their corresponding 7 centroids) to have a measure of the overall loss of semantics when the objects are substituted by their respective centroid, according to the clustering assignment. We observe that our proposal provides the best results.

4.4 Sorting operator

Some anonymisation methods (like *resampling* (Martínez et al. 2012a)) require from sorting input data before grouping them. Again, sorting categorical values is not straightforward since, in general, they are not ordinal (i.e. they do not take

values in an *ordered* range). In this section, we propose a sorting algorithm for categorical data. This operator takes advantage of the two previous operators: comparison (semantic distance) and aggregation (centroid). In this way the semantics is captured during the sorting process, by making use of knowledge ontologies.

To sort a set of values, a reference point is needed, so that values could be arranged according to their similarity/distance to that reference. Numerically, this is done according to the max/min value (i.e. the most extreme value) of the set. To define a sorting procedure for the categorical case, we also rely on the notion of the most extreme value, which corresponds to the one that, globally, is the *most distant* to all other values (conceptually, this is the opposite of the centroid as computed in section 2.3). Once this reference value is obtained, other values are sorted by iteratively picking those that are least distant to that extreme value.

To obtain the reference value/record as well as to compare it to other elements in the set, considering both the semantic and distributional features of data, we rely on the weighted semantic distance and the centroid calculus procedures proposed in sections 4.2 and 4.3 respectively. The sorting procedure is formalised in Algorithm 1.

Algorithm 1. Sorting procedure

Inputs: P (dataset)

Output: P' (P sorted)

```
1  Compute the centroid of all records in P
2  Consider the most distant record f to the centroid
3  Add f to P' and remove it from P
4  while (|P| > 0) do
5     Obtain the least distant record r to f
6     Add r to P' and remove it from P
7  Output P'
```

The method starts by picking the record that is the *most distant* (according to Eq. 4.2) to the *centroid* of the dataset (Eq. 4.6) (lines 1-2) as the reference point (*f*). Next, the rest of records are sorted by taking, at each iteration, the *least distant* one (Eq. 4.2) to the reference *f* (lines 4-6). Eq. 4.3 and Eq. 4.7 should be used, instead, in case of multivariate datasets. The evaluation of this method is left to the next chapters. In particular in chapter 7 it will be applied to the resampling method.

4.5 Summary

In this chapter, we proposed a general framework that enables the anonymisation of structured categorical data from a semantic perspective. The semantic framework proposes and formalises three operators (*comparison*, *aggregation* and *sorting*) that, by exploiting knowledge structures (like WordNet or SNOMED CT), and relying on the theory of *semantic similarity*, these operators enable a semantically-coherent interpretation of non-numerical attributes, while also considering their distributional features.

To deal with categorical data, an appropriate comparison operator is needed. We proposed a *comparison* operator that considers both semantic and distributional features of categorical data. The comparison operator between terms is based on the notion of semantic similarity and exploits ontologies as knowledge base.

An *aggregation* operator is usually needed in data analysis algorithms, both for discovering the central value of a dataset from which the data aggregation can be performed and also for constructing representative values of resulting clusters. Hence, the accuracy of these data analysis methods closely depends on the quality of the aggregation operator. As aggregation operator we propose the use of centroids. We proposed a new approach to construct centroids of datasets with textual attributes with special emphasis on the preservation of data semantics.

The proposed centroid construction method is evaluated using a real dataset consisting of answers to polls made by the *Observatori de la Fundació d'Estudis Turístics Costa Daurada* at the Delta de l'Ebre Catalan National Park. In comparison, other semantically-grounded methods (Abril et al. 2010; Guzman-Arenas et al. 2011) fail at minimising distances between records and the centroid due to the suboptimal (Guzman-Arenas et al. 2011) or even null (Abril et al. 2010) semantic assessment. The evaluation also showed that the, apparently coherent, strategy of using the LCS as the centroid (Abril et al. 2010) provided the worst results due to the need of generalising outliers. This shows the convenience of taking into account data distribution in conjunction with their semantics as our method does. In fact, methods based on using the most frequent tuple as the centroid (Bai et al. 2011; Cao et al. 2012; Domingo-Ferrer,Torra 2005; Erola et al. 2010; Torra 2004; Zhang et al. 2010) performed reasonably well because they tend to implicitly consider the distributional distances, but not the semantic distances.

Some masking methods require a sorting procedure. In this chapter we proposed a *sorting* operator appropriate for categorical data. This operator is based on the two previous operators: comparison and aggregation. The sorting criterion is the semantic similarity of concepts with respect to the centroid.

Comparison, aggregation and sorting operators compose the proposed semantic framework that can be used to adapt SDC methods, so that structured categorical

data could be *k*-anonymised (to minimise the disclosure risk) while retaining, as much as possible, their semantics and hence, their utility.

By means of the framework presented in this chapter and exploiting a knowledge base like WordNet or SNOMED CT, categorical terms can be coherently compared and aggregated. Since many SDC methods rely on these operators to anonymise data, our framework enables an accurate anonymisation of datasets according to their semantic and distributional features. This will contribute to minimise the information loss of masked data and, hence to retain the data utility.

The main publications related to contributions presented in this chapter are:

1J. Martínez, S., Valls, A., Sánchez, D.: Semantically-grounded construction of centroids for datasets with textual attributes. International journal: Knowledge-Based Systems. **35**(0), 160-172 (2012). *Impact Factor*: 2.422

2J. Martínez, S., Valls, A., Sánchez, D.: A semantic framework to protect the privacy of Electronic Health Records with non-numerical attributes. Journal of Biomedical Informatics. Accepted manuscript. *Impact Factor*: 1.792

Chapter 5
A new recoding anonymisation method

In this chapter, we use the semantic based framework presented in chapter 4 to design a recoding method preserving the semantic content of categorical data. Previous approaches neglected or only shallowly considered the semantic content of categorical attributes.

Data *recoding* provides a simple way to build *k*-anonymous datasets. The method addresses the problem of masking a subset of the categorical attributes of the input record set in a global manner. As stated in section 2.2, four types of attributes are distinguished: identifiers, quasi-identifiers confidential and non-confidential. Only the first two may lead to the re-identification of individuals. Identifiers are directly removed from the dataset because they refer to values that are unique for each individual (e.g. personal identification number or social security number). As a consequence, the masking process would be applied over tuples of categorical quasi-identifier attributes.

Unlike the exhaustive generalisation methods based on the VGHs analysed in section 2.5.3, this approach deals differently with the global masking process. Thanks to the wide coverage of ontologies as WordNet or SNOMED CT, we would be able to map categorical attribute values into ontological nodes that do not necessarily represent leaves of a hierarchy. As a result, semantically related concepts can be retrieved by going through the ontological hierarchy/ies to which the value belongs. These ontological hierarchies are designed in a much more general and fine-grained fashion than VGHs and, according to the agreement of domain knowledge experts, not as a function of the input data. This opens the possibility of substituting values by a much wider and knowledge-coherent set of semantically similar elements. To ensure scalability with regard to ontology size and input data, we bind the space of valid value changes to the set of value combinations that are present in the input dataset. When changing one value of a record for another, we can substitute a taxonomical subsumer with another one (this is the only case covered by the generalisation method) but also with a hierarchical sibling (with the same taxonomical depth) or a specialisation (located at a lower level). In fact, in many situations a specialisation can be more similar than a subsumer, due to their higher specificity concepts belonging to lower levels of a hierarchy have less differentiated meanings. As a result, the value change

would result in a higher preservation of the semantic of data. This is an interesting characteristic and an improvement on the more restricted data transformations supported by VGH-based generalisation methods.

5.1 A recoding method to mask categorical attributes

Finding the optimum anonymous partition requires the generation of all the possible value fusions for all the non-anonymous records, which has an exponential cost. To ensure the scalability of our approach, we opted for a greedy algorithm which selects, at each iteration a feasible value fusion. However, with an uninformed approach, the quality of the result would depend on the selection of the records at each step. To solve this, an exhaustive method that tests all the combinations can be used, with a factorial cost with respect to the number of non-anonymous records. This approach is again computationally too expensive because, as records are defined by unbounded textual attributes, they usually correspond to a high number of combinations, many of which are unique, thus leading to a large number of records that do not satisfy *k*-anonymity. The proposed method is based on the substitution of all quasi-identifier values of each record with the values of another record. To ensure the scalability of the method and guide the anonymisation process towards the minimisation of information loss, we have designed two heuristics (H) that ensure the selection, at each iteration, of the best set of records to transform:

H_1) From input dataset, select the set of records T with the lowest number of repetitions in the original set. The goal of this heuristic is to start the process with the records that less fulfil the *k*-anonymity property.

H_2) For each record $t \in T$ find the least distant record v in the input dataset. The aim of this heuristic is to maximise the quality of the anonymised dataset at each substitution.

To select the value that minimises the information loss resulting from the data substitution, the comparison operator (section 4.2) is used to select the most similar one. As a result of the replacement, quasi-identifier values for both records (the one to anonymise p and the most semantically similar one v) will take the same values and become indistinguishable; therefore, the *k*-anonymity level for both records will increase. By repeating the process iteratively for each non-anonymous record according to a certain value of *k*-anonymity, the input dataset will be anonymised.

The recoding algorithm presented in (Martinez et al. 2011) is given in Algorithm 2, highlighting in **bold** the steps in which the appropriate operators for numerical o categorical data are needed.

Algorithm 2. Recoding

Inputs: D (original dataset), k (level of anonymity)

Output: D^A (a transformation of D that fulfils the k-anonymity)

```
1   D^A := D
2   make a set P with the records of D^A with the min
    number of repetitions (H_1)
3   num_repetitions := min repetitions
4   while (num_repetitions < k) do
5     find the least distant record v ∈ D^A with respect
      to each t ∈ T (H_2)
6     replace each record t in D^A by the corresponding
      record v
7     make a set T with the records of D^A with the min
      number of repetitions (H_1)
8     num_repetitions := min repetitions
9   output D^A
```

First, the algorithm selects the group of records with the minimum number of repetitions (line 2). As long as this number is lower than *k,* the dataset is not *k*-anonymous. To increase the number of repetitions of value tuples, it looks for the non-minimal records with the least distant value tuples and with the lowest amount of repetitions (line 5). Then, original values are replaced by the least distant ones (line 6), increasing their anonymity level due to the higher amount of value repetitions. The process is repeated for the next least frequent record set (line 7) until the *k*-anonymity is fulfilled (line 4).

To compare numerical data is usually used the Euclidean distance. The adaptation of this method to categorical data is straightforward using the comparison operator proposed in the semantic framework in section 4.2. In line 5, this operator will provide the least *semantically* distant and least *frequent* record set to the reference value (Eq. 4.3), which is the record value with the lowest amount of repetitions (obtained in line 2). So, in this step we are considering both the data distribution as well as the semantic similarity of the terms.

Next section will evaluate the performance of this method. Analysing the computational cost of anonymisation, it is $O(p^3)$, where p is the number of different records in the dataset ($p \leq n$), where n is the number of records in the

dataset. In fact, the computationally most expensive step is the calculation of the semantic similarity between all the pairs of different records. Since each record has m values, this operation requires to execute $m \cdot p^2$ times the semantic similarity between a pair of single values. In the worst case, we require p iterations to build the valid partition (loop in line 4), so the final cost of the algorithm is $m \cdot p^2 \cdot p = m \cdot p^3$ times, where m is a relatively small number compared to p because the set of quasi-identifier attributes is usually small.

For large datasets, where p can be large if we deal with infrequent (unique) records, the scalability is more critical. For this reason we have optimised the implementation. Notice that the semantic similarity between records is measured in line 5 and is repeated at each iteration. As the set of different attribute values and distinct record tuples is known *a priori* and does not change during the masking process (unlike for generalisation methods), the similarities between all of them can be pre-calculated and stored. This avoids recalculating the similarity for already evaluated value pairs and, more generally, register pairs. In this way, the similarity measure is calculated *a priori* only $m \cdot p^2$ times, improving the efficiency with respect to the most expensive function of $O(p^2)$. As we illustrate in the evaluation section, with this modification the execution of the algorithm stays in the range of milliseconds for large datasets.

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

5.2 Evaluation

In this section is detailed the evaluation of the recoding method. We first study the contributions of the proposed heuristics with regards to the information loss. Afterwards we evaluate the data utility for data mining purposes of masked datasets comparing semantic and distributional approaches. Finally we have evaluated the disclosure risk and computational cost of the recoding method.

5.2.1 The dataset

We have evaluated this recoding method by applying it to a dataset consisting of answers to polls made by the *Observatori de la Fundació d'Estudis Turístics Costa Daurada* at the Delta de l'Ebre Catalan National Park. Visitors were asked to respond to several questions on their main reasons and preferences for visiting the park (see an extract in Table 10).

The dataset comprises 975 individual records and 26 variables. Two of the variables are categorical attributes expressing many semantic answers (the last two columns of Table 10) includes a set of textual answers expressed as a noun. Because of the variety of answers, the disclosure risk is high and individuals are easily identifiable. We therefore consider these answers as quasi identifiers that should be anonymised.

Table 10. Extract of sample microdata used for evaluation. The last two columns are textual attributes masked with our approach.

Age	Gender	Duration (in days) of the visit to the park	Number of companion	Origin	Reason for visiting the park	Main activities during the visit to the park
23	M	1	2	Spain	nature	fishing
26	M	3	1	Spain	landscape	sports
45	F	3	2	Belgium	sports	bicycling
56	M	1	0	France	nature	culture
54	F	2	0	Spain	nature	fishing
26	F	5	3	France	fishing	nature
45	F	1	1	Spain	relaxation	culture
30	M	2	0	Holland	holidays	visit
37	F	2	3	Spain	Second residence	beach

Considering these two attributes to be quasi-identifiers, we find a total of 211 different responses, 118 of which were unique (i.e. identifying a single person). Notice that if the person is identified, some confidential data may be released, such as the age or the number of accompanying persons (see Table 10). Fig. 12 shows the equivalence class structure defined by the values of the pair of attributes considered in this study. Note that this sample represents a much wider and more heterogeneous test bed than those reported in related works (Samarati,Sweeney 1998), (Li,Li 2008), which focused on bounded a fixed set of answers)(terms) in the categorical attributes.

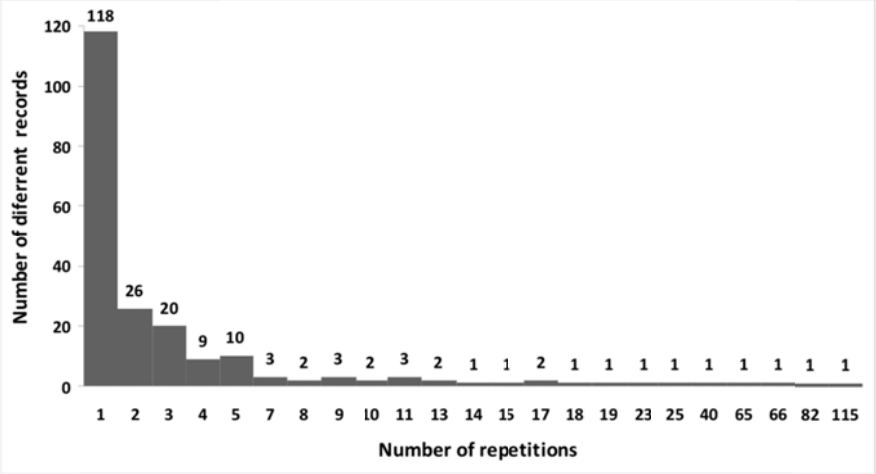

Fig. 12. Attribute distribution according to answer repetitions

The answer values for these two attributes are general and widely used concepts (i.e. sports, beach, nature, etc.). All of them are found in WordNet 2.1, which allows this ontology to be used to perform the semantic similarity measurement. However, as we are dealing with values represented by text labels we had to morphologically process them in order to detect different lexicalisations of the same concept (e.g. singular/plural forms). We applied the Porter Stemming Algorithm (Porter 1997) to both text labels of attributes (e.g. *sports*) and ontological labels (e.g. *sport*) in order to extract the morphological root of words (e.g. *sport*) and be able to map values to ontological concepts and detect conceptually equivalent values in the dataset (e.g. *relaxation=relax* as the morphological root of both words is *relax*).

5.2.2 Evaluation of the heuristics

In this section we evaluate the contribution of each of the designed heuristics in guiding the substitution process towards minimising the information loss from a semantic point of view. The quality of the masked dataset has been evaluated by measuring the information loss according to how semantically similar the masked values are, on average, compared to the original ones. Information loss has been computed and normalised as defined in Eq. 5.1. The same evaluation was repeated for different levels of *k*-anonymity.

$$semantic\_loss\left(D^A\right) = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m} sd\left(r_{ij}, r_{ij}^{\ A}\right)}{n * m} \qquad 5.1$$

To show the contribution of each heuristic in minimising the information loss of the results, we replaced the heuristic substitution by a naïve replacement that changes each sensitive record by a random one from the same dataset. Using the same basic algorithm presented in section 5.1, each random change will increase the level of *k*-anonymity until all records are anonymised. For the random substitution, records are ordered alphabetically in order to avoid depending on the initial order of data. The results obtained for the random substitution are the average of five executions. The two heuristics proposed in section 5.1 were gradually introduced instead of the random substitution in a way that enables the contribution by each heuristic to the resultant quality to be quantified. The results of this test are shown in Fig. 13, where: no heuristic at all is considered; only the first one is considered; the first one and the second one are considered.

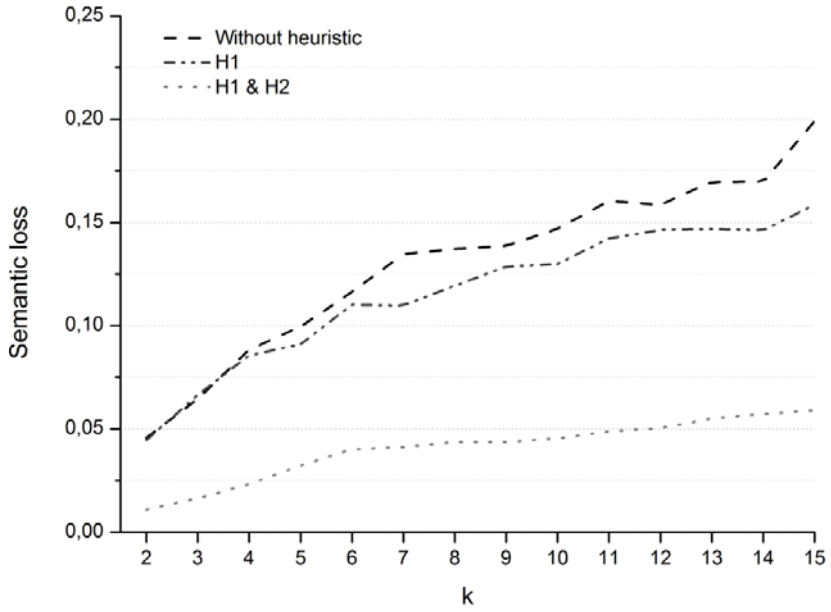

Fig. 13. Contribution of each heuristic to the anonymised dataset quality.

The results illustrated in Fig. 13 are consistent with what it is expected from the design of each heuristic. The first one, which only re-orders input data according to the degree of record repetition in order to prioritise the less anonymous records, leads to a slight improvement on the complete random substitution. The second one, which incorporates the semantic similarity function as a metric to guide the value fusion process towards the minimisation of the semantic loss, leads to the most significant improvement.

As a result of the heuristic fusion process, our approach considerably improves the naïve replacement. This is even more noticeable for a high *k*-anonymity level (above 5); when the two heuristics were used, we clearly outperformed the semantic loss of the random version. This is highly convenient and shows that our approach performs well regardless of the desired level of privacy protection.

5.2.3 Comparing semantic and distributional approaches

To show the importance of a semantically focused anonymisation, we compared it with a more traditional schema that focused on the distributional characteristics of the masked dataset (as stated at the section 2.5.3). This was done by using the Discernibility Model (DM) (Bayardo,Agrawal 2005) (Eq. 5.2). This measure is used to evaluate the distribution of *n* records (corresponding to *n* individuals) into *g* groups of identical values, generated after the anonymisation process. Concretely, DM assigns to each record a penalty based on the size of the group g_i to which it belongs after the generalisation. A uniform distribution of values in groups of similar size would optimize this metric:

$$C_{DM} = \sum_{i=1}^{n} |g_i|^2 \qquad (5.2)$$

We used DM in our algorithm instead of the semantic distance measure to guide the masking process. Both semantic and distributional approaches were compared by evaluating the semantic distance between the original and masked dataset, as stated in Eq. 5.1. The semantic distance *sd* is calculated as stated in Eq. 4.1.

Fig. 14 shows that the optimisation of the dataset distribution does not imply better preservation of the records' semantics. In fact, there is a noticeable semantic loss in the resulting dataset for *k*-anonymity values above 5 for the distributional approach. As we stated in the introduction, the utility of categorical information from the data analysis point of view strongly depends on its semantics. We can see that classical approaches that focus on providing uniform groups of masked values may significantly modify a dataset's meaning, thus hampering its exploitation.

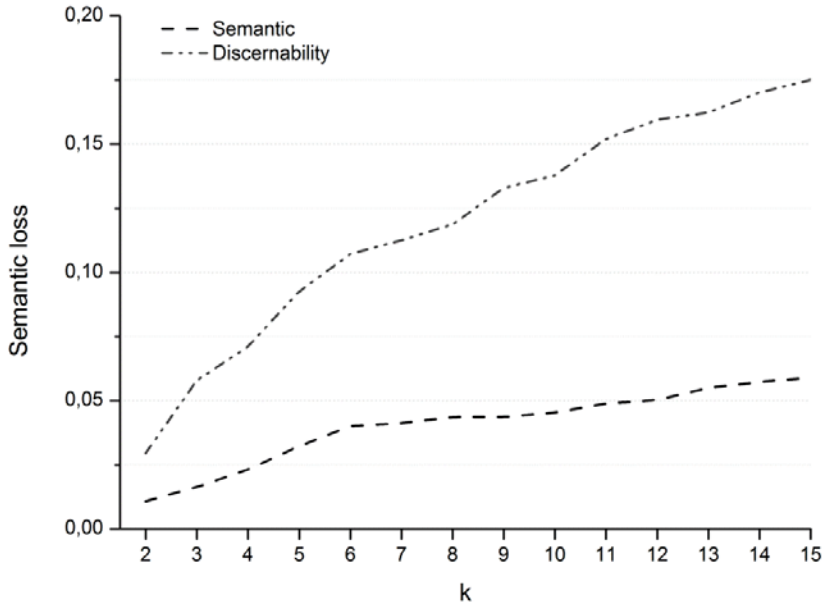

Fig. 14. Similarity against original data for semantic and distributional anonymisations.

5.2.4 Evaluation of data utility for semantic clustering

In order to evaluate the hypothesis that, from the data exploitation point of view, a semantic-driven anonymisation retains the utility of the original data better than distributional approaches, we then compared the utility of the dataset resulting from both approaches in a concrete data mining application.

As stated in the introduction, data acquired by statistical agencies are interesting for data analysis in order, for example, to extract user profiles, detect preferences or perform recommendations (Domingo-Ferrer 2008). Data mining and, more specifically, clustering algorithms are widely used for organising and classifying data into homogenous groups. Although clustering algorithms have traditionally focused on continuous-scale numerical or bounded categorical data, the increase in volume and the importance of unbounded textual data have motivated authors to develop semantically grounded clustering algorithms (He et al. 2008).

In (Batet et al. 2008) a hierarchical clustering algorithm is presented that can interpret and compare both numerical and categorical from a semantic perspective features of objects. In a similar approach to that used in the present study, ontologies are exploited as the base to map categorical features to semantically comparable concepts. The likenesses of the concepts are then assessed using semantic similarity measures. According to these similarities, an iterative

aggregation process of objects is performed based on Ward's method (Ward 1963). A hierarchical classification of non-overlapping sets of objects is therefore constructed from the evaluation of their individual features. The height of the internal nodes in the resulting dendogram reflects the distance between each pair of aggregated elements.

With this algorithm, and using WordNet as the background ontology, we evaluated the utility of data from the semantic clustering point of view. Fig. 15 shows the evaluation process. It consists on comparing the clusters obtained from the original dataset with those resulting from the execution of the clustering process, both for distributional (i.e. discernibility-based) and semantic (i.e. Wu and Palmer's similarity-based) anonymisation procedures. As result of the clustering process we obtained three different set of clusters:

- Set of clusters obtained from clustering of the original data.
- Set of clusters obtained from clustering of the anonymised data with the recoding method based on a distributional perspective.
- Set of clusters obtained from clustering of the anonymised data with the recoding method based on a semantic perspective.

A *k*-anonymity level of 5 was chosen for this comparison because it is a moderate privacy level that allows the retention of data utility.

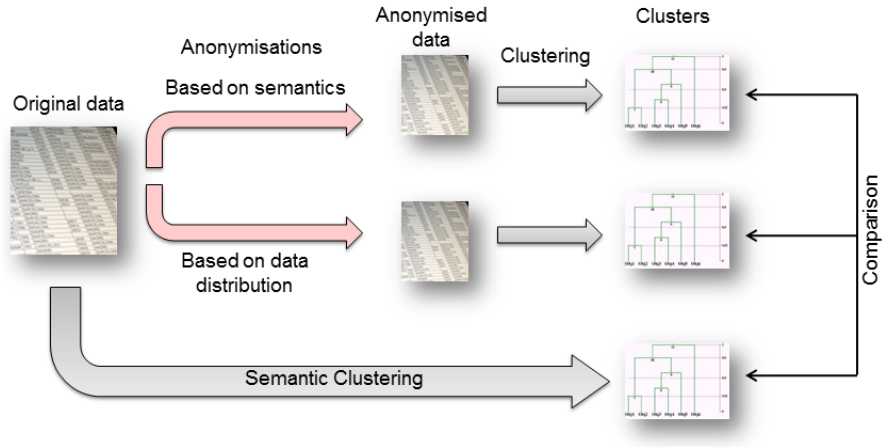

Fig. 15. Evaluation process of the data utility for data mining purposes

By quantifying the differences between the set of clusters (i.e. a partition) obtained from original data versus those obtained for both masking methods, we can determine which one best retains the semantics and, therefore, the utility of data. The resulting sets of clusters can be compared using the distance between partitions of the same set of objects defined in (De Mántaras 1991): considering two partitions of the same data set (in this case, the original and anonymised

versions), where P_A is a partition whose clusters are denoted as A_i, and P_B is a partition whose clusters are denoted as B_j, the distance is defined as:

$$d_{Part}(P_A, P_B) = \frac{2 * I(P_A \cap P_B) - I(P_A) - I(P_B)}{I(P_A \cap P_B)} \quad (5.3)$$

where $I(P_A)$ is the average information of P_A that measures the randomness of the distribution of elements over the n classes of the partition (similarly for $I(P_B)$), and $I(P_A \cap P_B)$ is the mutual average information of the intersection of two partitions. These are computed as

$$I(P_A) = -\sum_{i=1}^{n} P_i \log_2 P_i \quad (5.4)$$

$$I(P_B) = -\sum_{j=1}^{m} P_j \log_2 P_j \quad (5.5)$$

$$I(P_A \cap P_B) = -\sum_{i=1}^{n} \sum_{j=1}^{m} P_{ij} \log_2 P_{ij} \quad (5.6)$$

where the probabilities of belonging to the clusters are $P_i=P(A_i)$, $P_j=P(B_j)$, and $P_{ij}=P(A_i \cap B_i)$.

This distance evaluates whether the objects have been distributed in the same clusters when two different partitions (original and anonymised) are compared. Distance values are normalised in the [0..1] interval, where 0 indicates that both partitions have identical clusters and 1 indicates that the partitions are maximally different.

The distance between the original clusters and those obtained from both masking approaches are as follows.

Table 11. Distances between the clustering results

	Distance
Original data vs. Semantic anonymisation	0.26
Original data vs. Distributional anonymisation	0.57
Semantic vs. Distributional anonymisations	0.56

Table 11 shows how a semantically driven anonymisation produces a dataset that retains the semantics of the original data better than distributional approaches

(the distances in the resulting classification with respect to the original data are 0.26 and 0.57, respectively). Conclusions drawn from analysis of semantically anonymised data would therefore be more similar to those from the original data when the approach presented in this paper is used. As we stated in the introduction, this shows that semantics play an important role in the preservation of data utility. Note also the large differences between clusters resulting from each anonymisation schema, whose distance is a significant 0.56. This shows a high discrepancy in the way records are fused according to the different quality metrics. This result is consistent with that observed in section 4.2, where semantic and distributional anonymisations provided significantly different results.

5.2.5 Record linkage

Data utility is an important dimension when aiming to anonymise data and minimise information loss. From the point of view of privacy protection, however, disclosure risk should be also minimised. The latter can be measured as a function of the probability of reidentifying the masked dataset with respect to original data.

This dimension is commonly evaluated by means of *Record Linkage* (RL) as defined in (Torra,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:

$$RL = \frac{\sum_{i=1}^{m} P_{rl}(r_i^A)}{m} \cdot 100 \tag{5.7}$$

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

$$P_{rl}(r_i^A) = \begin{cases} 0 & if\ r_i \notin G \\ \frac{1}{|G|} & if\ r_i \in G \end{cases} \tag{5.8}$$

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.

We have calculated the record linkage percentage for different levels of *k*-anonymity comparing the original registers with respect to the semantic anonymisation and then with the distributional version of the method. The RL probabilities are illustrated in Fig. 16.

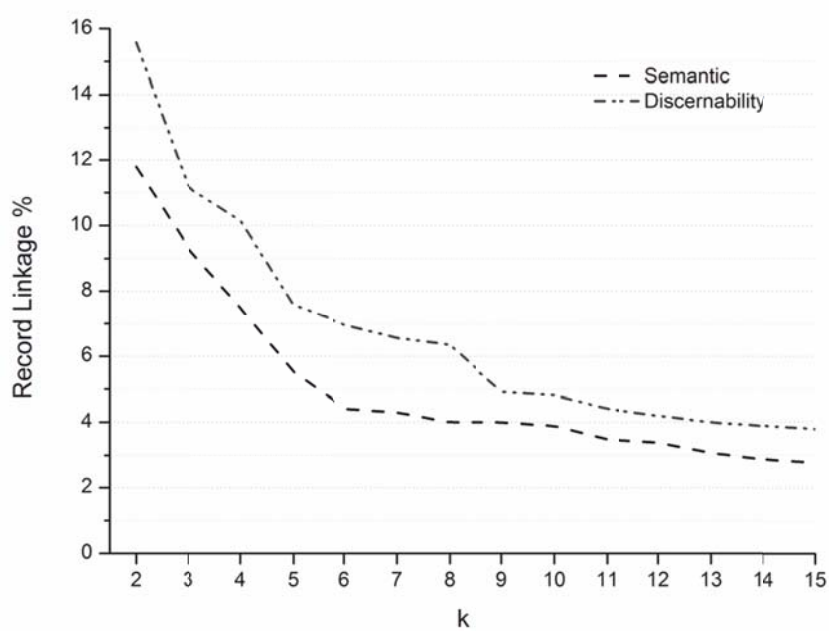

Fig. 16. Record Linkage percentage for semantic and dicernability-based anonymisations.

Both approaches follow a similar trend, i.e. RL probability decreases as *k* increases. We can also see that the degree of record linkage is quite stable for *k* values of 5 and over. The main difference is that our method gives lower probabilities of record re-identification than a distributional approach, especially for small values of *k*. Compared to the distributional approach, this allows the degree of *k*-anonymity to be lowered (resulting in less information loss) while a comparable level of disclosure risk is maintained.

In conclusion, these results show that an anonymisation process that is focused on the preservation of data semantics does not contradict the goal of a privacy preservation method, i.e. to minimise the risk of disclosure.

5.2.6 Execution time study

From a temporal perspective, executing our method over a 2.4 GHz Intel Core processor with 4 GB RAM, the run time of the anonymisation process for the test dataset ranged from 1.2 to 1.6 seconds (according to the desired level of *k*-anonymity) (see Fig. 17). The pre-calculation of the semantic similarities between all value pairs of each attribute in the dataset took 6.33 minutes.

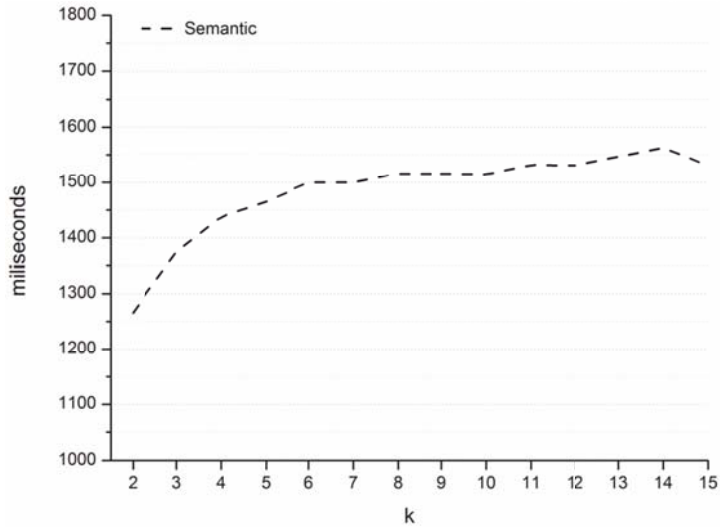

Fig. 17. Anonymisation process runtime according to the level of *k*-anonymity.

We can clearly see how, as stated in section 4.2, similarity computation is the most computationally expensive function and how minimising the number of calculations noticeably optimises runtime.

Run times are also much lower than those reported by related works that required several hours (Xu et al. 2006a), (Li,Li 2008) to perform the anonymisation of the data even for generalisation schemas, very limited VGHs and bounded categorical data (3-4 levels of hierarchical depth and an average of a dozen values (Li,Li 2008)). In contrast, we were able to mask much bigger and fine grained data in much less time while considering large and wide ontologies such as WordNet, with thousands of concepts and a maximum depth of 16 levels (as explained in section 3.1.1). This shows the scalability of our method for large and heterogeneous databases.

5.3 Summary

In this chapter, we applied the semantic framework proposed in chapter 4 to develop a new recoding method. This global masking method is based on the exploitation of wide and general ontologies in order to properly find the most appropriate substitution on each record from a conceptual point of view rather than from a symbolic one. The algorithm uses heuristics to guide the search on the set of possible value substitutions towards the preservation of the semantics of the dataset. This has been demonstrated with several tests conducted with real textual data from visitors to a Catalan National Park. Our results indicate that, compared with a classical approach based on optimisation of the distribution of the data, ours retains the quality and utility of data better from a semantic point of view. This was illustrated when we exploited masked data using a clustering process. The partitions generated from the original dataset and the anonymised data are more similar with our semantic method than with classical approaches.

Finally, we have taken special care to ensure the applicability and scalability of the method when dealing with large and heterogeneous data. By enabling the exploitation of already available ontologies, we avoid the need to construct tailor-made hierarchies according to data labels such as VGH-based schemas, which suppose a high cost and limit the method's applicability. Moreover, the non-exhaustive heuristic algorithm based on constrained value substitutions achieved a good scalability with regard to the size, heterogeneity and number of attributes of input data and to the size, depth and branching factor of the ontology.

The main publications related to contributions presented in this chapter are:

3J. Martinez, S., Sanchez, D., Valls, A., Batet, M.: Privacy protection of textual attributes through a semantic-based masking method. International Journal: Information Fusion 13(4), 304-314 (2011). *Impact Factor*: 1.467

1C. Martinez, S., Sanchez, D., Valls, A.: Ontology-Based Anonymization of Categorical Values. In: Torra, V., Narukawa, Y., Daumas, M. (eds.) Modeling Decisions for Artificial Intelligence, vol. 6408. Lecture Notes in Computer Science, pp. 243-254. Springer Berlin / Heidelberg, (2010). CORE B

2C. Martinez, S., Sanchez, D., Valls, A., Batet, M.: The Role of Ontologies in the Anonymization of Textual Variables. In: the 13th International Conference of the Catalan Association for Artificial Intelligence 2010, pp. 153-162.

3C. Martínez, S., Valls, A., Sánchez, D.: Anonymizing Categorical Data with a Recoding Method Based on Semantic Similarity. In: Hüllermeier, E., Kruse, R., Hoffmann, F. (eds.) Information Processing and Management of Uncertainty in Knowledge-Based Systems. Applications, vol. 81. Communications in Computer and Information Science, pp. 602-611. Springer Berlin Heidelberg, (2010).

Chapter 6
A new semantic microaggregation anonymisation method

Among the plethora of anonymisation methods, *microaggregation* stands as a natural approach to satisfy the *k*-anonymity property in statistical databases (Domingo-Ferrer,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. Because the search for the optimal partition when considering multivariate data is NP-hard (Oganian,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 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 thesis, 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.

6.1 Microaggregation methods

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. 2010a). In fact, microaggregation methods satisfy the *k*-anonymity property (Domingo-Ferrer,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,Mateo-Sanz 2002; Lin et al. 2010b; Torra,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,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,Wei 2008), authors propose different sub-optimal heuristics to perform this task. In (Chiu,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,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,Shao 2007). From these, the MDAV (*Maximum Distance Average Vector*) method (Hundepool et al. 2003), on which this work 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.

6.1.1 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. Since clusters 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 (Hundepool et al. 2003; Domingo-Ferrer,Torra 2005). 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,Torra 2005; Erola et al. 2010; Huang et al. 2010a; Lin et al. 2010b; Nin et al. 2008a; Domingo-Ferrer et al. 2008).

The behaviour of the MDAV method is detailed in algorithm 3. 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 3. 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
```

```
    Form a cluster in D^A with s and the k-1 closest
    records to s
    Remove these records from D
end while
if (|D| ≥ 2*k) then
    Compute the centroid x̄ of remaining records in D;
    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
```

6.1.2 Limitations of MDAV when dealing with categorical data

The basic MDAV method, however, presents some limitations 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 information loss resulting from the replacement of the individual records by the cluster centroid (Domingo-Ferrer,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 example job or city of living because, due to their discrete nature, modalities tend to repeat and, hence, it would be desirable to put as many repetitions as possible into the same cluster to maximise its cohesion.

Fig. 18 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. 18-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. 18-B. This second clustering seems much more natural, with more homogeneous clusters.

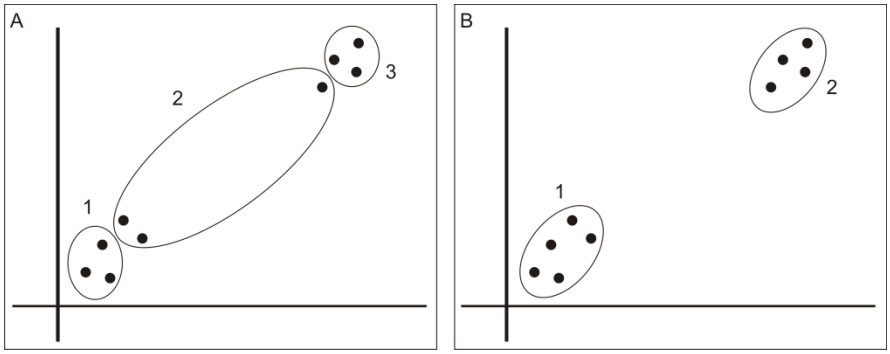

Fig. 18. A comparative example of microaggregation with $k = 3$. **A.** Fixed-sized clustering. **B.** Variable-sized clustering.

Some authors have proposed modifications of the MDAV algorithm to support variable-sized clusters (Domingo-Ferrer,Mateo-Sanz 2002; Laszlo,Mukherjee 2005; Lin et al. 2010b). 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 3), 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 3). 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.

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 (Torra 2004; Domingo-Ferrer,Torra 2005). 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.

As repeatedly argued in this thesis, and shown in previous sections, 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).

Recently, some authors have considered the semantics of categorical data during the microaggregation process. In (Abril et al. 2010), the MDAV algorithm is applied to categorical attributes, computing the distance between records using the Wu & Palmer similarity measure (Wu,Palmer 1994)(see Eq. 3.3) and WordNet (Fellbaum 1998)(see section 3.1.1) as the ontology. The Wu & Palmer measure evaluates the similarity between two concepts (c_1 and c_2). The measure ranges from 1 for identical concepts to 0. Hence, this similarity measure is converted into a distance function as follows:

$$dis_{w\&p}(c_1, c_2) = 1 - sim_{w\&p}(c_1, c_2) \tag{6.1}$$

where $sim_{w\&p}$ is the wu & Palmer semantic similarity measure described in section 3.2.1 (Eq. 3.3).

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 Least Common Subsumer (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,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 (Torra 2004; Domingo-Ferrer,Torra 2005), 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.

In conclusion, current works present the following limitations:

- Size of clusters is too constrained
- Similarity measure is too simple or even non semantic.
- A centroid based on the root node or LCS is too poor to represent the cluster.

6.2 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 the MDAV algorithm and the application of the proposed semantic framework (section 3.3) for 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 6.1.2, due to the discrete nature of categorical data, it would be desirable to consider their distribution to create cohesive clusters. In section 6.2.1, 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 distance: considering that textual data should be interpreted according to their underlying semantics (Torra 2011), we propose to use the comparison operator (section 4.2) to guide 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 6.2, the construction of an appropriate and representative centroid is crucial to guide the microaggregation process and to minimise the information loss. We propose to use the aggregation operator (section 4.3) to calculate the centroid of multivariate categorical datasets, exploiting a knowledge base as well as considering the data distribution.

6.2.1 Adaptive microaggregation

As stated in section 6.1.2, 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 (Domingo-Ferrer,Torra 2005; Abril et al. 2010) or *2k-1*, for variable-sized ones (Domingo-Ferrer,Mateo-Sanz 2002; Lin et al. 2010b). 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,Tsai 2007; Lin,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 stated in section 3.3.1 as follow:

Remind that a multivariate dataset *MV* with *p* indistinguishable records and *m* attributes is represented 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 6. 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 6, Fig. 19 shows the advantage, compared with related works, of 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. 19-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. 19-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 $\omega_i > 2k\text{-}1$, 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. 19-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.

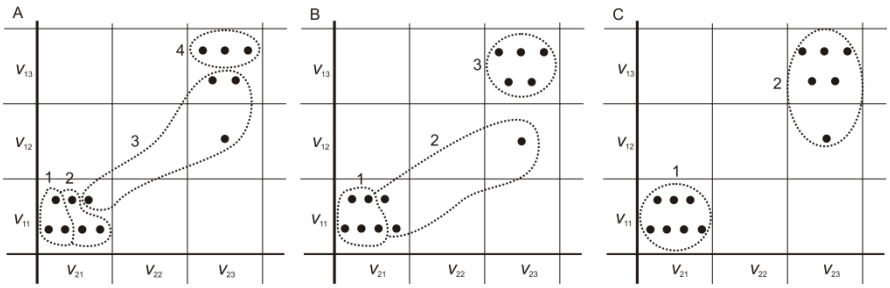

Fig. 19. An example of microaggregation with $k = 3$. **A.** Fixed-sized clustering. **B.** Variable-sized clustering with a maximum size of *2k-1*. **C.** Adaptive clustering without maximum size restriction.

To incorporate this adaptive behaviour during the clustering construction, the MDAV method, shown in Algorithm 3, has been modified as shown in Algorithm 4, highlighting in **bold** the steps in which the operators that we later adapted to the semantic case. We called the algorithm *Semantic Adaptive MDAV* (SA-MDAV).

Algorithm 4. 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 ($|D| \geq k$) do
3 **Compute the centroid** $\bar{X}$ of all tuples in D;
4 Select the **most distant** tuple r to the centroid $\bar{X}$
5 Form a cluster C in D^A with the tuple r. **Calculate centroid** $\bar{C}$
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. Calcu-
       late 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 3), 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.

The semantic-based operators defined in chapter 4 are used (as marked in bold), whenever a centroid needs to be obtained (either of the whole dataset, line 3, or of a particular cluster, lines 5, 10, 14 and 18), the aggregation operator in section 4.3 can be applied. Since clusters are built according to the selected centroid and records are replaced by cluster centroids, the fact that centroids minimise the accumulated semantic distances of the aggregated terms helps to minimise the information loss. Again, the comparison operator presented in section 4.2, can be used to obtain the most semantically distant record *r* to the dataset centroid (line 4, where the reference value is the centroid), the most semantically distant record to *r* (line 13, where *r* is the reference value and line), the least semantically distant records to build a cluster around *r* or *s* (lines 8 and 16, where the reference values are the cluster centroid) and the least semantically distant centroid for the remaining records (line 22). Since our operator considers all record value repetitions at once, identical records can be clustered together, obtaining more cohesive groups.

Note also that, as formalised above, input data is managed according to distinct value tuples with associated frequencies of appearance. Considering the algorithm design (without the centroid calculation cost, that is considered in evaluation section 6.3.4), 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.

6.3 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 6.3.1 and 6.3.2, 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 6.3.3. In section 6.3.4, 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.

6.3.1 The datasets

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. 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 (see Table 12) used for evaluation tests.

Table 12. 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

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". This dataset has been analysed in the section 5.2.1 as part of the evaluation of the recoding method. See section 5.2.1 for details.

Dataset 2 is the well-known *Adult Census* (Hettich,Bay 1999), which is publicly available in the UCI repository[3] and has been often used in the past for evaluating privacy-preserving methods (Iyengar 2002; Fung et al. 2005; Lin,Wei 2008; Domingo-Ferrer et al. 2006; Lin et al. 2010b). 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 12. For evaluation purposes, we have used the training set consisting on 30,162 records, after removing rows with missing values. Fig. 20 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 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.

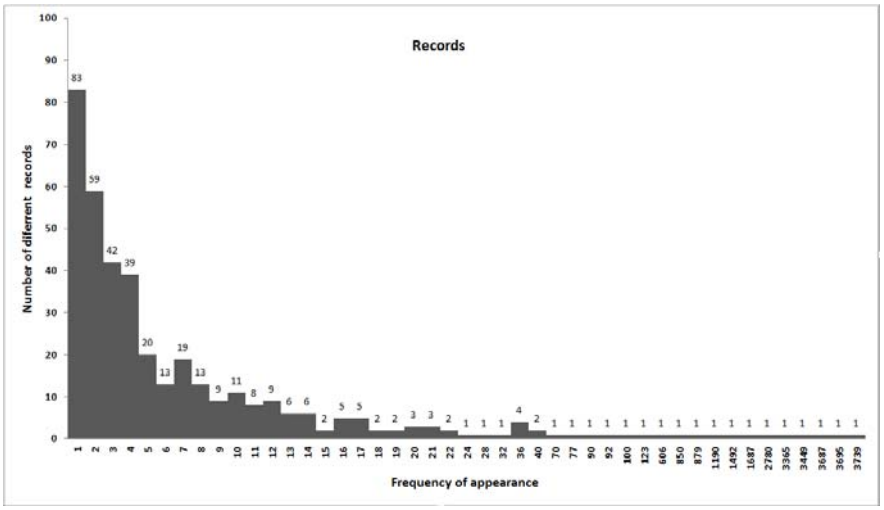

Fig. 20. The Dataset 2 frequency distribution of distinct value tuples

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

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 (Loukides,Shao 2007; He,Naughton 2009; Domingo-Ferrer et al. 2006; Abril et al. 2010) 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,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.

6.3.2 Evaluation measures

As stated in section 6.1, 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,Mateo-Sanz 2002; Lin et al. 2010b; Torra,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.

$$SSE = \sum_{i=1}^{g} \sum_{j=1}^{n_i} \left(\sum_{l=1}^{m} \frac{dis_{w\&p}(x_{ijl}, \overline{x}_{il})}{m} \right)^2 \tag{6.2}$$

where g is the number of clusters, n_i is the number of elements in the ith cluster, x_{ijl} is the value of the l attribute of the jth element of the cluster i and $\overline{x}_{il}$ denotes the value of the l attribute of the centroid of the cluster ith. As a distance measure, we employ $dis_{w\&p}$ defined in Eq. 3.3, 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:

$$SST = \sum_{i=1}^{g} \sum_{j=1}^{n_i} \left(\sum_{l=1}^{m} \frac{dis_{w\&p}(x_{ij}, \overline{x}_{l})}{m} \right)^2 \tag{6.3}$$

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

$$L = \frac{SSE}{SST} \times 100 \tag{6.4}$$

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. To evaluate the disclosure risk we have used the Record Linkage (RL) measure described in section 5.2.5 as part of the evaluation of recoding method. See section 5.2.5 for details.

In related works to microaggregation, sometimes a *balance score* between the information loss and the disclosure risk, is calculated as the weighted average of these two complementary measures (Eq. 6.5). 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; Torra 2004; Domingo-Ferrer,Torra 2001b; 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 low information loss and low record linkage, de-identification. So, we have also made an analysis of this score for this anonymisation method.

$$score = \alpha L + (1 - \alpha) RL \tag{6.5}$$

6.3.3 Analysis of SA-MDAV

As stated in Section 6.2, 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 6.2. To focus only on the algorithmic differences between MDAV and SA-MDAV, in all the three settings, the distance between records was calculated with the comparison operator proposed in section 4.2 (Eq. 4.2) and the selection of centroids was done with the aggregation operator proposed in Section 4.3 (Eq. 4.6). The different versions are:

- *S-MDAV (Semantic MDAV)*: The dataset is microaggregated using the basic MDAV (Algorithm 3, in Section 6.1.1). 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 6.2. The difference with the classical MDAV relies on the use of the semantic operator 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 6.2. 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 6.3.2 was computed; results are shown in Fig. 21.

Results shown in Fig. 21 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.

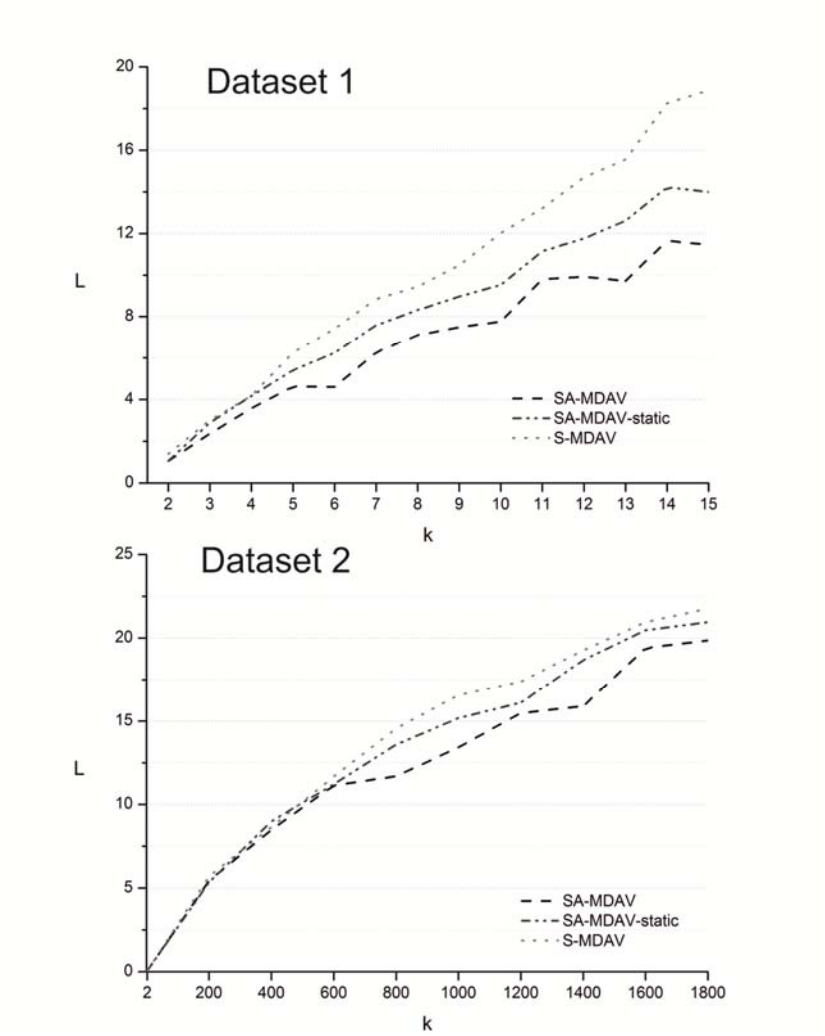

Fig. 21. A comparison of the three algorithm versions according to Information Loss (L)

Secondly, when cluster size is adapted to data distribution (SA-MDAV vs. S-MDAV), the minimisation of information loss (*L*, Fig. 21) 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. 21. 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.

6.3.4 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,Torra 2005), as detailed in Section 6.1.2, 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. 3.3) 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. 22), disclosure risk (RL, as shown in Fig. 24), quality score (Fig. 25 and Fig. 26) and also runtime (Fig. 27).

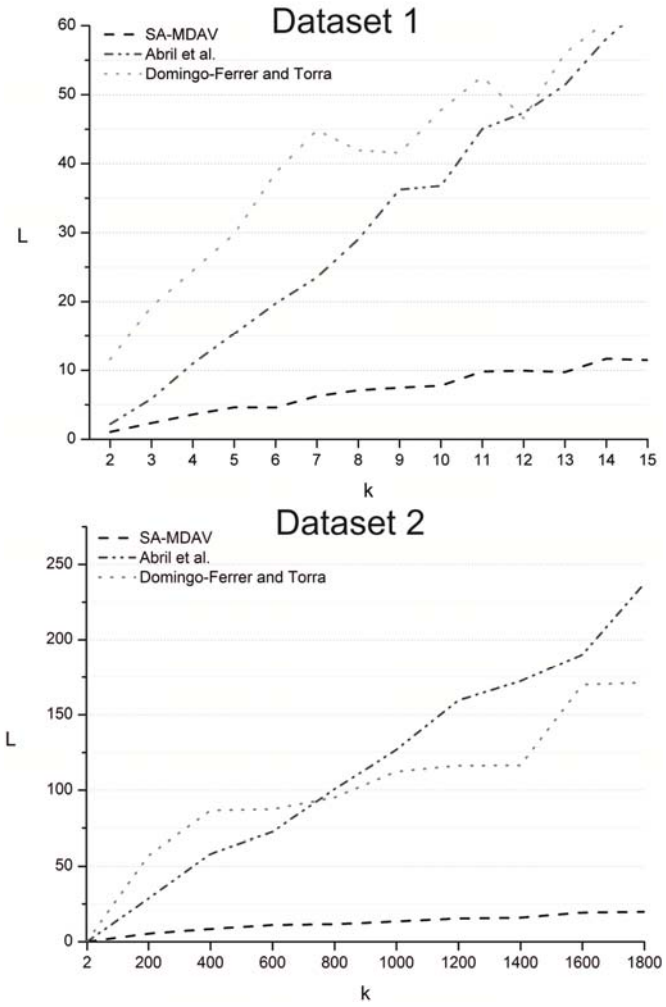

Fig. 22. A comparison of Information Loss (L) values for the evaluated methods

With regards to the information loss metric (Fig. 22), 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. 23 we show the knowledge structure evaluated from WordNet when computing the first initial centroid, first execution of line 3 in Algorithm 4. 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 4.3, 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.

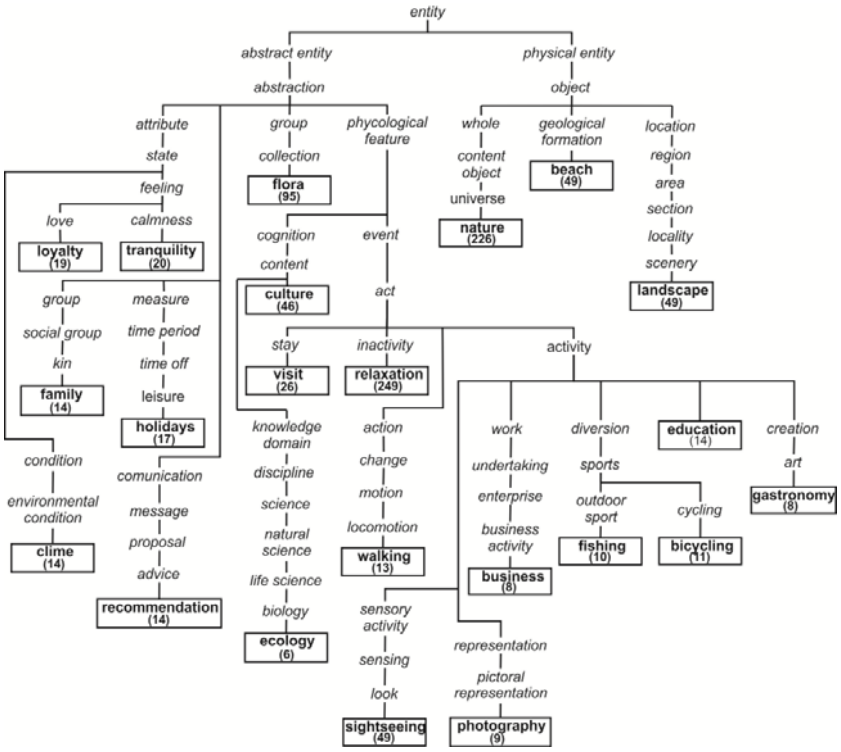

Fig. 23. The knowledge structure extracted from WordNet for input values of an attribute of Dataset 1 (in bold). The numbers in brackets represent the amount of appearances of each value.

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. 24) 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 matching 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.

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 4.3. Centroids aim to minimise the sum of semantic distances between all the aggregated values, considering also their distribution, as a weighting factor. This process generates centroids that can be either more new general concept 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).

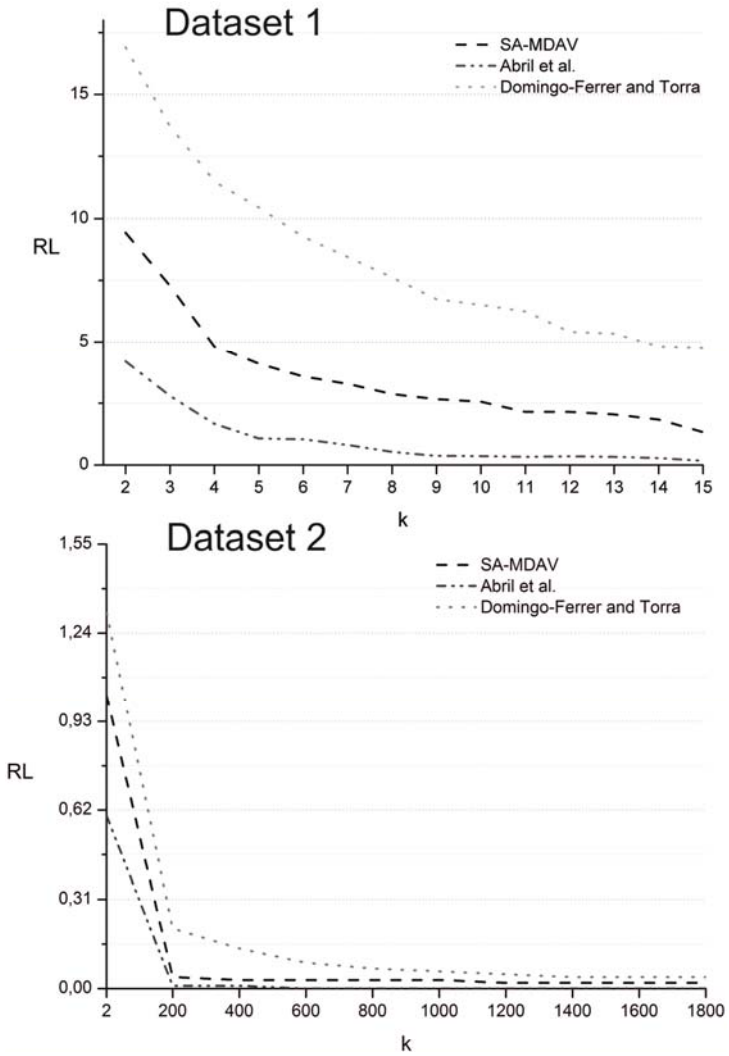

Fig. 24. A comparison of RL percentage for the evaluated methods.

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. 6.5). First, as shown in Fig. 25, 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; Torra 2004; Domingo-Ferrer,Torra 2001b; Yancey et al. 2002).

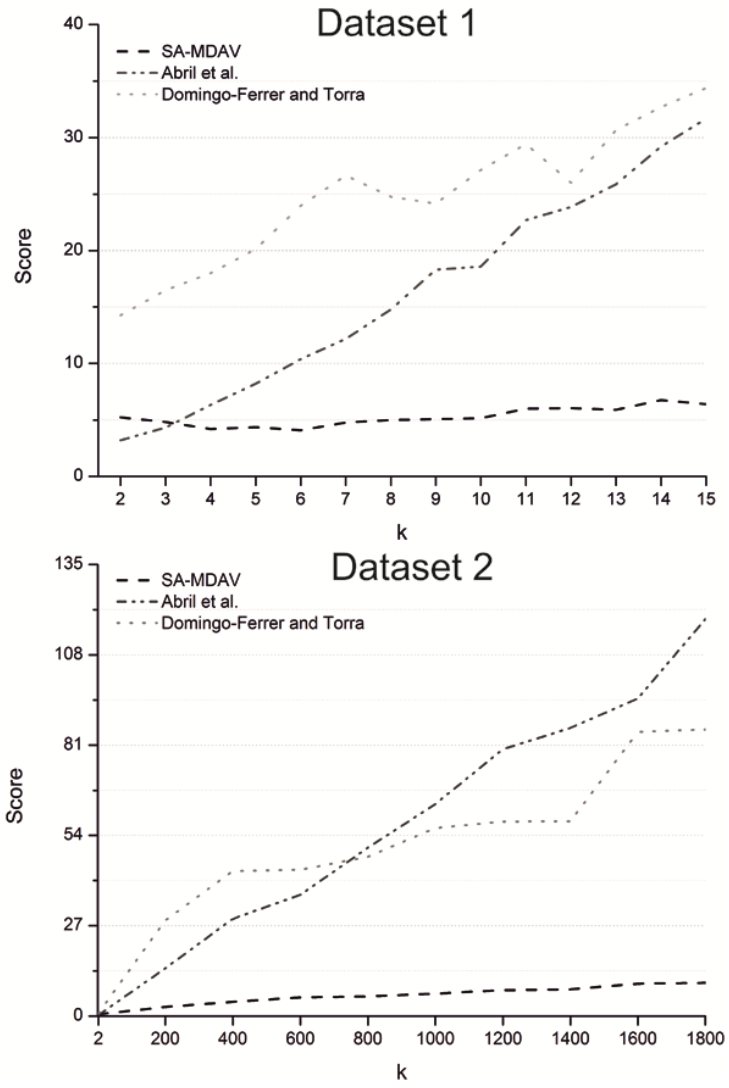

Fig. 25. Score with an equilibrated balance (α=0.5) between information loss and disclosure risk.

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 (Eq. 6.5) 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. 26, 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. 24.

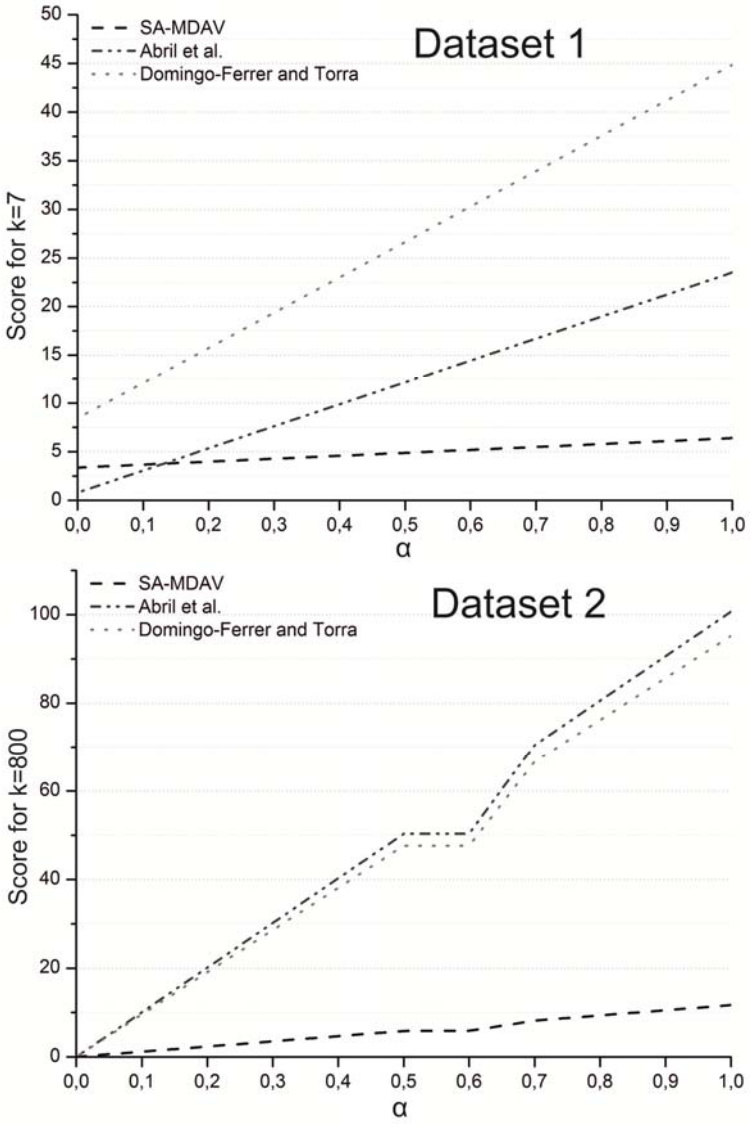

Fig. 26. Score values when varying the relative weight between information loss and disclosure risk.

Finally, an analysis of the execution time has been done. To understand the runtime results shown in Fig. 27, 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 6.1.1, 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 4.3. Furthermore, as stated in section 6.2.1, 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. 12 and Fig. 20, and on the number of centroid candidates, which depends on the ontology.

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.

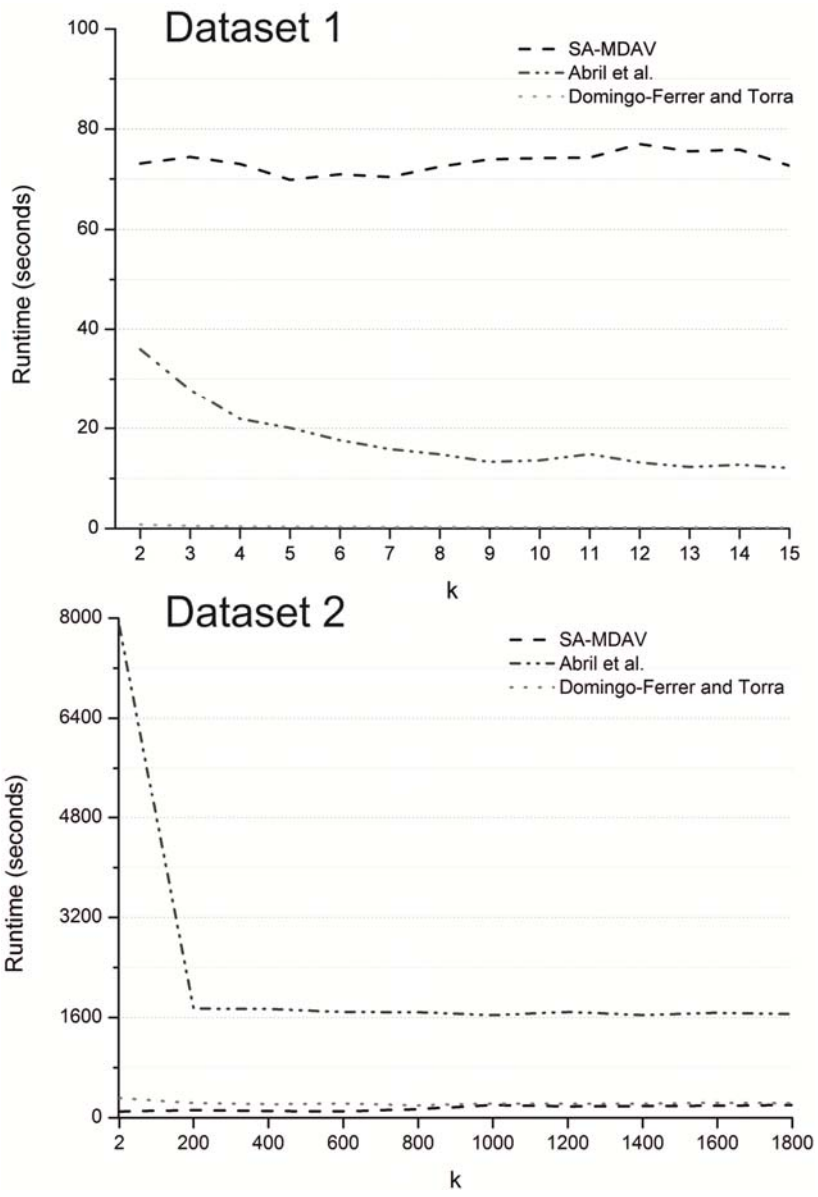

Fig. 27. A runtime comparison of the evaluated methods.

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.

6.4 Summary

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,Mateo-Sanz 2002; Domingo-Ferrer,Torra 2005; Torra 2004; Lin et al. 2010b). 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 5.2, 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 defining the SA-MDAV method.

The proposal relies on a hierarchical knowledge structure that represents the taxonomical relations of the values that appear in the dataset. The proposal focuses on the following aspects:

- Adaptive microaggregation. SA-MDAV creates clusters of different size according to the data distribution.
- Semantic distance. SA-MDAV uses the semantic comparison operator to guide the cluster construction process.
- Semantic centroid. SA-MDAV uses the semantic aggregation operator to calculate the centroid of multivariate categorical datasets.

The evaluation, performed over two different datasets with categorical 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 SA-MDAV, even though being heuristic and subject to sub-optimal choices to preserve its scalability, improves related works by a considerable margin, both when considering the absolute information loss and also when evaluating the balance between information loss and disclosure risk. 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.

The main publications related to contributions presented in this chapter are:

4J. Martínez, S., Sánchez, D., Valls, A.: Semantic Adaptive Microaggregation of Categorical Microdata. International Journal: Computers & Security 31(5), 653-672 (2012). *Impact Factor*: 0.868

Chapter 7
A new semantic resampling method

In this chapter, it is proposed an anonymisation method based on a classic data re-sampling algorithm for numerical data. Applying the semantic framework proposed in chapter 3, the re-sampling algorithm is adapted to be able to deal with non-numerical data from a semantic perspective. Moreover, we introduced some changes on the sampling process in order to ensure the k-anonymity property.

This chapter is focused on Resampling (Heer 1993), a method that, even though has not gained as much research attention as other masking algorithms like Microaggregation (Domingo-Ferrer,Mateo-Sanz 2002), it has demonstrated to be fast, which is an interesting feature when dealing with large-scale datasets and to retain a high utility of data for numerical attributes with respect to other methods (Herranz et al. 2010b; Karr et al. 2006). In a nutshell, Resampling is based on taking *t* independent samples from the original attributes, sorting them and finally building a new masked record as the average of the ranked values of the *t* samples. Even though this method reduces the disclosure risk, it does not ensure the *k*-anonymity property on the masked dataset.

The application of this method to numerical attributes (e.g. salaries, outcomes) is straightforward by using some ordering criterion on numbers and then calculating the arithmetic average of the ranked values of the different samples. However, it is not used for categorical data due to the lack of appropriate aggregation and sorting operators which requires a semantic interpretation of their meaning (e.g. sports like *swimming* or *sailing* are more similar than *swimming* and *trekking*) as it is argued in this thesis.

In this chapter is presented a resampling method that is able to deal with non-numerical attributes from a semantic perspective, applying appropriate sorting and averaging operators defined in our proposed framework (see chapter 3). As far as we know, no other semantically-grounded masking resampling methods have been proposed in the literature. Moreover, to guarantee a theoretical level of privacy, the resampling method will be designed so that it fulfils the well-known *k*-anonymity property (on the contrary to classic resampling methods (Herranz et al. 2010b; Domingo-Ferrer,Mateo-Sanz 1999)).

7.1 The original resampling method

Resampling was defined in (Jones,Adam 1989) as a method for perturbation of output data that consists on modifying only the answers to some queries while leaving the original data unchanged. Later, in Heer (Heer 1993) resampling was applied to anonymise contingency tables based on bootstrapping. A cell-oriented version resampling procedure was proposed later in (Domingo-Ferrer,Mateo-Sanz 1999). However, in this case the masked data file generated does not rely on resampling the original data, but on generating binomial random perturbations that preserve some statistics (obtaining an asymptotically unbiased estimate of the original data). Resampling has been also applied to anonymise input data (i.e. original records of the published dataset). In (Herranz et al. 2010b) a distributed version of the resampling method is proposed for dealing with the case where the original data is partitioned among several providers. All these methods focus on numerical data.

In this thesis, as done in related works (Domingo-Ferrer,Mateo-Sanz 1999; Herranz et al. 2010b), we base our method in the Heer's resampling approach (detailed in Algorithm 5 and Fig. 28). Briefly, being *n* the number of records in the dataset, the method takes *t* samples with replacement (i.e. values can be taken more than once). Each sample is sorted in increasing order. Then, the masked dataset is obtained by taking, as first value, the average of the first values of all samples, as second value, the average of the second values of all samples, and so on.

Algorithm 5. Resampling

Inputs: D (original dataset), t (number of samples)

Output: D^A (a masked version of D)

1 Take t samples $S_1,\ldots,S_t$ of n records of D (with replacement)

2 Sort each sample S_i in $S_1,\ldots,S_t$

3 Make a set P_i with t elements with the records at the i-th position of the sorted samples

4 Add the average of each P_i to D^A

5 output D^A

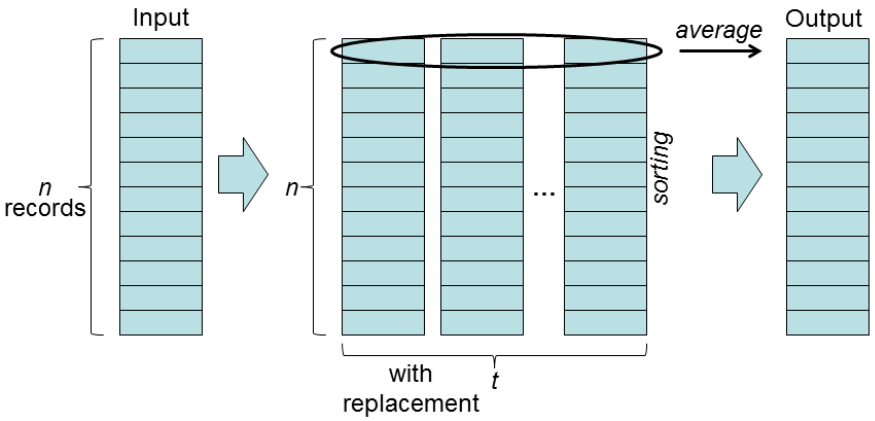

Fig. 28. Resampling process

Comparison studies (Herranz et al. 2010b; Karr et al. 2006) show that Heer's resampling achieves a high utility score with respect to other masking techniques, but with a slightly higher disclosure risk. This is related to the fact that, unlike other masking methods (Domingo-Ferrer,Mateo-Sanz 2002; Abril et al. 2010; Torra 2004), the Heer's approach was designed without considering the *k*-anonymity property (formalized years later in (Samarati,Sweeney 1998)). Hence, resampled results cannot guarantee a –theoretical- level of privacy.

A new version of resampling has been designed, in to tackle this issue. The new resampling method fulfils the *k*-anonymity property while it is also able to deal with non-numerical data from a semantic perspective. Two issues not considered in previous works (Herranz et al. 2010b; Domingo-Ferrer,Mateo-Sanz 1999).

7.2 Semantic *k*-anonymous resampling for categorical data

This section proposes a new resampling method (named S*k*Resampling) focused on minimizing the information loss when masking categorical data while ensuring the fulfilment of the k-anonymity property. It is based on the Heer's resampling method (Heer 1993) with the following modifications:

- *K-anonymous resampling*: the original sampling method has been modified (as detailed in section 6.3.1) so that masked records fulfil the k-anonymity property.
- *Semantic resampling of categorical data*: in order to semantically interpret non-numerical data during the resampling process, we have

applied the semantic framework described in chapter 3. Concretely, it has been used the comparison, aggregation and sorting operators.

7.2.1 k-anonymous resampling

To guarantee the k-anonymity on the masked dataset we modified Heer's method (Algorithm 5) as detailed in Algorithm 6. Let D be the input dataset, k is the desired level of k-anonymity and n is the number of records in D.

Algorithm 6. Resampling

Inputs: D (original dataset), k (level of anonymity)

Output: D^A (a transformation of D that fulfils the k-anonymity)

1 take k samples of n/k records of D

2 take the (n mod k) remaining records

3 **sort** each sample

4 make a set P_i with k elements with the records at the i-th position of the sorted samples

5 generate the **centroid** of each set P_i

6 add the (n mod k) remaining records to the set with the **least distant centroid**

7 recalculate the **centroid** of the modified set

8 replace each record in the sets by its **centroid** and add it to D^A

9 output D^A

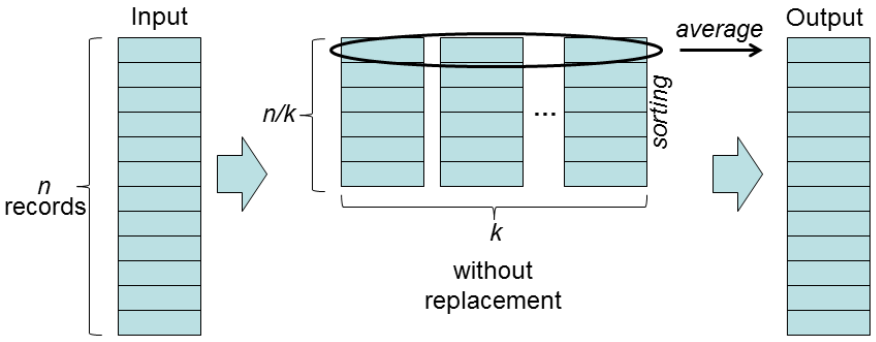

Fig. 29. *k*-anonymous resampling process.

The algorithm proceeds as follows (see algorithm 6 and Fig. 29). First, to guarantee the *k*-anonymity property, it is required to create *k* samples, so that *k* records can be replaced by their average in a later stage and become *k*-indistinguishable. To do so, *k* samples of *n*/*k* records are created (line 1), without replacement (i.e. each record is taken only once). It is important to avoid record repetitions to ensure that all records fulfil the k-anonymity. Then these samples are sorted (line 3) and sets P_i are created with the records at the i^{th} position of all samples (line 4). The idea is that, by sorting samples, similar records are put together at similar positions of different samples. Finally, the anonymised dataset is obtained by replacing all records of each P_i by the centroid of P_i (line 8). Since, by definition, P_i contains at least *k* records, this process generates a *k*-anonymous dataset. Note that, when taking *n/k* records for sample, the remaining *n mod k* records (line 2) should be also treated. These records are added to the set with the closest centroid (line 6), recalculating its centroid (line 7) before the replacement (line 8).

In this case, the sorting operator proposed in section 4.4 can be applied to line 3 so that records are arranged according to their semantic similarity. The reference value to perform the sorting process is the centroid of each sample. Again, the proposed centroid procedure in section 4.3 can be applied to compute semantically coherent centroids of each set (lines 5 and 7), and the comparison operator proposed in section 4.2 can be used to select the least semantically distant centroid for the remaining records (line 6).

7.3 Evaluation

In this section, the evaluation of the proposed semantic and *k*-anonymous *Sk*Resampling method is detailed. As evaluation data, we used the well-known *Adult Census* dataset (Hettich,Bay 1999). This dataset has been analysed in the section 6.3.1 as part of the evaluation of the microaggregation method. Two non-numerical attributes have been taken, corresponding to “occupation” (14 distinct

modalities) and "native-country" (41 modalities). This gives 30,162 records (after removing rows with missing values) distributed into 388 different responses, 83 of them being unique (Fig. 20). Attribute values have been mapped to WordNet (Fellbaum 1998), which has been used as the knowledge base that enables the semantic interpretation of data.

Because the theoretical level of privacy is guaranteed by the fulfilment of the *k*-anonymity property, the evaluation has been focused on the quantification of the information loss. To do so, we have employed the *sum of square errors* (SSE), which has been extensively used to evaluate anonymisation methods (Domingo-Ferrer,Mateo-Sanz 2002; Abril et al. 2010). It is defined as the sum square of distances between each original record and their corresponding masked version (Eq. 6.2). Hence, the higher the SSE is, the higher the information loss.

To evaluate the benefits that the semantic S*k*Resampling approach brings with regards to the preservation of the data utility in comparison with non-semantic methods, we configured four settings that, while based on the same resampling algorithm (Algorithm 4), they vary the criteria used to compare, sort and replace the original values:

- Setting 1: no semantics are considered when masking the dataset. On one hand, values are compared using the equality predicate (0 if identical, 1 otherwise). On the other hand, the centroid is selected as the mode (i.e., the most frequently occurring value in the set). This mimics the behaviour of anonymising approaches dealing with non-numerical data in a categorical fashion (Torra 2004; Domingo-Ferrer,Mateo-Sanz 2002).
- Setting 2: semantics are considered when comparing values. The Wu & Palmer measure (Eq. 6.1) applied to WordNet as knowledge structure are used. The centroid is kept the same as in setting 1. Comparing this to setting 1, one can quantify the influence in information loss of the distance calculation method.
- Setting 3: the same as setting 2, but taking as centroid the concept in WordNet that taxonomically subsumes all elements in the set (i.e. the Least Common Subsumer -LCS-, as done in (Abril et al. 2010)). This approach considers the semantics of data in the centroid calculation but neglect their distribution in the taxonomical tree.
- Setting 4 (S*k*Resampling): the method presented in section 7.2.1 is applied. It shows the contribution of our proposal on the information loss.

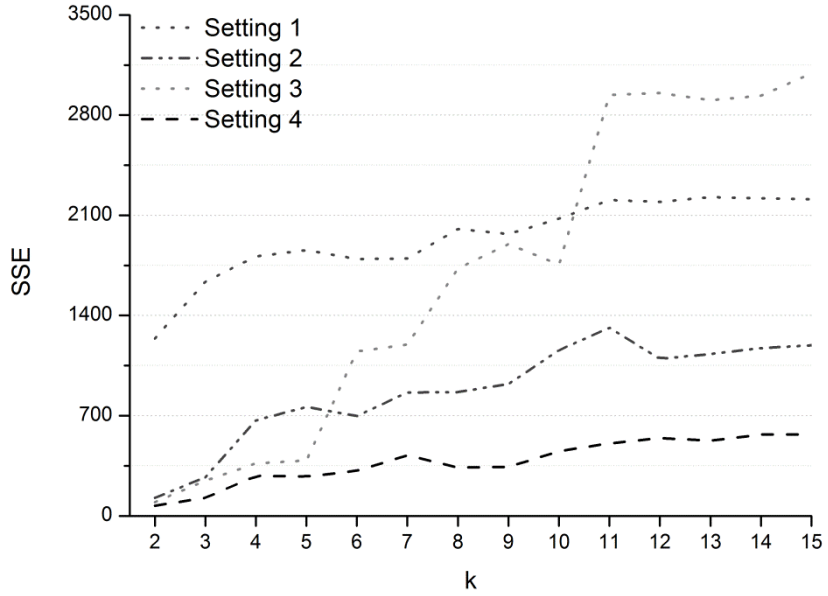

Fig. 30. Comparison of Information Loss (SSE) values for the evaluated settings.

The information loss values (i.e. *SEE*) obtained in the four settings while varying the *k*-anonymity level from 2 to 15 are shown in Fig. 30.

First of all we can see that our approach (*setting 4*) achieves the lowest information loss for all the *k*-anonymity values. Compared to other settings, we observe that a non-semantic method (*setting 1*) results in the highest information loss for most *k*-values. This shows that managing and transforming non-numerical data only by means of their distribution features (i.e. repetitions, mode) hardly preserves the meaning of masked data and hence their utility. By considering semantics in distance calculations (*setting 2*), information loss is significantly reduced showing the limitations of comparing attribute values according to the equality/inequality of their textual labels. Regarding the results obtained by *setting 3* (i.e. centroid computed as the LCS), we observe that, even though it is able to retain more information than non-sematic centroid calculation methods (i.e. mode) for low *k*-values, it produces severely worse results when a higher anonymity degree is required. The fact that the LCS is used as the centroid causes that very general abstraction are picked when a larger and more heterogeneous set of values should be generalized (i.e. higher *k*-anonymity). In comparison, our method (*setting 4*) produces more accurate centroids, by considering both the semantics and the data distribution.

7.4 Summary

In this chapter, we have adapted a classic resampling method to always fulfil the *k*-anonymity property, in order to guarantee a certain level of privacy. Compared to other methods, resampling is faster since the sampling process is done randomly, which makes it especially suitable for very large datasets. By contrast, this randomness may negatively influence the information loss.

While considering the semantics of the data by means of the application of semantically-grounded comparison, sorting and averaging operators, we obtain an effective method for anonymising non-numerical data. Several authors have shown the benefits that a semantic treatment of data can bring with regards to the preservation of the utility of masked data (Torra 2011; Martinez et al. 2011). The evaluation results sustain our hypothesis, being able to minimise the information loss in comparison with non-semantic approaches.

The main publications related to contributions presented in this chapter are:

4C. Martínez, S., Sánchez, D., Valls, A.: Towards k-anonymous non-numerical data via Semantic Resampling. In: Information Processing and Management of Uncertainty in Knowledge-Based Systems, Catania, Italy 2012, pp. 519-528. S. Greco et al. (Eds.).

Chapter 8
A new Semantic Record Linkage method

Record linkage methods evaluate the disclosure risk of revealing confidential information in anonymised datasets that are publicly distributed. Concretely, they measure the capacity of an intruder to link records in the original dataset with those in the masked one. In the past, masking and record linkage methods have been developed focused on numerical or ordinal data. In the previous chapters, we have used the traditional record linkage method based on matching in the evaluation stage. However, this approach does not take into account the distance of the values from a semantic perspective. In this chapter, we propose a new record linkage method specially tailored to accurately evaluate the disclosure risk for categorical attributes, from a semantic point of view. Our method, named Semantic Record Linkage, relies on ontologies to interpret the semantics of data and propose coherent record linkages.

8.1 Introduction

The goal of a privacy-preserving method is avoiding that an intruder re-identifies the identity of an individual from the published data, associating his confidential information. A typical way to achieve some degree of anonymity is to satisfy the *k*-anonymity property (see section 2.2).

Ideally, any masking method should minimize the information loss to maximize data utility (Martínez et al.). *Information loss* is a quality measure of the reduction of the utility in the masked data, when compared to the original one.

However, analysing only the Information Loss is not enough. Another important aspect of any masking method is the minimization of the *disclosure risk*. It measures the capacity of an intruder to obtain the information contained in the original dataset from the masked one (Domingo-Ferrer,Torra 2001a). To compute the disclosure risk, many works (Domingo-Ferrer,Torra 2001a; Torra,Domingo-Ferrer 2003; Winkler 2004; Murillo et al. 2012) consider *record linkage* (RL) methods. These try to link the records in the original dataset with those of in the masked one. Two kinds of record linkage methods are usually considered in the literature(Nin et al. 2008b). On the one hand, *distance-based*

record linkage computes a distance measure between original and masked records, linking each masked one to the closest in the original dataset. For numerical data, an Euclidean distance is typically used. On the other hand*, probabilistic record linkage* bases the matching on the expectation-maximization algorithm (Mclachlan,Krishnan 1997) which is based on the amount of coincidences between masked and original datasets.

Classical RL methods have been defined independently of the masking method used to anonymise input data. However, some works (Nin et al. 2008a; Medrano-Gracia et al. 2007; Torra,Miyamoto 2004; Nin et al. 2008b) have shown that it is possible to increase the amount of linkages by designing tailored RL methods for concrete masking schemas. In (Torra,Domingo-Ferrer 2003; Nin et al. 2008a, b) authors show that ad-hoc designed RL methods increase the disclosure risk when assuming that input data have been anonymised by means of a micro-aggregation process. In (Nin et al. 2008b) a similar work is proposed, in which an especially designed RL method increases the amount of linkages when input data is masked by with rank swapping (Dalenius,Reiss 1982). Using especially tailored RL methods one assumes the worst possible scenario for privacy protection and, hence, better evaluates the potential disclosure risk.

Both generic and ad-hoc RL methods proposed in the literature are focused only on numerical and ordinal data. However, as stated above, several masking methods for categorical attributes have been proposed in recent years.

In this chapter, we present a new distance-based RL method designed to measure the disclosure risk of masking methods based on the generalization of categorical attributes. Our method (called *Semantic Record Linkage,* SRL) relies on the semantic operators proposed in our framework (chapter 4) to propose linkages between original and masked datasets, discovering the most semantically similar records. This supposes an improvement over methods based solely on the number of term coincidences. Considering that the knowledge structure used by the masking method to propose generalizations remains hidden to the intruder, we use general-purpose taxonomies/ontologies to better interpret categorical values (Yang et al. 2009; Cheng et al. 2010).

SRL has been applied to evaluate the disclosure risk of a classical generalization method applied to a real dataset of categorical data, and it has been compared against a non-semantic RL approach relying on counting term coincidences. We have evaluated SRL versus related works based on VGH in order to better analyse the repercussion that the use of different dimensioned VGH has during the anonymisation process in the disclosure risk of the masked datasets. Results show that a semantically-grounded RL method increases the risk of re-identification compared to existing methods and, hence, it better evaluates the potential disclosure risk of masked data.

In the section 2.5.3, we reviewed works proposing anonymisation schemas focused on categorical attributes. By analysing and understanding their behaviour, we will be able to propose a specially tailored RL method that evaluates better the potential risk of disclosure of these methods.

8.2 Enabling semantically-grounded record linkage

A Record Linkage method, especially when tailored for a specific masking method, can be seen as a reverse engineering process, in which an intruder tries to guess and undo the data transformations performed during the anonymisation process. In the case of masking methods dealing with categorical data, two elements influence how the anonymisation is performed: the underlying knowledge structure used to propose generalizations, and the quality criteria used to decide the one that minimizes the information loss. Obviously, both elements are variables that remain hidden to the intruder. In consequence, the RL method should either guess them from input data or substitute them by other elements that are general enough to be applicable even when the masking criteria vary.

Classic generalization methods use ad-hoc taxonomical structures constructed according to input attribute labels to propose value generalizations. To undo generalizations, an accurate RL requires a similar knowledge base. Considering that the design and structure of the VGH depends on the way in which the anonymiser structured the knowledge, it is neither feasible nor scalable to guess the VGH. Instead, we use available knowledge structures that aim to be general enough to cover most of the concepts that may appear in a domain: ontologies. Ontologies are formal and machine-readable structures, representing a shared conceptualization of a knowledge domain, expressed by means of semantic relationships. They have been successfully applied in many areas dealing with textual resources (Sánchez et al. 2010) and knowledge management (Valls et al. 2010). Ontologies present several advantages compared with VGHs. Widely used ontologies provide a taxonomical structure much larger and finer grained than VGHs, being created from the consensus of a community of knowledge experts. They represent knowledge in an objective, coherent and detailed manner. This contrasts with the ad-hoc, overspecified and coarse nature of VGHs which can be hardly assessed.

With such ontologies, attribute values (i.e. words) presented in input datasets can be mapped to ontological nodes (i.e. concepts) via simple word-concept label matching. In this manner, the hierarchical tree to which each textual value belongs can be explored to retrieve possible generalizations and/or specializations that can assist the RL process. From a domain independent point of view, one can use a general ontology like WordNet, as it has been used in the previous chapters of this thesis.

Once we have selected the knowledge source in which the RL will rely, it is necessary to define a criterion to match records between the masked and original datasets. Distance-based RL methods define a measure by means of which the closest records are matched. This measure should be as similar as possible to the quality metric used to anonymise data. In the case of generalization methods, one can assume that the anonymiser has selected the generalization that minimizes the

information loss. From a semantic point of view, information loss is a function of the difference between the degree of generality of the original and masked values. So it can be seen as a measure of semantic alikeness. On the contrary to related works (Domingo-Ferrer,Torra 2001a; Torra,Domingo-Ferrer 2003; Winkler 2004; Yancey et al. 2002) focused on numerical and ordinal data, which evaluate textual values in a categorical way, we rely on the semantic similarity (see chapter 3) to properly compute the semantic distance between textual labels and guide the RL process.

8.3 New Semantic RL method for VHG schemas

In this section, we propose a new record linkage method for categorical attributes relying on a semantic interpretation of the values through the semantic operators proposed in the framework (Chapter 4). The method, named Semantic Record Linkage (SRL), is designed for dealing with anonymisation schemas when dealing with categorical.

Starting from a dataset in which each record corresponds to an individual, let us consider the typical anonymisation scenario used in works like (Nin et al. 2008b; Domingo-Ferrer,Torra 2001b) consisting on:

1. Identifier attributes (*e.g.* ID-card numbers) have been removed from the dataset.
2. If an attribute is considered confidential (*e.g.* salary) then it is not modified.
3. The anonymisation is applied to quasi-identifier non-confidential attributes (*e.g.* job, city-of-living, personal preferences).

The resulting dataset D consists on n records, each of them composed by m quasi-identifier non-confidential attributes and c confidential attributes. Let us have that D^A is the publishable and, therefore, anonymised version of D, containing n records with m anonymised quasi-identifier non-confidential attributes and the initial c confidential attributes. Let us consider that an intruder gathers information about the set of individuals in D, and builds a dataset E that contain the same m non-confidential quasi-identifiers that appear in D, together with some identifier attributes. Assuming that some (or all) of the records in E correspond to individuals that are also in D, the intruder can access confidential data (*e.g.* salary) if he is able to link a record $r_k^E \in E$ with the anonymised (and published) record $r_i^A \in D^A$, so that r_k^E and r_i^A correspond to the same individual, disclosing his identity. This can be achieved by using the common non-confidential attributes in E and D^A; that is $E \cap D^A$. The amount of correct record linkages evaluates the disclosure risk of the privacy-preserving method.

According to this scenario, the proposed *Semantic Record Linkage* method (SRL) can be applied to the set of quasi-identifier non-confidential attributes if they consist on categorical values. As stated in section 7.2, the method relies on ontologies to assess the semantic similarity between textual values. The linkage is done calculating the minimum distance between the values that the intruder knows (*i.e.* the textual attributes in *E*) and the anonymised attributes published (*i.e.* the textual attributes in D^A obtained by a generalization process from the original values in *D*). The linkage method is formalized as follows.

Let us have that *D* is composed by *n* records, $r_i = \left(r_{i1},...,r_{im}\right)$, D^A consist on the same number of anonymised records, $r_i^A = \left(r_{i1}^A,...,r_{im}^A\right)$, and *E*, owned by the intruder, has some records $r_k^E = \left(r_{k1}^E,...,r_{km}^E\right)$, where r_{ij} and r_{ij}^A and r_{kj}^E are categorical values. Then we define the set of linked records (*LR*) with respect to each r_k as:

$$LR_{r_k^E} = \left\{ l \;\middle|\; l = \underset{\forall i=1..n, r_i^A \in D^A}{argmin} \left(sd_O\left(r_k, r_i^A\right)\right) \right\} \qquad 8.1$$

The intruder searches for the least distant record to r_k in D^A. Because the result may be non-unique (*i.e.* equally similar records), we obtain a set of linked records *LR*.

The SRL method relies on the measurement of the semantic distance between the categorical values that appear in each record in order to estimate their alikeness. Considering that a generalization-based masking method tries to minimize the information loss by suggesting the closest subsumer that satisfies the *k*-anonymity (see section 7.2), our SRL method hypotheses that the semantically closest record in *D* for an anonymised one $r_i^A \in D^A$ should be r_i (*i.e.* the original version of r_i^A). The similarity between two records r_i and r_k is defined as the arithmetic average of the semantic similarity between each of their attribute values as follows:

$$sd\left(r_i, r_k\right) = \frac{\sum_{j=1}^{m} sd\left(r_{ij}, r_{kj}\right)}{m} \qquad (8.2)$$

where the function *sd* corresponds to any of the semantic distance measures presented in section 3.2.

Now, the *disclosure risk* (DR) of a privacy-preserving method can be measured as the difficulty in finding correct linkages between original and masked datasets. This is done by counting the amount of correct linkages that the intruder is able to perform between E and D^A. DR is evaluated for the worst possible case, assuming that E contains all the m records of D^A and all the n non-confidential quasi-identifier attributes (Torra,Domingo-Ferrer 2003). DR is calculated as the percentage of the average probability of linking each record r_k in D^A denoted as $p_{D^A}(r_k)$, as follows.

$$DR = \frac{\sum_{k=1}^{n} p_{D^A}(r_k)}{n} \cdot 100 \tag{8.3}$$

where $p_{D^A}(r_k)$ is the probability of making a correct linkage calculated as follows:

$$p_{D^A}(r_k) = \begin{cases} 0 & if\ r_k^A \notin LR \\ \dfrac{1}{|LR|} & if\ r_k^A \in LR \end{cases} \tag{8.4}$$

being $LR \subset D^A$ the set of records with minimum distance with respect to each record r_k, (Eq. 8.1) and assuming that $r_k{}^A$ in D^A and r_k correspond to the same individual.

8.4 Evaluation

In this section, we test the behaviour of the SRL method in the evaluation of the disclosure risk of generalization schemas dealing with categorical data, comparing it to a non-semantic approach relying on the matching of textual labels (as previous works).

8.4.1 Evaluation data

The dataset used for evaluation consists on a set of real answers to polls made by the "Observatori de la Fundació d'Estudis Turístics Costa Daurada" at the Catalan National Park "Delta de l'Ebre". This dataset has been analysed in the section

5.2.1 as part of the evaluation of the recoding method. The two categorical attributes available in the dataset, columns "reason for visiting the park" and "main activities during the visit to the park" have been considered as non-confidential quasi-identifiers (two last columns in Table 13), so they will be anonymised and used to perform the record linkage afterwards.

8.4.2 Masking method

To evaluate the SRL method, we have implemented a generalization algorithm that aims to depict the classic generalisation methods discussed in section 2.5.3. Due to the size of the data used during the evaluation, we opted by a non-exhaustive method based on a best-first search strategy (similar to (Martinez et al. 2011; Li,Li 2008; He,Naughton 2009; Martinez et al. 2010a)). To reproduce the best scenario from the data utility point of view, we used a quality measure that quantifies the number of generalization steps performed at each transformation (like in (Samarati,Sweeney 1998; Sweeney 2002b), as discussed in section 2). It is important to note that, on the contrary to simpler approaches (Sweeney 2002b; Samarati,Sweeney 1998; Bayardo,Agrawal 2005; Iyengar 2002), the search space of the implemented algorithm is not constrained, and each value can be changed by any of the concepts that generalize it. This configures a more realistic but also challenging scenario.

Both the best-first search algorithm and the quality measure rely on a hierarchical structure that defines the possible generalizations for each value found in the dataset. In the same manner as the methods described in section 7.2, we have constructed ad-hoc VGHs. For this dataset, 25 distinct terms appear in the two attributes considered. In consequence, 25 leaves taxonomically connected through generalization concepts are contained in the VGH. To evaluate the influence of the VGH design, we have constructed two different VGHs. The first one (Fig. 31, denoted as VGH2), incorporates up to two levels of generalization. The second on (Fig. 32, named VGH3) models a finer grain classification.

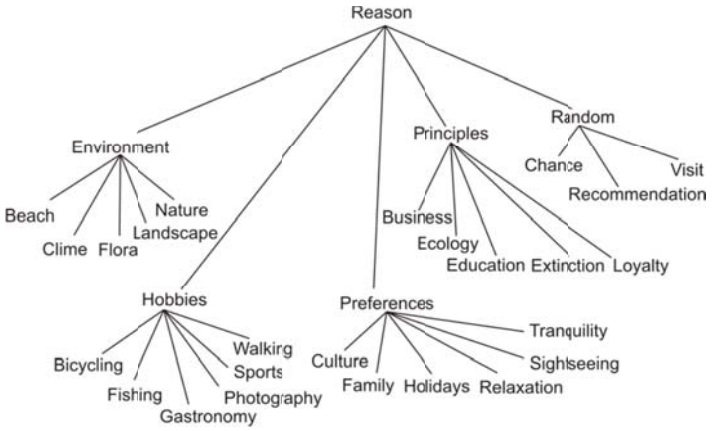

Fig. 31. VGH2, modeling up to two levels of generalization per label.

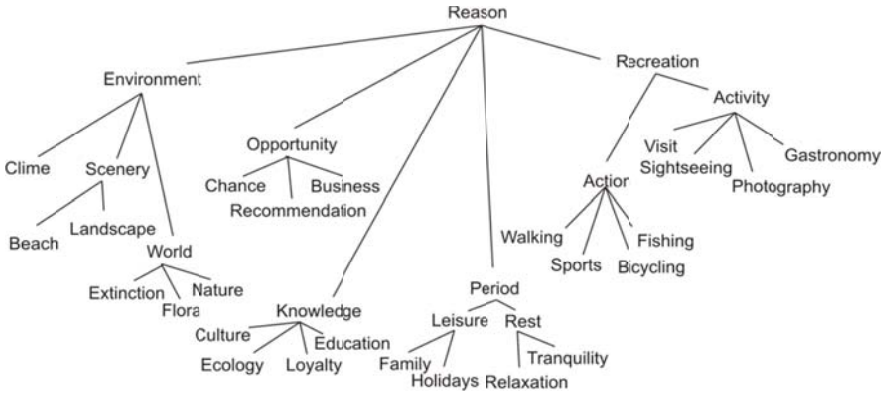

Fig. 32. VGH3, modelling up to three levels of generalization per label.

It is important to note that, in addition to the quality metric, the design of the VGH used to assist the masking process has a direct influence in the information loss and their utility from a semantic point of view. One may construct a VGH that progressively propose fine grained generalizations for attribute labels (e.g. sailing -> water sport -> sport -> activity). In this case, each generalization produces a lower loss of semantics than coarser taxonomical structures (e.g. sailing -> activity). The disclosure risk in detailed VGHs, however, may increase because the generalizations are less abstract and can be more easily linked with the original labels. So, there is a trade-off between information loss and disclosure risk: when one decreases, the other tend to increase. Finding the equilibrium is a difficult task that should be carefully considered.

8.4.3 Evaluation of RL

The results obtained from the anonymisation have been evaluated by means of our SRL method, using WordNet (see section 3.1.1) as ontology and three of the semantic distance measures introduced in section 3.2 as the criteria to propose linkages. Concretely we used as semantic distance between two concepts c_1 and c_2 (in Eq. 8.2) concretely: dis_{pL} (Eq. 3.1), $sim_{w\&p}$ (Eq. 3.3) and dis_{logSC} (Eq. 3.9). Note that the Wu & Palmer measure evaluates the similarity between two concepts (c_1 and c_2). The measure ranges from 1 for identical concepts to 0. Hence, this similarity measure is converted into a distance function as:

$$dis_{w\&p}(c_1, c_2) = 1 - sim_{w\&p}(c_1, c_2) \tag{8.5}$$

where $sim_{w\&p}$ is the wu & Palmer semantic similarity measure described in Eq. 3.3.

To test the adequacy of the SRL we have compared it against a non-semantic implementation of RL (named *Matching-based Record Linkage*, MRL). The MRL method represents the expected behaviour of record linkage without background knowledge and dealing with textual data in a categorical fashion, like in (Domingo-Ferrer,Torra 2001a; Torra,Domingo-Ferrer 2003; Winkler 2004; Yancey et al. 2002). In this case, to build the set of linked records *LR* (as in Eq. 8.1), the record similarity can only be based on the terminological matching of textual labels. It searches for records with exactly the same values in E and D^A and assigns them a maximum similarity value. Formally, the record similarity is:

$$dis_{MRL}(r_i, r_k) = \begin{cases} 1 & \text{if } r_i = r_k \\ 0 & \text{if } r_i \neq r_k \end{cases} \tag{8.6}$$

8.4.4 Results

The first study regards to the results obtained when using different semantic similarity measures in comparison to a non-semantic approach (MRL). Fig. 33 shows the evaluation of the disclosure risk (definition 4.3) of the generalization method for k-anonymity values from 2 to 20. For the SRL method, the three semantic similarity measures introduced in section 3 have been used. On the left, it is shown the percentage of correct record linkages obtained with a dataset masked using VGH2 while on the right the results when using a more detailed knowledge structure, VGH3 are given.

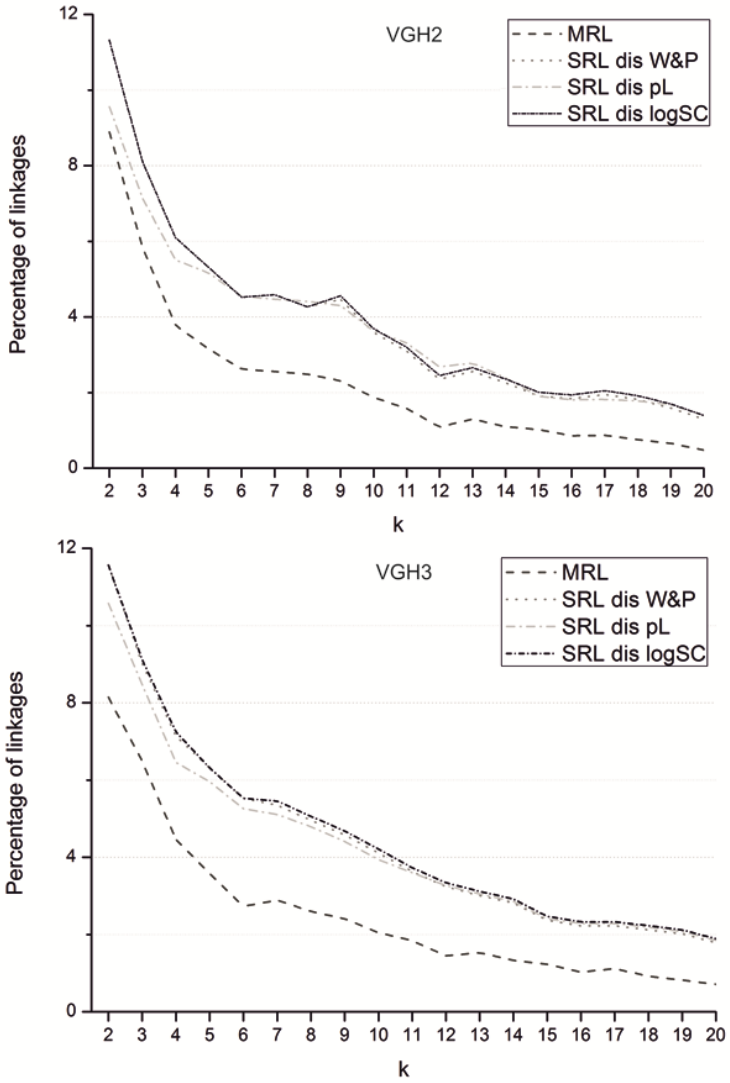

Fig. 33. Disclosure risk evaluated by means of SRL and MRL methods, using VGH2 and VGH3 as the knowledge base.

Several conclusions can be extracted. First, the proposed SRL method is able to improve the amount of correct linkages proposed by the non-semantic approach (MRL). The differences are more evident for values of *k* from 4 to 8, because larger *k*-anonymity levels imply a higher loss of information. In this interval, the amount of correct linkages obtained with SRL almost double those achieved by the MRL approach. Moreover, the decreasing of the number of linkages as the *k*-value increases is coherent to that what it is represented in the distribution of the

dataset (most of the records have a number of repetitions between 1 and 5, see section 4.3). This explains the abrupt decrease in the number of linkages for same range of *k*-values. It is interesting to note that, regardless the *k*-value, the SRL method will always outperform the MRL counterpart. In fact, for *k*-values higher than the number of maximum repetitions of any record (118 in our case), the number of linkages obtained by the MRL method will be zero, due to all the labels in the masked dataset will be generalized. The SRL method, on the contrary, will always propose record linkages with a probability of correct linkage depending on the number of total records.

Regarding the SRL method, the differences when using each semantic distance measure are minor, even though the approach by Wu and Palmer (Wu,Palmer 1994) and logSC (Batet et al. 2011) provided a slightly higher amount of linkages. This is coherent to what was evaluated in (Batet et al. 2011), in which former measures improved the similarity assessment accuracy of path-based ones by a considerable margin, when compared with human ratings of term similarity. In our case, semantic similarity measures are only used to rank pairs of terms and select the most similar ones. Results in (Batet et al. 2011) showed that, even though the similarity assessment of each measure may be different, the relative order of the resulting raking is quite similar. Consequently, the selection of the semantic similarity function does not have a noticeable effect in the results.

The second analysis studies the differences in the disclosure risk when using VGH2 or VGH3. More correct record linkages are obtained when using the most detailed knowledge structure. Fig. 34 shows the increment (in percentage) of correct linkages of the SRL method with respect to the basic MRL in both cases. We quantify among a 10-25% improvement in the amount of record linkages when using the SRL method applied to the dataset masked according to VGH3. It is worth to note that when using a more detailed knowledge base to guide the anonymisation process (VGH3), the masked values are more similar to the original ones, due to the lower level of abstraction introduced by the generalization process. In consequence, a RL method that is able to evaluate this semantic difference reveals a higher disclosure risk. On the contrary, a non-semantic RL approach obtains similar disclosure risk because, in both cases (VGH2 and 3), the original labels have been changed.

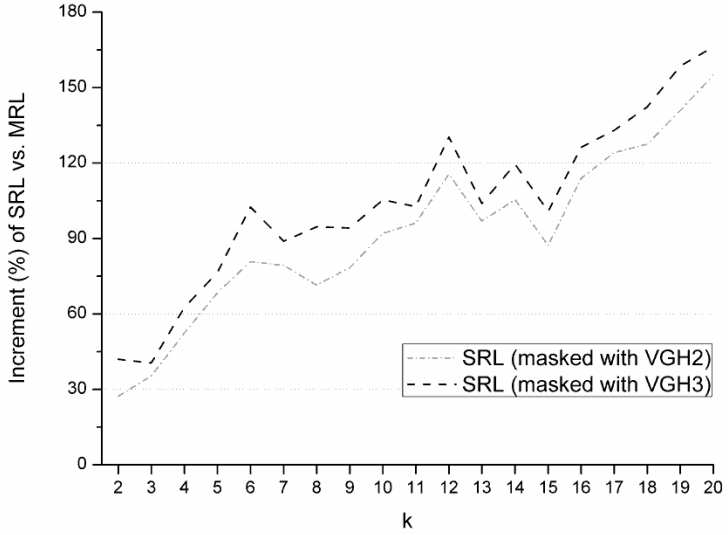

Fig. 34. Increment in the amount of correct linkages of SRL (using LogSC distance measure) with respect to the MRL, masking data according to VGH2 and VGH3.

Finally, we can see that the difference in percentage between the SRL for both VGH2 and VGH3 (with respect to MRL) is maintained in the range 10-25% along *k* values, stating that this difference is independent of the level of privacy. The relative difference between SRL and MRL, on the contrary, increases significantly as does the *k*-value. One may conclude that, from the point of view of minimizing the risk of disclosure, one should use simpler hierarchies of concepts (with few levels of generalization), due to the higher level of abstraction of the values. However, as stated in the introduction, anonymisation methods should also maximize the utility of data, minimizing the information loss.

The third analysis studies the information loss when using knowledge structures with different levels of detail, measuring how semantically similar the masked records are with respect to the original ones. The *semantic information loss* of D^A with respect to D has been computed as the semantic distance between the original and anonymised datasets as follows:

$$information\_loss_D\left(D^A\right) = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m} dis_{LogSC}\left(r_{ij}, r_{ij}^A\right)}{n*m} \tag{8.7}$$

where dis_{LogSC} has been computed as defined in Eq. 3.9, following a similar criteria as the one used to guide the anonymisation process.

Again, WordNet has been used as ontology, enabling an objective comparison of the semantic differences when using each VGH. Fig. 35 shows the evolution of information loss for each VGH, according to the *k*-anonymity level.

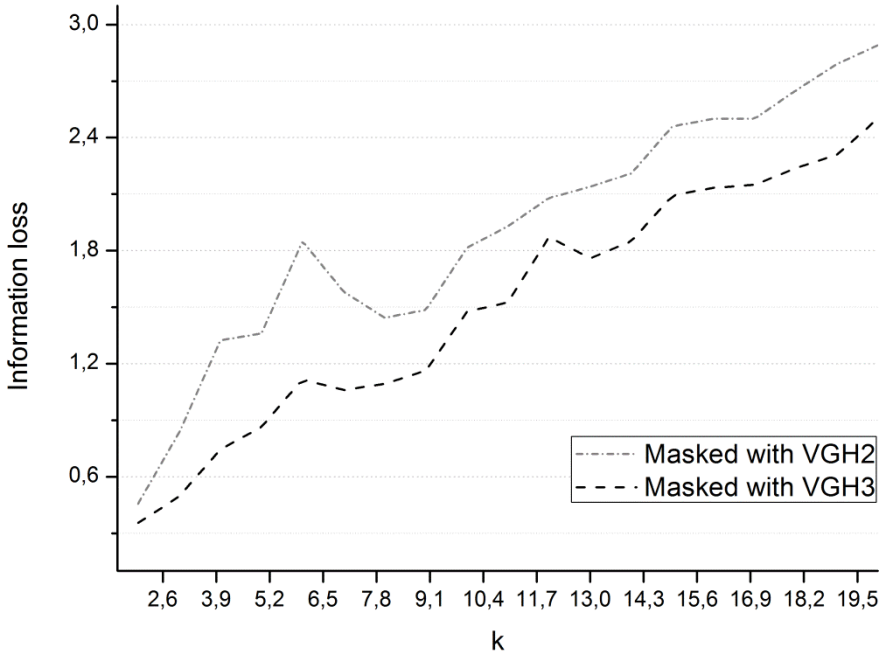

Fig. 35. Information loss according to the type of VGH used during the anonymisation.

On one hand, we observe a higher information loss when a simpler VGH is used. Hierarchies with fewer nodes produce more abstract masked values as a result of each generalization step. When compared to a detailed ontology like WordNet, the semantic distance of the masked data is higher. In consequence, data utility will decrease because it is related to the preservation of the semantics of textual values (Martinez et al. 2010c) . It is also worth noting that even though using a detailed knowledge base to guide the anonymisation process is desirable from the data utility point of view, the fact that larger and finer grain generalizations are available also increases the search space of possible value transformations. Due to the algorithmic design of generalization methods, the use of a source as large as WordNet is not feasible. The search space of possible generalizations for each value would be so high that even methods based on heuristic searches will not scale with large amounts of data(Martinez et al. 2010a; Martinez et al. 2011).

On the other hand, we notice a linear trend in the increasing of information loss according to the level of *k*-anonymity. Fig. 35 shows that the use of more detailed knowledge structures (such as VGH3) decreases information loss. Notice that the results obtained show an opposite trend with respect to the ones obtained when

evaluating the disclosure risk (Fig. 34). This indicates that there is a trade-off between the preservation of data utility and the disclosure risk. The differences in the curve shapes (almost linear for information loss vs. inverted log for disclosure risk) suggest that it is not convenient to protect data with high *k*-anonymity values, because the consequent loss in data utility will be comparatively higher than the decrease in the disclosure risk.

8.5 Summary

The work presented in this chapter proposes a new record linkage method based on semantic similarity theory and using ontologies to evaluate masking methods based on categorical attribute values. It has been evaluated for the case of masking with generalisation, using different knowledge base structures (VGH). Results show the convenience of using semantically-grounded RL methods compared to non-semantic algorithms. The tests have also gone a step further, evaluating the influence of the knowledge bases both in the information loss and in the disclosure risk.

The main contributions presented in this chapter are published in:

5J. Martínez, S., Sánchez, D., valls, A.: Evaluation of the Disclosure Risk of Masking methods Dealing with Textual Attributes. International Journal of Innovative Computing, Information & Control 8(7(A)), 4869-4882 (2012).

5C. Martínez, S., Valls, A., Sánchez, D.: An ontology-based record linkage method for textual microdata. In, vol. 232. Frontiers in Artificial Intelligence and Applications, pp. 130-139. (2011).

Chapter 9
A comparative study in the medical domain

After presenting three new anonymisation methods based on semantic knowledge provided by ontologies, this chapter applies those methods in real data from the Medical domain. In particular, we address the problem of protecting the privacy of patients when datasets of Electronic Health Records (EHR) are given to third parties, who will perform some kind of data analysis. This kind of datasets contains confidential information regarding clinical outcomes, such as diseases and treatments, which should not be disclosed. Since those terms are particular of the medical field, we will use a standard medical ontology to support the anonymisation processes.

The chapter first explains the dataset used for the analysis and the interest of applying masking methods in this particular kind of data. Afterwards, the semantic recoding, microaggregation and resampling methods are applied and a comparison of the results is done. Evaluation takes into account the information loss and the risk of reidentification from a semantic perspective, as well as the computational cost of the three methods.

9.1 Application scenario

Recent advances in Electronic Health Record (EHR) technology have significantly increased the amount of clinical data electronically available. These data, consisting of medical and scientific documents and also of digitalised patient health records, are valuable resources for clinical and translational research. The analysis of the health care experience captured in clinical databases may lead to improved continuity in patient assessment, improved treatment, avoidance of adverse drugs reactions, and in ensuring that people at risk receive appropriate support services (Elliot et al. 2008) (Malin,Sweeney 2004).

Since medical information is usually associated to individuals, privacy must be ensured when data is made available for secondary use. This is explicitly stated by the Data Protection Act 1998 and the Human Rights Act 1998, which consider clinical data as “sensitive”. Nowadays, EHRs are collected and maintained by public and private institutions that made them available for clinicians and

researchers. Those institutions should guarantee that health information associated to patients is only made public with patient's authorisation. Exemptions are allowed in Section 39 of the Data Protection Act for medical purposes as well as statistical or historical research (Elliot et al. 2008). Moreover, the US Health Insurance Portability and Accountability Act (HIPAA) privacy rule permits publishing personal health information for public-health purposes without patient consent, if individual's privacy is "sufficiently" guaranteed (Meystre et al. 2010). To guarantee this privacy, the HIPAA defines 18 data elements, named Protected Health Information (PHI) (GPO, US: 45 C.F.R. 164 Security and Privacy 2008), which must be removed to consider clinical data de-identified. PHI includes names, census, geographical and financial information and biometrics, among others.

While data de-identification (by encrypting or removing identifying attributes) prevents linking confidential data and patient's identity, it provides a false appearance of anonymity. Patient disclosure could still happen through statistical matching of remaining data. Several studies demonstrated that it is still possible to identify a patient by combining quasi-identifier attributes which, when considered individually, did not seem problematic (Domingo-Ferrer 2008; Elliot et al. 2008). There have been cases of disclosure in a priori protected clinical data, such as the identification of the clinician and the patient in a late abortion case through the analysis of released tabular data (Rogers 2005). The identification was possible due to the low amount of late abortions in the region in which both the patient and the clinician were located. In (Malin,Sweeney 2004) authors developed a technique for re-identifying seemingly anonymous genomic data by analysing combinations of, a priori, non-identifying attributes. Medical data, due to their variability and high dimensionality, is very prone to the appearance of identifying combinations of attribute values.

The disclosure of patient electronic health records supposes a serious threat due to the amount of interested parties, including insurance companies, journalists or any attacker with partial knowledge about the patient. Disclosure may produce economic consequences because it may damage people's ability to get jobs, insurances or mortgages, and also legal consequences for clinicians responsible of data releasing, which are exposed to the potential claims that may be derived (NHS 2002). Moreover, the awareness of the disclosure risks inherent to data release may lead to future reluctances and lack of trust in making data available for research. The fact that medical data does not get published or that it is excessively perturbed (e.g. encrypted or partially supressed) to minimise disclosure risk to the extent that it does not meet a level of accuracy, may produce a severe impact in its utility, hampering the benefits that can be extracted from its analysis (Purdam,Elliot 2007).

To avoid these problems, privacy preserving methods guaranteeing a *really sufficient* level of privacy must be developed. These should remove direct or formal identifiers (such as PHI elements), and also generate some distortion (i.e., masking) on the combinations of potentially identifying values, so called *quasi-identifiers* (e.g., rare diagnostics or personalised treatments). It is also equally

important and crucial to assure the quality of the published data. Therefore, this distortion should be done in a way the anonymised data retains its utility as much as possible, so that similar conclusions can be extracted from the analysis of the original and the anonymised version of the dataset.

Different techniques can be envisioned according to the type of data to deal with. Medical data can be presented as unstructured textual documents or as structured patient records collecting values for a set of normalised attributes (e.g., symptoms, diagnosis and treatment observed in a visit to a certain patient). In the former case, document sanitisation methods have been developed (Meystre et al. 2010); the latter case refers to privacy protection in structured databases (Domingo-Ferrer 2008), which is the focus of this thesis.

Different techniques can be identified according to the algorithmic principles in which SDC methods rely to create anonymised datasets (Domingo-Ferrer 2008). Few approaches to anonymisation have been applied to medical data. The simplest methods rely on the *supression* of records whose attributes represent unique or rare value combinations (Samarati,Sweeney 1998; Sweeney 2002a). This strategy has been usually applied in the past to anonymise medical data (Ohno-Machado et al. 2004; Elliot et al. 2008). Even though, it results in a high information loss since the utility of supressed records is completely lost. Moreover, in heterogeneous datasets (such as health records), many records could be removed.

Instead of removing, more sophisticated methods could be applied, such as the ones developed in this thesis:

- Recoding (chapter 5)
- Microaggregation (chapter 6)
- Resampling (chapter 7)

Another important point is that in the medical domain many potentially identifying data is expressed by means of non-numerical attributes, such as categorical information collected in the patients exploration and treatment (Meystre et al. 2010). The preservation of semantics is crucial to ensure the utility of anonymised results (Torra 2011; Martinez et al. 2010b) in EHR.

Due to the importance of terminology and knowledge in clinical assessment, the medical domain has been very prone to the development of large and detailed knowledge structures. ICD-9/10, MeSH or SNOMED CT (Spackman et al. 1997; Nelson et al. 2001) are paradigmatic examples of structured medical terminologies, containing thousands of taxonomically structured medical terms. Details about SNOMED ontology can be seen in section 3.1.2.

Therefore, the goal of this chapter is make a comparative analysis of the semantic anonymisation methods proposed in this thesis in this particular field of application. As far as we know, there are no precedents using these structures to semantically anonymise non-numerical medical data.

9.2 Evaluation and comparison

In this section, we evaluate the contribution of our framework presented in chapter 4 through the SDC methods discussed in Chapter 5 (recoding), Chapter 6 (microaggregation) and Chapter 7 (resampling) during the anonymisation of clinical data with non-numerical attributes. Results are compared according to retained data utility and disclosure risk against the classic non-semantic anonymisation and a naïve method based on data suppression.

9.2.1 The dataset

As evaluation data, we used a structured database containing inpatient information provided by the California Office of Statewide Health Planning and Development (OSHPD) collected from licensed hospitals in California[4]. Specifically, we used the latest patient discharge dataset (4th quarter of 2009) of the hospital with the largest amount of records (i.e., Cedars Sinai Medical Center, Los Angeles County).

Table 13. Example of clinical data used for evaluation. Numbers in parenthesis represent the ICD-9 codes

ID	Age Range	Patient ZIP code	Principal diagnosis cause of admission	Other condition that coexist at the time of admission
*	50-54	916**	abstinent alcoholic (291.81)	metabolic acidosis due to salicylate (276.2)
*	65-69	913**	infected spinal fixation device (996.67)	uric acid renal calculus (592.0)
*	65-69	903**	aneurysm of thoracic aorta (441.2)	cardiac oedema (428.0)
*	>=85	902**	fibroma of ovary (218.9)	chronic osteoarthritis (715.9)
*	30-34	917**	acute fulminating appendicitis (540.9)	body mass index 40+ - severely obese (V85.4)

Prior to publication, the OSHPD has masked or removed some attributes that resulted in unique combinations of certain demographic variables (see an example on the first three columns of Table 13), as suggested by the PHI rules for data anonymisation (GPO, US: 45 C.F.R. 164 Security and Privacy 2008). For example, specific patient age has been masked in a range of 20 categories. Other

[4] http://www.oshpd.ca.gov/HID/Products/PatDischargeData/PublicDataSet/index.html

attributes, such as the hospital identification number have been directly removed from the dataset, whereas the later digits of the ZIP code were also removed. However, clinical data related for example to diagnoses are published "as is"; see the last two columns of Table 13. As stated in the introduction, this kind of information (e.g., rare diseases or combinations of several ones) can also be used to disclose patient identities by means of statistical inference, especially if the attacker has additional knowledge about a certain patient (Malin,Sweeney 2004; Rogers 2005).

In the experiments, we focused on two categorical attributes corresponding to the *principal diagnosis* and *other conditions* of the patient at the time of the admission (i.e., m=2), which are stored as ICD-9 codes in the original data file. After removing records with missing information, a total of 3006 individual records is available for testing (i.e., n=3006). Data distribution for these two attributes is shown in Fig. 36. A total of 2331 different combinations of values (i.e., p=2331 tuples) can be found, from which a significant amount (2073) are unique. As demonstrated in previous works (such as the late abortion identification case (Rogers 2005)), when scarce or even unique combinations of this type of clinical attributes appear, patient private information disclosure may happen if a third party knows other patient's data as for example, the hospital name, its address or the period of hospitalisation. Due to the sensitive nature of these attributes, we considered them as quasi-identifiers that should be masked prior to publication. At the same time, considering that patient diagnoses are valuable information for clinical research, its anonymisation should preserve, as much as possible, the utility of data. Finally, since values for these attributes are non-numerical, a structured medical knowledge base has been used to extract and interpret their semantics, as proposed by our framework. Since ICD-9 codes were available for each condition, we translated them into SNOMED-CT concepts, using a publicly available mapping file[5]. After this mapping, the SNOMED-CT ontology can be used in the anonymisation and evaluation process. See details about SNOMED ontology in section 3.1.2. Its size and fine-grained taxonomical detail make it especially suitable to assist semantic similarity assessments (Batet et al. 2011; Sanchez,Batet 2011; Pedersen et al. 2007).

[5] http://www.nlm.nih.gov/research/umls/licensedcontent/snomedctarchive.html

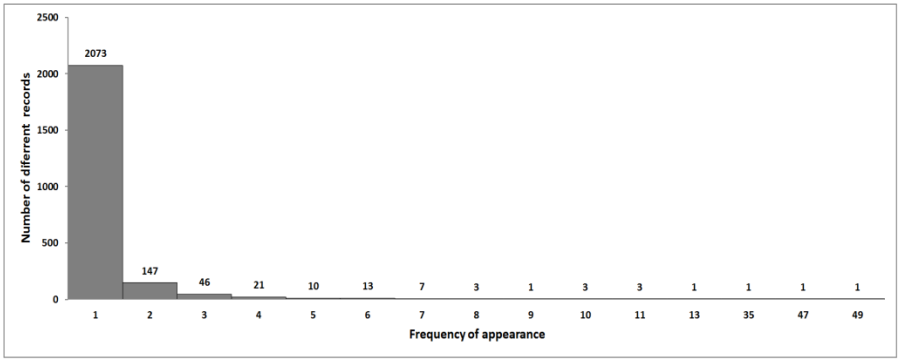

Fig. 36. Distribution of distinct value tuples for the *principal* and other *conditions* attributes.

9.2.2 Evaluating the preservation of data semantics

Taking into consideration that the preservation of data semantics is crucial to retain the utility of anonymised non-numerical data (Torra 2011; Martinez et al. 2010b), in our tests, we measured the quality of anonymised data by quantifying the *semantic information loss* caused by replacing original values by their masked versions.

In the literature, *information loss* of SDC methods focusing on the *k*-anonymity property is usually measured as the *Sum of Squared Errors* (SSE) (Abril et al. 2010; Domingo-Ferrer,Mateo-Sanz 2002; Domingo-Ferrer 2008; Lin et al. 2010b; Nin et al. 2008a). It is defined as the sum of squares of the *distances* between the original records and their masked version (Eq. 9.1). Hence, the lower the SSE is, the lower the information loss, and the higher the data utility will potentially be. To measure the information loss from a *semantic* perspective, we computed SSE scores using the semantic distance dis_{logSC} defined in Eq. 3.9 and SNOMED CT as the knowledge base. Since we are dealing with multivalued data, the average of distances for all attribute values is considered, as follows:

$$SSE = \sum_{i=1}^{n} \left(\frac{\sum_{j=1}^{m} dis_{log\,SC}(r_{ij}, r_{ij}{}^{A})}{m} \right)^2 \tag{9.1}$$

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

In the following, we evaluate the benefits of our semantic framework regarding the preservation of data semantics in comparison with a non-semantic approach. The three SDC methods introduced in chapter 5 (recoding), chapter 6 (microaggregation) and chapter 7 (resampling) will be tested under two different configurations:

Using classical non-semantic operators. In this case values are compared using the equality test (i.e., 0 distance if they are identical and 1 otherwise), whereas the centroid is the most frequent record (i.e., mode). This setting depicts the behaviour of classical anonymising approaches dealing with non-numerical attributes from a non-semantic perspective (Domingo-Ferrer,Torra 2005; Torra 2004).

Using the semantic operators proposed in our framework (chapter 3). In this setting, the three semantic operators proposed in chapter 4 are used, configuring a semantically-grounded anonymisation.

Since the three SDC methods aim at fulfilling the *k*-anonymity property, an analysis with respect to the *k* anonymity level has been done. Considering the data distribution shown in Fig. 36, the *k*-level has been set from 2 to 15, so that for k=15, up to a 90% of the total amount of records will be masked.

In order to have a reference point for the analysis of the results of the methods, we have also implemented a naïve algorithm based on *supressing* those records not fulfilling the *k*-anonymity property, that is, records whose value tuples are repeated less than *k* times. As stated in section 9.1, even though this approach produces a high information loss, it has been applied in the past to anonymise structured datasets (Samarati,Sweeney 1998; Sweeney 2002a) and, more specifically, medical data (Elliot et al. 2008; Ohno-Machado et al. 2004).

Results of the three SDC algorithms (for semantic and non-semantic settings) and the suppression method are shown in Fig. 37. We can see the differences in –semantic- SSE scores (Eq. 9.1) obtained for the dataset described in section 9.2.1 for *k*-values between 2 and 15.

As expected, SSE scores grow as *k*-values increase because, in order to guarantee higher levels of privacy, more records must become indistinguishable and, hence, more changes on the original data are required.

Regarding the non-semantic setting, results show that managing and transforming non-numerical data without considering their semantic features worse preserves the meaning of original data. On the contrary, our semantic approach, that considers both semantics and data distribution, produces significantly lower information loss figures. This shows the benefits of exploiting available medical knowledge bases like SNOMED CT, so that data semantics can be considered (and better preserved) during the anonymisation process.

For some methods the improvement brought by our framework is more significant than for others. The case of microaggregation is the most noticeable, since the semantic framework allows retaining more than a 50% more semantic information than a non-semantic approach. This is coherent, since the microaggregation method heavily relies on semantic operators to group records and to aggregate them. On the other side, the resampling method shows the lowest improvement since it first performs a random sampling of input data, which cannot be optimised from a semantic perspective.

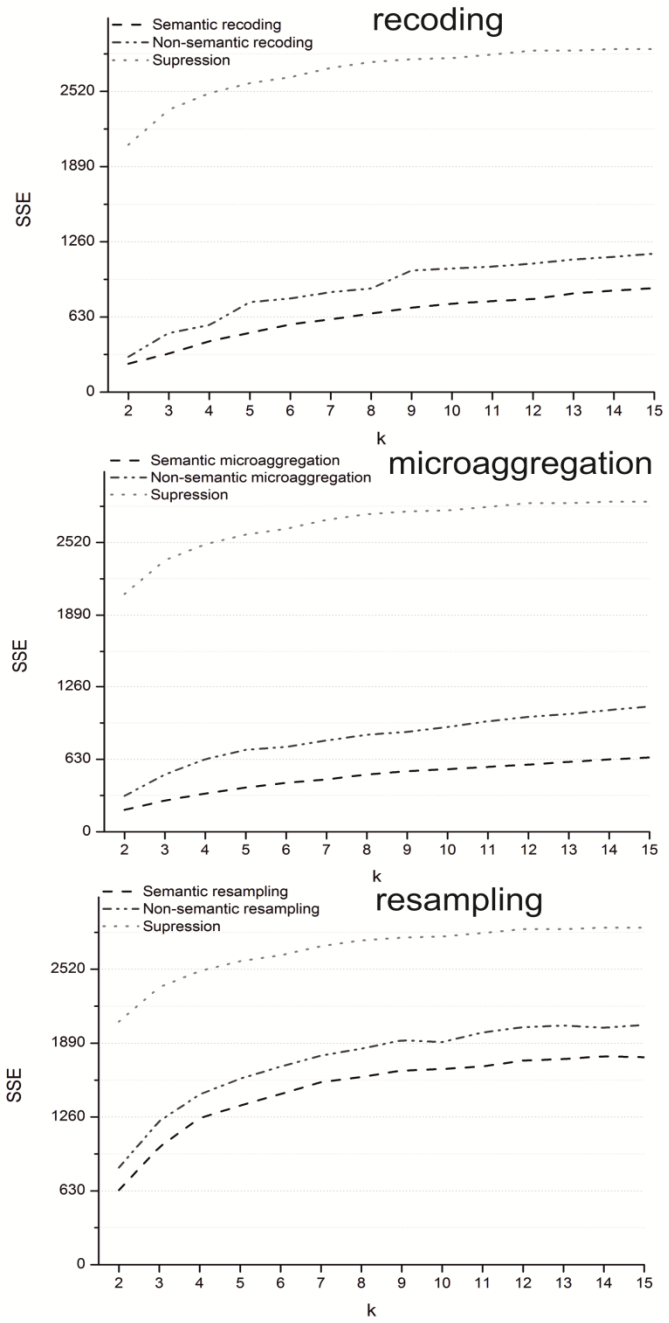

Fig. 37. Semantic Information Loss (SSE) for the three SDC methods (under semantic and non-semantic settings) and the approach based on data suppression for different *k*-anonymity levels.

Information loss obtained when *suppressing* non-anonymous records is significantly higher than what was obtained in the semantic approach. Analysing, for example, SSE values of k=15 (which corresponds to the maximum level of privacy), we can see that for the microaggregation algorithm, the SSE when using the semantic framework is 645, whereas the SSE obtained with suppression is 2875. As discussed in section9.2.1, since most of the records have quite low frequency of appearance in the original EHR, even for low k-values, a high percentage of input data is removed. This severely affects data utility, hampering posterior analyses. This shows the importance of applying utility-preserving SDC methods to anonymise the data and, more concretely, taking into consideration both their distributional and semantic features, as the proposed framework pursues.

9.2.3 Comparing the anonymisation methods

This second analysis aims at studying the convenience of using each SDC method for EHR anonymisation when implemented with our semantic framework. The three SDC algorithms have been compared under the perspectives of information loss, disclosure risk and runtime.

In addition to the semantic information loss measure introduced above, we also computed a standard utility function focusing solely on the preservation of the original data distribution. In this last case, the well-known *KL-divergence* (Kullback,Leibler 1951) score has been considered. Being $f(x_i)$ and $f^*(x_i)$ the probability distributions of an original record x_i in the original and masked datasets, respectively, the *KL-divergence* between both datasets is defined as:

$$KL = \sum_{i=1}^{n} f(x_i) \cdot \ln \frac{f(x_i)}{f^*(x_i)} \tag{9.2}$$

A smaller *KL* score indicates a higher similarity between distributions of records between original and masked datasets.

SSE and KL scores for the different methods using the proposed framework are shown in Fig. 38 for the same k values as above.

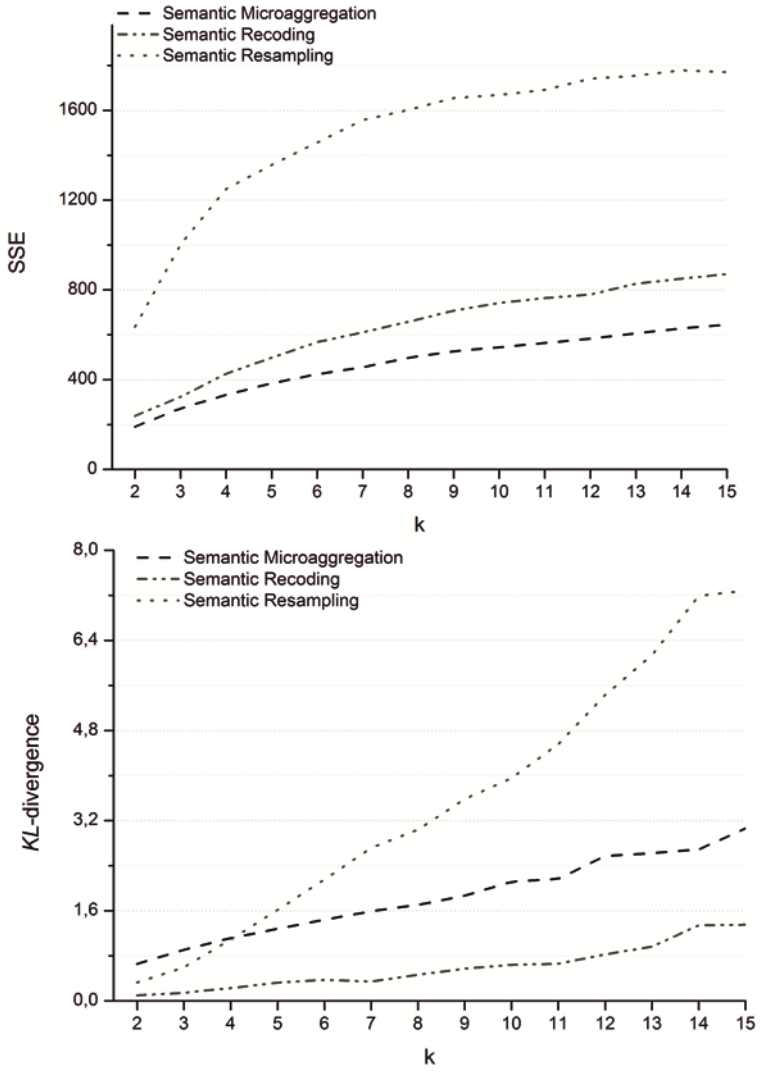

Fig. 38. SSE (semantic) and KL-divergence (distributional) scores for the three SDC methods applying the proposed framework across different levels of k-anonymity.

In terms of semantic preservation (SSE), the best method is *microaggregation* (which is coherent to results obtained by related works (Domingo-Ferrer et al. 2006; Lin et al. 2010a)) closely followed by *recoding*. This makes sense since both methods heavily rely on semantic operators to aggregate or replace record values. *Resampling*, as discussed above, firstly executes a random data sampling that cannot be optimised from a semantic perspective, hampering data utility.

Differences are also reflected on the shape of the SSE function as *k*-values growth. Both *microaggregation* and *recoding* grow almost linearly with respect to *k*. *Resampling*, otherwise, shows an almost logarithmic shape, which is coherent to the fact that the sampling is done on sets of size *n/k*.

Focusing solely on the preservation of data distribution, KL scores show a different picture. In this case, *recoding* provides the best results, followed by *microaggregation* and *resampling*. The fact that both *microaggregation* and *resampling* aggregate records with their centroids, which are synthetically constructed according to the background ontology, may cause that new record values/tuples (i.e. value generalisations and new value tuple combinations) not found in the original dataset appear in the masked version. Even though these new records are semantically similar to original ones, the KL score is penalised, since original values are not found in the masked version. This fact significantly alters the probability distribution of masked data with regards to the original one. On the contrary, *recoding* method systematically replaces records for already existing values. Hence, the probability of finding an original record in the masked dataset will increase, resulting in more similar data distributions. *Resampling*, again, provides the worst results (especially for high *k* values) due to the randomness of the sampling process. This produces less cohesive groups and, hence, more general centroids that will more likely correspond to generalisations rather than to values found in the input dataset.

Another dimension to consider is the Disclosure Risk (DR). As stated above, DR is the chance to disclosure the identity of an individual and it is opposite to information loss because there is a trade-off between the preservation of data utility and the disclosure risk as show in chapter 8. DR is calculated by means of Record Linkage as the percentage of correct linkages between the original and masked datasets. To compare the disclosure risk of the three SDC algorithms we have measured the Record Linkage of the datasets obtained from the anonymisation. The Record Linkage have been evaluated by means of our SRL method (see section 8.3), using SNOMED CT (see section 3.1.2) as ontology and the semantic distance measure dis_{logSC} defined in Eq. 3.9 as the criteria to propose linkages.

Analysing the *SRL* results of the masked dataset for the three SDC methods (Fig. 39) can be extracted several conclusions. First, resampling is the method with the lowest percentage of record linkages from k=4. This method replaces group values in a random way and then aggregates them by its centroid. As a result, most values in the original dataset are replaced in the masked dataset. For this reason the chance of proposing a correct linkage is very low, hence, the SRL values tend to be very low. The opposite case applies for the microaggregation approach. The values are grouped in homogeneous clusters then they are replaced by centroids more semantically close to values. This increases the chance of make corrects linkages using the semantic matching proposed by SRL. The recoding method obtain the best result for k=2 and remains between the results of the other two methods for the rest of *k*. In comparison, recoding method presents an average behaviour. The fact that the recoding method replaces records for already existing

values implies that the semantic approach of SRL not increase the number of correct linkages.

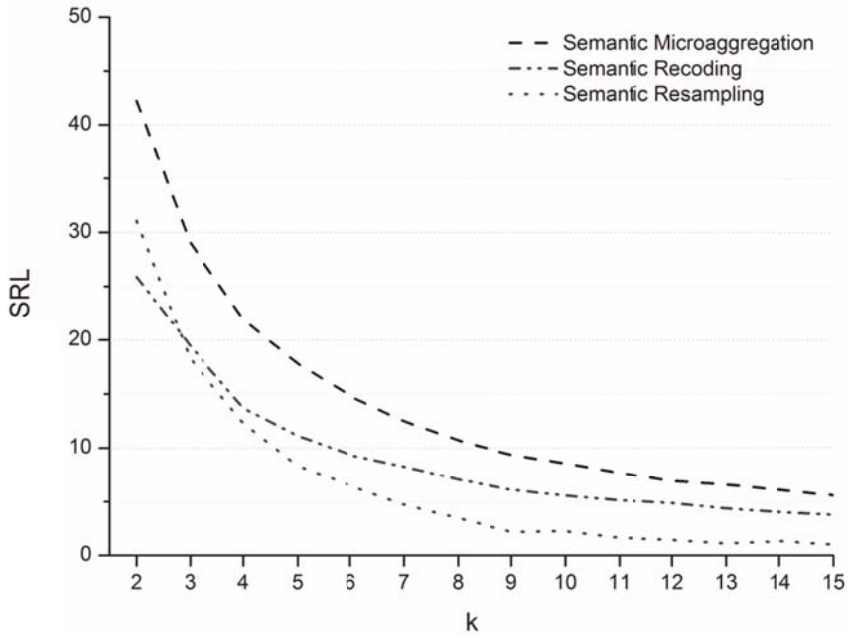

Fig. 39. A comparison of SRL percentage for the evaluated methods.

To evaluate the balance score between the information loss and the disclosure risk of the different methods, we measured the overall score (Eq. 6.5). The lower the score is, the higher the quality of the method because both low information loss and low disclosure risk are achieved. First, as shown in Fig. 40, we consider an equal balance between the data utility and the disclosure risk, an average with α=0.5, as done in section 6.3.2. To calculate the score value, we measured the information loss in function of L (Eq. 6.4) for the three methods.

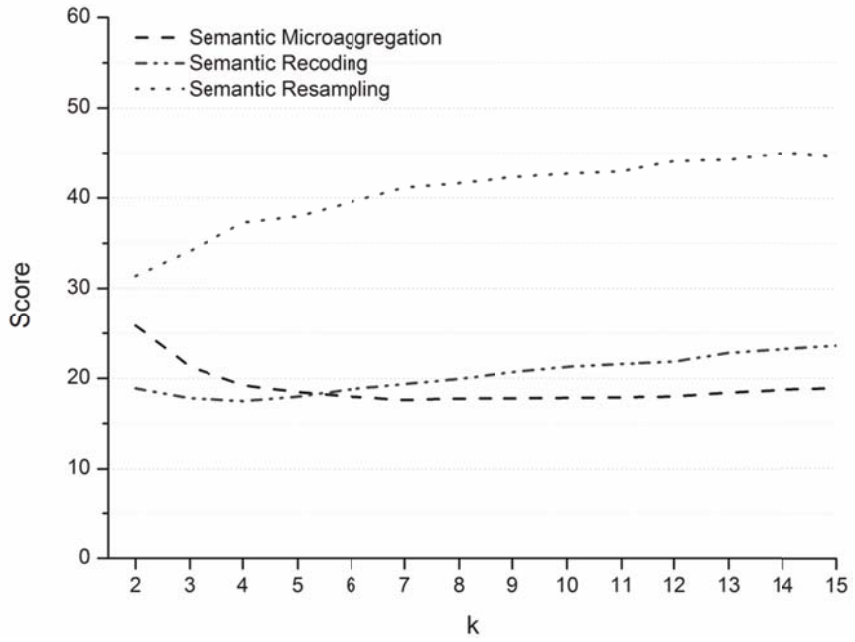

Fig. 40. Score with an equilibrated balance (α=0.5) between information loss and disclosure risk.

The conclusion is that, recoding method provides the best results for low *k* levels (*k*<=5) and microaggregation method works better for high *k* levels (*k*>5). Thus, depending on the level of desired *k*-anonymity be convenient to choose a method or another. It is also relevant to note that the score is maintained almost constant as *k*-values grow stating that the quality of the methods scales well as the privacy requirements increase. On the other hand, even though the resampling method resulted in lowest SRL values, it provides the worst balance between information loss and disclosure risk, a circumstance that is the goal of a SDC method. This is due to the god results obtained by resampling on disclosure risk not compensate the bad results obtained on information loss.

We have also studied the behaviour of the overall score when varying the parameter α 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*=7) has been fixed. As it can be seen in Fig. 41, recoding method achieves the best results for lows α and microaggregation method obtains the best results for highs α. This means that if we need to give more importance to information loss with an adequate level of disclosure risk, we will use the microaggregation method. On the contrary, to give more weight to disclosure risk with a moderate level of information loss, we will select the recoding method. The resampling method obtains the minimal score for almost all the cases due to the poor results obtained in information loss. As shown

in Fig. 41, resampling method obtains the best result only when data utility is not taken into account.

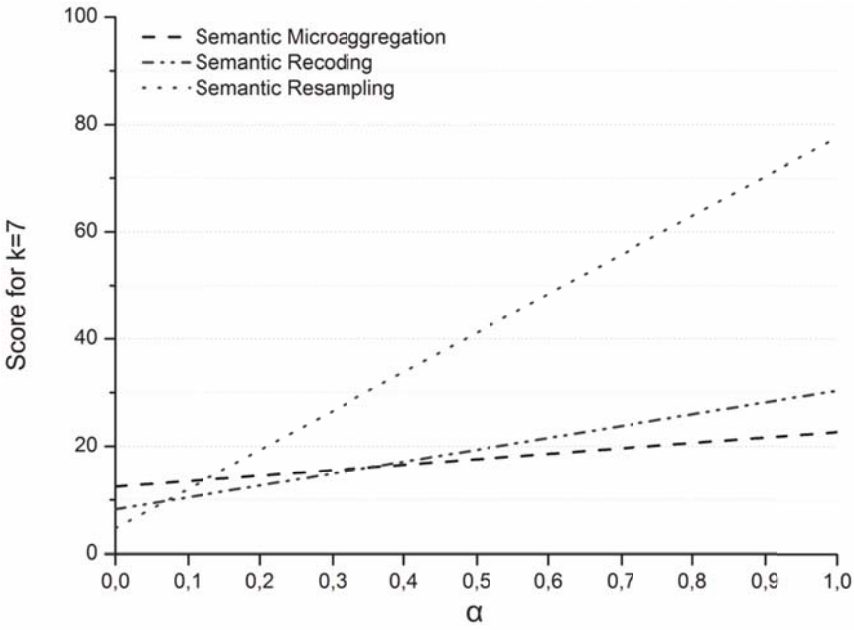

Fig. 41. Score values when varying the relative weight between information loss and disclosure risk.

Finally, the computational efficiency of data anonymisation is a relevant feature to consider when resources are limited because EHRs are likely to contain large amounts of data. Fig. 42 shows the comparison for the three SDC methods executed on a 2.4 GHz Intel Core processor with 4 GB RAM.

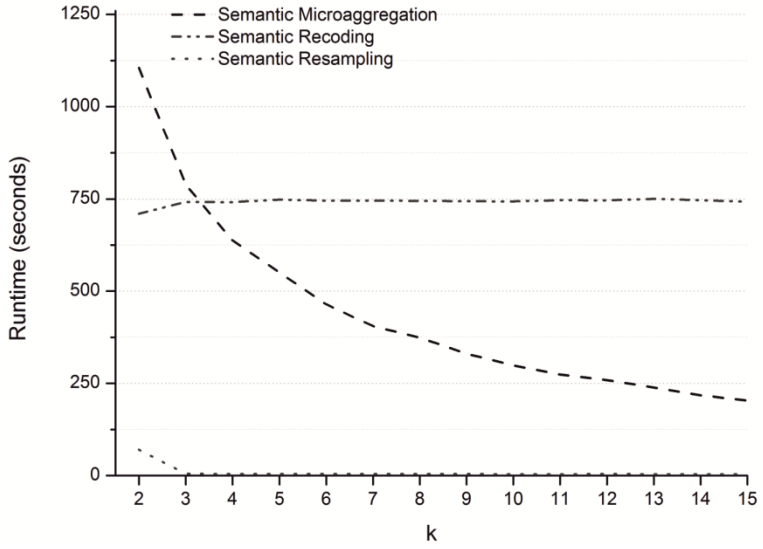

Fig. 42. Runtime (in seconds) for the three SDC methods applying the proposed framework across different levels of *k*-anonymity.

Runtime figures show that the fastest method is *resampling*, with an almost negligible runtime. *Microaggregation* is the slowest for *k* values below 4, whereas it surpasses *recoding* for higher *k*-values, with an almost inverse logarithmic shape. Results are coherent to the computational costs of the different methods. First, *microaggregation* scales $O(p^2/k)$, where p is the number of distinct records. Since the number of needed clusters lowers as k increases (since each cluster groups, at least, k records), the number of microaggregation iterations and, hence the runtime, lowers as k increases. *Recoding* replaces non-*k*-anonymous records by other similar ones, which implies a cost of $O(p^2)$ for each iteration, resulting in a total of $O(p^3)$ in the worst case (i.e. all records are non-*k*-anonymous). Since the asymptotic computational cost can neglect the value of k, the runtime is almost constant with respect to this parameter.

Finally, as resampling firstly randomly divides input data in n/k samples, it scales linearly as $O(n/k)$, resulting in the lowest runtime, which is even lower when k values increase.

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

large to be non-computationally feasibly anonymised by means of *microaggregation* or *recoding* methods, the *resampling* method could be considered thanks to its high efficiency, at the expenses of a higher information loss (both semantic and distributional).

9.3 Summary

Appropriate measures to protect the privacy of EHRs should be taken before making them available for clinical research. On the one hand, original data should be anonymised in a way that the chance of patient disclosure, even when applying statistical methods, is sufficiently reduced. On the other hand, anonymised datasets should retain as much information as possible, so that they are still useful for research tasks. SDC methods aim at balancing these two complementary dimensions, up to now, anonymisation with regards to clinical terms has not been exploited.

Considering the amount and importance of categorical data in EHRs, in this chapter we applied the general framework proposed in chapter 4 that provides semantically-grounded *comparison*, *aggregation* and *sorting* operators. Exploiting a structured medical knowledge base like SNOMED-CT, and relying on the theory of *semantic similarity*, these operators enable a semantically-coherent interpretation of non-numerical attributes, while also considering their distributional features. Since many SDC methods (particularly those focused on fulfilling *k*-anonymity) rely on these basic operators to anonymise data, they can directly apply the proposed semantic framework to produce semantically-grounded anonymisations for categorical data. As shown in the evaluation performed with a real clinical dataset, the use of this framework considerably improves the degree of semantic preservation for all the considered methods, in comparison with non-semantic approaches based solely on information distribution. As a result, the meaning of data and, hence, the utility of the anonymised data from a semantic perspective is better preserved.

The main contributions presented in this chapter are published in:

2J. Martínez, S., Valls, A., Sánchez, D.: A semantic framework to protect the privacy of Electronic Health Records with non-numerical attributes. Journal of Biomedical Informatics. Accepted manuscript. *Impact Factor*: 1.792

Chapter 10
Conclusions and future work

Statistical Disclosure Control (SDC) methods aim at publishing data while preserving the confidentiality of individuals. While published data files has to be as close to the original as possible in order to maintain their utility, the disclosure risk has to be minimised in order to protect the individuals. This implies a trade-off between utility and privacy.

In this thesis, we have studied the results obtained when retaining the semantics of the datasets with categorical attributes, showing that this aspect is crucial in order to retain its analytical utility. Many SDC masking methods have been designed in the past, but most of them focus solely on numerical attributes. The application of these methods to categorical data is not straightforward due to the limitations on defining suitable operators to deal with non-numerical data. There have been some previous works using ad-hoc structures like Value Generalisation Hierarchies (VGH) but they present some drawbacks: (1) VGH are constructed manually in function of input data and human intervention is required, (2) they offer a rough and biased knowledge model and (3) the quality of the results depends on the structure of VGH. Considering the importance of data semantics in anonymisation tasks, we have studied the definition of new operators that incorporate the semantic knowledge offered by ontologies, in order to be used in making methods.

The first contribution of this thesis is the **definition of a general framework that enables the anonymisation of categorical data from a semantic perspective.** We have formalised a set of operators that exploit knowledge structures to enable a semantic-coherent interpretation of categorical attributes without neglecting their distributional features. The proposed semantic framework is composed by the follow operators:

- **Comparison operator**: calculates the distance between terms considering both semantic and distributional features of categorical data.
- **Aggregation operator**: we proposed a new approach to construct centroids of datasets with categorical attributes with special emphasis on the preservation of data semantics and taking also into account the data distribution. Note that the quality of the masked data closely depends on

the quality of the aggregation operator because, usually, data is masked by aggregating subsets of the original dataset.

- **Sorting operator**: using the above to operators it is possible to define semantically-coherent total orders for categorical data that are required for certain anonymization tasks.

As first conclusion, we have demonstrated that **by means of the proposed framework that exploits knowledge structures as WordNet or SNOMED CT, categorical terms can be coherently managed taking into account its semantics.**

The next contribution of this thesis consists on using the above framework to **design and/or adapt three SDC algorithms for categorical data**:

- **Recoding algorithm**: the algorithm uses the semantic comparison operator in order to properly find the most appropriate substitution of sensitive values aiming to minimise both disclosure risk and information loss. The algorithm also applies heuristics during the masking process to make the anonymisation computationally feasible and scalable when dealing with large datasets.
- **Microaggregation algorithm**: applying the comparison and aggregation operators, we proposed an adaption of the classic MDAV microaggregation algorithm that properly deals with categorical data while preserving original semantics and data distribution.
- **Resampling algorithm**: applying the three proposed operators the algorithm is able to treat categorical data from a semantic perspective. Compared to others methods, resampling is faster and appropriate for large datasets in which others methods cannot work.

These contributions have been thoroughly evaluated with real datasets. Evaluation results sustain our hypothesis, obtaining as second conclusion that these **semantic-based masking algorithms are able to minimise the information loss in comparison with non-semantic approaches.**

Another important aspect of any masking method is the minimisation of disclosure risk. We proposed and evaluated **a new ontology-based record linkage method**, which focuses on evaluating the disclosure risk of semantically-grounded anonymisation methods. From the results of the tests, we can obtain a third conclusion: **the privacy achieved by the masking algorithms can be more accurately evaluated from a semantic perspective.**

The contributions of this Ph. D thesis have been comprehensively **applied and evaluated within the context of Electronic Health Record anonymisation**. Exploiting medical structured knowledge like SNOMED CT and the proposed semantic framework we enable a semantically coherent interpretation of medical terms. This semantic information is used during the masking process to maintain as much utility as possible, which is crucial for clinical and translational research.

Disclosure risk has been also evaluated with the proposed record linkage method, in addition to the computation cost, which is relevant when dealing with large datasets. The evaluation of these three aspects permits to compare the three anonymisation methods under different perspectives. Considering the results, we can conclude that semantic microaggregation and recoding methods are the best approaches to mask categorical data. Concretely, microaggregation achieves a high level of *k*-anonymity and recoding achieves a low level of *k*-anonymity. Finally, resampling method could be applied when input databases are too large to be anonymised in a computationally feasible time through microaggregation or recoding methods.

To sum up, **as a general conclusion of the thesis we have shown that well-defined general purpose semantic structures, as ontologies are a good source of information to interpret the semantics of terms and their use is crucial to retain the utility of data during the anonymisation process**.

Finally, we list here the papers have been published about the work done in this Ph. D Thesis:

Journals:

1J. Martínez, S., Valls, A., Sánchez, D.: Semantically-grounded construction of centroids for datasets with textual attributes. International journal: Knowledge-Based Systems. 35(0), 160-172 (2012). *Impact Factor: 2.422*

2J. Martínez, S., Valls, A., Sánchez, D.: A semantic framework to protect the privacy of Electronic Health Records with non-numerical attributes. Journal of Biomedical Informatics. Accepted manuscript. *Impact Factor: 1.792*

3J. Martinez, S., Sanchez, D., Valls, A., Batet, M.: Privacy protection of textual attributes through a semantic-based masking method. International Journal: Information Fusion 13(4), 304-314 (2011). *Impact Factor: 1.467*

4J. Martínez, S., Sánchez, D., Valls, A.: Semantic Adaptive Microaggregation of Categorical Microdata. International Journal: Computers & Security 31(5), 653-672 (2012). *Impact Factor: 0.868*

5J. Martínez, S., Sánchez, D., valls, A.: Evaluation of the Disclosure Risk of Masking methods Dealing with Textual Attributes. International Journal of Innovative Computing, Information & Control 8(7(A)), 4869-4882 (2012).

Conferences:

1C. Martinez, S., Sanchez, D., Valls, A.: Ontology-Based Anonymization of Categorical Values. In: Torra, V., Narukawa, Y., Daumas, M. (eds.) Modeling Decisions for Artificial Intelligence, vol. 6408. Lecture

Notes in Computer Science, pp. 243-254. Springer Berlin / Heidelberg, (2010). *CORE B*

2C. Martinez, S., Sanchez, D., Valls, A., Batet, M.: The Role of Ontologies in the Anonymization of Textual Variables. In: the 13th International Conference of the Catalan Association for Artificial Intelligence 2010, pp. 153-162.

3C. Martínez, S., Valls, A., Sánchez, D.: Anonymizing Categorical Data with a Recoding Method Based on Semantic Similarity. In: Hüllermeier, E., Kruse, R., Hoffmann, F. (eds.) Information Processing and Management of Uncertainty in Knowledge-Based Systems. Applications, vol. 81. Communications in Computer and Information Science, pp. 602-611. Springer Berlin Heidelberg, (2010). *CORE C*

4C. Martínez, S., Sánchez, D., Valls, A.: Towards k-anonymous non-numerical data via Semantic Resampling. In: Information Processing and Management of Uncertainty in Knowledge-Based Systems, Catania, Italy 2012, pp. 519-528. S. Greco et al. (Eds.). *CORE C*

5C. Martínez, S., Valls, A., Sánchez, D.: An ontology-based record linkage method for textual microdata. In, vol. 232. Frontiers in Artificial Intelligence and Applications, pp. 130-139. (2011).

As **future work**, we can identify some general research lines.

Regarding the knowledge base used to guide the anonymisation process, it would be interesting to study how the different methods behave with other ontologies with different sizes and granularities. The possibility of combining several ontologies as background knowledge could be also considered, in order to complement knowledge modelled for each of them. To do so, methods to integrate and coherently compare knowledge and semantics of different ontologies can be used (Batet et al. 2012; Sanchez et al. 2012; Sánchez et al. 2012a; Sánchez,Batet 2013). Regarding the semantic interpretation of terms, linguistic techniques like morpho-syntactic analyses and part-of-speech tagging can be applied to extend this approach not only to simple categorical attributes (i.e. words) but to more complex sentences. This will permit to use the proposed methods in a wider range of applications, such as private information retrieval in Web Search Engines (Sánchez et al. 2013).

Regarding the privacy model, research on the application of the proposed semantic framework to other models and methods can be done. Particularly it is important to note that privacy models such as *k-anonymity*, even though providing a more robust anonymisation in front of statistical disclosure attacks than the sole removal of identifying attributes, present some limitations. First, an intruder can discover sensitive values when there is little diversity on those sensitive attributes. Second, attacks based on the background knowledge that a potential attacker may have, can compromise the dataset privacy (Machanavajjhala et al. 2007). Some authors in the literature proposed complementary models to *k*-anonymity. *l*-

diversity (Machanavajjhala et al. 2007) guarantees stronger privacy introducing certain level of diversity in the sensitive values. A limitation of *l*-diversity model is that it does not consider semantic meanings of sensitive values. Another alternative to the *k*-anonymity is the *t*-closeness model (Li et al. 2007). A dataset have *t*-closeness if all equivalence classes have *t*-closeness. An equivalence class have *t*-closeness if the distance between the distribution of a sensitive attribute in this class and this distribution in the whole dataset is no more than a threshold *t*. We argue that the proposed framework can be applied in a future research to these models in order to adapt them to be able to manage categorical data.

Related with the privacy model, recently, a more robust privacy model named *differential privacy* have been published. *Differential privacy* (Dwork 2006) ensures that released data are insensitive to any individual's data (i.e. the alteration of one input records). Hence, individual data remains uncertain for an attacker. To achieve this, an amount of noise is usually added to the output to introduce uncertainty, a process that implies a loss of information. The approach proposed in this thesis could be extended to fulfil differential privacy by adding appropriate noise to, for example, the record counts of each aggregated group (Mohammed et al. 2011) or to computed centroids. In this case, a trade-off between the degree of data aggregation and the amount of noise should be carefully considered to minimise information loss.

Finally, the semantic framework could be also applied to other privacy-preserving methods, such as *rank swapping* (Nin et al. 2008b) or *k*-ward (Domingo-Ferrer,Mateo-Sanz 2002). In those cases, we would also consider the definition of additional operators such as variance or co-variance for categorical data. These statistical operators can be useful for anonymised data evaluation for data mining purposes where it can be interesting to maintain these statistics during the masking process (Domingo-Ferrer 2012).

References

Abril, D., Navarro-Arribas, G., Torra, V.: Towards semantic microaggregation of categorical data for confidential documents. Paper presented at the Proceedings of the 7th international conference on Modeling decisions for artificial intelligence, Perpignan, France, (2010)

Aggarwal, C.C., Yu, P.S.: Privacy-Preserving Data Mining: Models and Algorithms. Springer Publishing Company, Incorporated, (2008)

Bai, L., Liang, J., Dang, C.: An initialization method to simultaneously find initial cluster centers and the number of clusters for clustering categorical data. Knowl.-Based Syst. **24**(6), 785-795 (2011). doi:10.1016/j.knosys.2011.02.015

Batet, M.: Ontology-Based Semantic Clustering. Universitat Rovira i Virgili (2011)

Batet, M., Isern, D., Marin, L., Martínez, S., Moreno, A., Sánchez, D., Valls, A., Gibert, K.: Knowledge-driven delivery of home care services. J. Intell. Inf. Syst., (in press) (2010). doi:10.1007/s10844-010-0145-0

Batet, M., Sanchez, D., Valls, A.: An ontology-based measure to compute semantic similarity in biomedicine. J Biomed Inform **44**(1), 118-125 (2011)

Batet, M., Sánchez, D., Valls, A., Gibert, K.: Semantic similarity estimation from multiple ontologies. Applied Intelligence, 1-16 (2012). doi:10.1007/s10489-012-0355-y

Batet, M., Valls, A., Gibert, K.: Improving classical clustering with ontologies. Paper presented at the Proceedings of the 4th World conference of the IASC, Japan, (2008)

Bayardo, R.J., Agrawal, R.: Data Privacy through Optimal k-Anonymization. Paper presented at the the 21st International Conference on Data Engineering, (2005)

Bechhofer, S., Harmelen, F.v., Hendler, J., Horrocks, I., McGuiness, D., Patel-Scheinder, D., Stein, L.: OWL Web Ontology Language Reference. http://www.w3.org/TR/owl-re (2009)

Bollegala, D., Matsuo, Y., Ishizuka, M.: Measuring semantic similarity between words using web search engines. Paper presented at the Proceedings of the 16th international conference on World Wide Web, Banff, Alberta, Canada, (2007)

Byun, J.-W., Kamra, A., Bertino, E., Li, N.: Efficient k-anonymization using clustering techniques. Paper presented at the Proceedings of the 12th

international conference on Database systems for advanced applications, Bangkok, Thailand, (2007)

Cao, F., Liang, J., Li, D., Bai, L., Dang, C.: A dissimilarity measure for the k-Modes clustering algorithm. Knowl.-Based Syst. **26**(0), 120-127 (2012). doi:10.1016/j.knosys.2011.07.011

Chakaravarthy, V., Gupta, H., Roy, P., Mohania, M.: Efficient techniques for document sanitization. In: CIKM '08: Proceeding of the 17th ACM conference on Information and knowledge management, Napa Valley, California, USA 2008, pp. 843-852. ACM

Cheng, R.C., Bau, C.T., Tsai, M.Y., Huang, C.Y.: Web Pages Cluster Based on the Relations of Mapping Keywords to Ontology Concept Hierarchy. International Journal of Innovative Computing, Information and Control **6**(6), 2749–2760 (2010)

Chiu, C.-C., Tsai, C.-Y.: A k-Anonymity Clustering Method for Effective Data Privacy Preservation. Paper presented at the Proceedings of the 3rd international conference on Advanced Data Mining and Applications, Harbin, China, (2007)

Dalenius, T., Reiss, S.P.: Data-swapping: A technique for disclosure control. Journal of Statistical Planning and Inference **6**(1), 73-85 (1982). doi:Doi: 10.1016/0378-3758(82)90058-1

De Mántaras, R.L.: A Distance-Based Attribute Selection Measure for Decision Tree Induction. Mach. Learn. **6**(1), 81-92 (1991). doi:10.1023/a:1022694001379

Ding, L., Finin, T., Joshi, A., Pan, R., Cost, R.S., Peng, Y., Reddivari, P., Doshi, V., Sachs, J.: Swoogle: a search and metadata engine for the semantic web. Paper presented at the Proceedings of the thirteenth ACM international conference on Information and knowledge management, Washington, D.C., USA, (2004)

Domingo-Ferrer, J.: Microaggregation for database and location privacy. Paper presented at the Proceedings of the 6th international conference on Next Generation Information Technologies and Systems, Kibbutz Shefayim, Israel, (2006)

Domingo-Ferrer, J.: A Survey of Inference Control Methods for Privacy-Preserving Data Mining. In: Aggarwal, C.C., Yu, P.S. (eds.) Privacy-Preserving Data Mining, vol. 34. Advances in Database Systems, pp. 53-80. Springer US, (2008)

Domingo-Ferrer, J.: Marginality: A Numerical Mapping for Enhanced Exploitation of Taxonomic Attributes. In: Torra, V., Narukawa, Y., López, B., Villaret, M. (eds.) Modeling Decisions for Artificial Intelligence. Lecture Notes in Computer Science, pp. 367-381. Springer Berlin Heidelberg, (2012)

Domingo-Ferrer, J., Martínez-Ballesté, A., Mateo-Sanz, J., Sebé, F.: Efficient multivariate data-oriented microaggregation. The VLDB Journal **15**(4), 355-369 (2006). doi:10.1007/s00778-006-0007-0

Domingo-Ferrer, J., Mateo-Sanz, J.M.: Resampling for statistical confidentiality in contingency tables. Computers & Mathematics with Applications **38**(11-12), 13-32 (1999). doi:10.1016/s0898-1221(99)00281-3

Domingo-Ferrer, J., Mateo-Sanz, J.M.: Practical Data-Oriented Microaggregation for Statistical Disclosure Control. IEEE Trans. on Knowl. and Data Eng. **14**(1), 189-201 (2002). doi:10.1109/69.979982

Domingo-Ferrer, J., Seb\, F., \#233, Solanas, A.: A polynomial-time approximation to optimal multivariate microaggregation. Comput. Math. Appl. **55**(4), 714-732 (2008). doi:10.1016/j.camwa.2007.04.034

Domingo-Ferrer, J., Torra, V.: Disclosure control methods and information loss for microdata. In: Elsevier (ed.) Confidentiality, Disclosure, and Data Access: Theory and Practical Applications for Statistical Agencies. pp. 91-110. (2001a)

Domingo-Ferrer, J., Torra, V.: A quantitative comparison of disclosure control methods for microdata. In: P. Doyle, J.L., J. Theeuwes and L. Zayatz (ed.) Confidentiality, Disclosure and Data Access. pp. 111–133. Amsterdam: North-Holland (2001b)

Domingo-Ferrer, J., Torra, V.: Ordinal, Continuous and Heterogeneous k-Anonymity Through Microaggregation. Data Min. Knowl. Discov. **11**(2), 195-212 (2005). doi:10.1007/s10618-005-0007-5

Dwork, C.: Differential privacy. In: in ICALP 2006, pp. 1-12. Springer

Elliot, M., Purdam, K., Smith, D.: Statistical disclosure control architectures for patient records in biomedical information systems. J Biomed Inform **41**(1), 58-64 (2008). doi:10.1016/j.jbi.2007.05.002

Erola, A., Castella-Roca, J., Navarro-Arribas, G., Torra, V.: Semantic microaggregation for the anonymization of query logs. Paper presented at the Proceedings of the 2010 international conference on Privacy in statistical databases, Corfu, Greece, (2010)

European project IST-2000-25069 ESSproject, t.F., 2001-2009: Computational aspects of statistical confidentiality, 2009. http://neon.vb.cbs.nl/casc/..\casc\index.htm (2009)

European_Comission: European Privacy Regulations. http://ec.europa.eu/justice_home/fsj/privacy/index_en.htm (2012)

Everitt, B.S., Landau, S., Leese, M., Stahl, D.: Cluster Analysis, 5th Edition. Wiley Series in Probability and Statistics, (2011)

F.C.Statistical-Methodology: Report on statistical disclosure limitation methodology. Technical report, Statistical and Science Policy, Office of

Information and Regulation Affairs, Office of Management and Budget. In. (2005)

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

Fung, B.C.M., Wang, K., Yu, P.S.: Top-Down Specialization for Information and Privacy Preservation. Paper presented at the Proceedings of the 21st International Conference on Data Engineering, (2005)

Gomez-Perez, A., Fernandez-Lopez, M., Corcho, O.: Ontological Engineering, 2nd printing. Springer-Verlag, (2004)

Gouweleeuw, J.M., Kooiman, P., Willenborg, L.C.R.J., DeWolf, P.P.: Post randomization for statistal disclousure control: Theory and implementation. In. Voorburg: Statistics Netherlands, (1997)

GPO, US: 45 C.F.R. 164 Security and Privacy 2008. http://www.access.gpo.gov/nara/cfr/waisidx_08/45cfr164_08.html

Greenacre, M., Hastie, T.: Dynamic visualization of statistical learning in the context of high-dimensional textual data. Web Semantics: Science, Services and Agents on the World Wide Web **8**(2-3), 163-168 (2010). doi:DOI: 10.1016/j.websem.2010.03.007

Guarino, N.: Formal Ontology and Information Systems. In: (ed), G.N. (ed.) 1st Int. Conf. on Formal Ontology in Information Systems, Trento, Italy 1998, pp. 3-15. IOS Press

Guo, L., Wu, X.: Privacy Preserving Categorical Data Analysis with Unknown Distortion Parameters. Trans. Data Privacy **2**(3), 185-205 (2009)

Guzman-Arenas, A., Cuevas, A.-D., Jimenez, A.: The centroid or consensus of a set of objects with qualitative attributes. Expert Syst. Appl. **38**(5), 4908-4919 (2011). doi:10.1016/j.eswa.2010.09.169

Guzman-Arenas, A., Jimenez-Contreras, A.: Obtaining the consensus and inconsistency among a set of assertions on a qualitative attribute. Expert Syst. Appl. **37**(1), 158-164 (2010). doi:10.1016/j.eswa.2009.05.010

Han, J.: Data Mining: Concepts and Techniques, 2nd. ed. Morgan Kaufmann, cop., San Francisco (Calif.) (2005)

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

He, Z., Xu, X., Deng, S.: k-ANMI: A mutual information based clustering algorithm for categorical data. Inf. Fusion **9**(2), 223-233 (2008). doi:http://dx.doi.org/10.1016/j.inffus.2006.05.006

Heer, G.R.: A bootstrap procedure to preserve statistical confidentiality in contingency tables. In: Lievesley, D. (ed.) International Seminar on

Statistical Confidentiality, Luxembourg 1993, pp. 261-271. Eurostat (1993)

Herranz, J., Matwin, S., Nin, J., Torra, V.: Classifying data from protected statistical datasets. Computers & Security **29**(8), 875-890 (2010a). doi:DOI: 10.1016/j.cose.2010.05.005

Herranz, J., Nin, J., Torra, V.: Distributed Privacy-Preserving Methods for Statistical Disclosure Control

Data Privacy Management and Autonomous Spontaneous Security. **5939**, 33-47 (2010b). doi:10.1007/978-3-642-11207-2_4

Hettich, S., Bay, S.D.: The UCI KDD Archive. In. (1999)

HIPAA: Health insurance portability and accountability 2010. http://www.hhs.gov/ocr/hipaa/ (2010)

Huang, K.L., Kanhere, S.S., Hu, W.: Preserving privacy in participatory sensing systems. Computer Communications **33**(11), 1266-1280 (2010a). doi:10.1016/j.comcom.2009.08.012

Huang, T., Yu, Y., Guo, G., Li, K.: A classification algorithm based on local cluster centers with a few labeled training examples. Knowl.-Based Syst. **23**(6), 563-571 (2010b). doi:10.1016/j.knosys.2010.03.015

Huang, Z.: Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values. Data Min. Knowl. Discov. **2**(3), 283-304 (1998). doi:10.1023/a:1009769707641

Hundepool, A., Wetering, A.V.d., Ramaswamy, R., Franconi, L., Capobianchi, A., DeWolf, P.P., Domingo-Ferrer, J., Torra, V., Brand, R., Giessing, S.: 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. In. (2003)

Iyengar, V.S.: Transforming data to satisfy privacy constraints. Paper presented at the KDD, (2002)

Jiang, J.J., Conrath, D.W.: Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy. In: International Conference Research on Computational Linguistics (ROCLING X), 1997 1997, p. 9008

Jin, X., Zhang, N., Das, G.: ASAP: Eliminating algorithm-based disclosure in privacy-preserving data publishing. Information Systems **36**(5), 859-880 (2011). doi:10.1016/j.is.2011.03.001

Jones, D.H., Adam, N.R.: Disclosure avoidance using the bootstrap and other resampling schemes. In: Proceedings of the Fifth Annual Research Conference, U.S. Bureau of the Census, Washington, DC 1989, pp. 446-455

Karr, A.F., Kohnen, C.N., Oganian, A., Reiter, J.P., Sanil, A.P.: A Framework for Evaluating the Utility of Data Altered to Protect Confidentiality. The American Statistician **60**, 224-232 (2006)

Keikha, M., Khonsari, A., Oroumchian, F.: Rich document representation and classification: An analysis. Knowl.-Based Syst. **22**(1), 67-71 (2009). doi:10.1016/j.knosys.2008.06.002

Kokar, M.M., Matheus, C.J., Baclawski, K.: Ontology-based situation awareness. Inf. Fusion **10**(1), 83-98 (2009). doi:DOI: 10.1016/j.inffus.2007.01.004

Kullback, S., Leibler, R.: On Information and Sufficiency. The Annals of Mathematical Statistics **22**(1), 79-86 (1951). doi:citeulike-article-id:3245942

Lassila, O., van Harmelen, F., Horrocks, I., Hendler, J., McGuinness, D.L.: The semantic Web and its languages. Intelligent Systems and their Applications, IEEE **15**(6), 67-73 (2000). doi:10.1109/5254.895864

Laszlo, M., Mukherjee, S.: Minimum Spanning Tree Partitioning Algorithm for Microaggregation. IEEE Trans. on Knowl. and Data Eng. **17**(7), 902-911 (2005). doi:10.1109/tkde.2005.112

Leacock, C., Chodorow, M.: Combining local context with WordNet similarity for word sense identification. In: WordNet: A Lexical Reference System and its Application 1998

LeFevre, K., DeWitt, D.J., Ramakrishnan, R.: Incognito: efficient full-domain K-anonymity. Paper presented at the Proceedings of the 2005 ACM SIGMOD international conference on Management of data, Baltimore, Maryland, (2005)

LeFevre, K., DeWitt, D.J., Ramakrishnan, R.: Mondrian Multidimensional K-Anonymity. Paper presented at the the 22nd International Conference on Data Engineering, (2006)

Li, N., Li, T., Venkatasubramanian, S.: t-Closeness: Privacy Beyond k-Anonymity and l-Diversity. In: ICDE'07 2007, pp. 106-115

Li, T., Li, N.: Towards optimal k-anonymization. Knowl. Data Eng. **65**(1), 22-39 (2008). doi:DOI: 10.1016/j.datak.2007.06.015

Lin, D.: An Information-Theoretic Definition of Similarity. Paper presented at the Proceedings of the Fifteenth International Conference on Machine Learning, (1998)

Lin, J.-L., Chang, P.-C., Liu, J.Y.-C., Wen, T.-H.: Comparison of microaggregation approaches on anonymized data quality. Expert Syst. Appl. **37**(12), 8161-8165 (2010a). doi:10.1016/j.eswa.2010.05.071

Lin, J.-L., Wei, M.-C.: An efficient clustering method for k-anonymization. Paper presented at the Proceedings of the 2008 international workshop on Privacy and anonymity in information society, Nantes, France, (2008)

Lin, J.-L., Wen, T.-H., Hsieh, J.-C., Chang, P.-C.: Density-based microaggregation for statistical disclosure control. Expert Syst. Appl. **37**(4), 3256-3263 (2010b). doi:10.1016/j.eswa.2009.09.054

Little, E.G., Rogova, G.L.: Designing ontologies for higher level fusion. Inf. Fusion **10**(1), 70-82 (2009). doi:DOI: 10.1016/j.inffus.2008.05.006

Loukides, G., Shao, J.: Capturing data usefulness and privacy protection in K-anonymisation. Paper presented at the Proceedings of the 2007 ACM symposium on Applied computing, Seoul, Korea, (2007)

Macqueen, J.B.: Some Methods for classification and analysis of multivariate observations. In: Procedings of the Fifth Berkeley Symposium on Math, Statistics, and Probability 1967, pp. 281-297. University of California Press

Machanavajjhala, A., Kifer, D., Gehrke, J., Venkitasubramaniam, M.: *l*-diversity: Privacy beyond *k*-anonymity. ACM Trans. Knowl. Discov. Data **1**(1), 3 (2007). doi:10.1145/1217299.1217302

Malin, B., Sweeney, L.: How (not) to protect genomic data privacy in a distributed network: using trail re-identification to evaluate and design anonymity protection systems. J. of Biomedical Informatics **37**(3), 179-192 (2004). doi:10.1016/j.jbi.2004.04.005

Maniruzzaman, M., Chaves, J., McGee, C., Ma, S., Jr, R.S.: CHTE quench probe system: a new quenchant characterization system. Paper presented at the 5th International Conference on Frontiers of Design and Manufacturing (ICFDM), (2002)

Martinez, S., Sanchez, D., Valls, A.: Ontology-Based Anonymization of Categorical Values. In: Torra, V., Narukawa, Y., Daumas, M. (eds.) Modeling Decisions for Artificial Intelligence, vol. 6408. Lecture Notes in Computer Science, pp. 243-254. Springer Berlin / Heidelberg, (2010a)

Martínez, S., Sánchez, D., Valls, A.: Towards k-anonymous non-numerical data via Semantic Resampling. In: Information Processing and Management of Uncertainty in Knowledge-Based Systems, Catania, Italy 2012a, pp. 519-528. S. Greco et al. (Eds.)

Martinez, S., Sanchez, D., Valls, A., Batet, M.: The Role of Ontologies in the Anonymization of Textual Variables. Paper presented at the Proceeding of the 2010 conference on Artificial Intelligence Research and Development: Proceedings of the 13th International Conference of the Catalan Association for Artificial Intelligence, (2010b)

Martinez, S., Sanchez, D., Valls, A., Batet, M.: The Role of Ontologies in the Anonymization of Textual Variables. In: the 13th International Conference of the Catalan Association for Artificial Intelligence 2010c, pp. 153-162

Martinez, S., Sanchez, D., Valls, A., Batet, M.: Privacy protection of textual attributes through a semantic-based masking method. Inf. Fusion **13**(4), 304-314 (2011). doi:DOI: 10.1016/j.inffus.2011.03.004

Martínez, S., Sánchez, D., Valls, A., Batet, M.: Privacy protection of textual attributes through a semantic-based masking method. Information Fusion (In Press). doi:DOI: 10.1016/j.inffus.2011.03.004

Martínez, S., Valls, A., Sánchez, D.: Semantically-grounded construction of centroids for datasets with textual attributes. Knowl.-Based Syst. **35**(0), 160-172 (2012b). doi:10.1016/j.knosys.2012.04.030

Mclachlan, G.J., Krishnan, T.: The EM Algorithm and Extensions. Wiley Series in Probability and Statistics, (1997)

Medrano-Gracia, P., Pont-Tuset, J., Nin, J., Muntés-Mulero, V.: Ordered Data Set Vectorization for Linear Regression on Data Privacy. In: Torra, V., Narukawa, Y., Yoshida, Y. (eds.) Modeling Decisions for Artificial Intelligence, vol. 4617. Lecture Notes in Computer Science, pp. 361-372. Springer Berlin / Heidelberg, (2007)

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 **10**(1), 70 (2010). doi:citeulike-article-id:7585186

Miller, G.A., Charles, W.G.: Contextual correlates of semantic similarity. Language and Cognitive Processes **6**(1), 1-28 (1991). doi:10.1080/01690969108406936

Mohammed, N., Chen, R., Fung, B.C.M., Yu, P.S.: Differentially private data release for data mining. Paper presented at the Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining, San Diego, California, USA, (2011)

Murillo, J., Abril, D., Torra, V.: Heuristic Supervised Approach for Record Linkage. In: Torra, V., Narukawa, Y., López, B., Villaret, M. (eds.) Modeling Decisions for Artificial Intelligence. Lecture Notes in Computer Science, pp. 210-221. Springer Berlin Heidelberg, (2012)

Nelson, S.J., Johnston, D., Humphreys, B.L.: Relationships in Medical Subject Headings. In: Publishers, K.A. (ed.) Relationships in the Organization of Knowledge. pp. 171-184. New York (2001)

NHS: Share with care, peoples views on consent and confidentiality of patient information, London, NHS Information Authority. (2002)

Nin, J., Herranz, J., Torra, V.: On the disclosure risk of multivariate microaggregation. Knowl. Data Eng. **67**(3), 399-412 (2008a). doi:DOI: 10.1016/j.datak.2008.06.014

Nin, J., Herranz, J., Torra, V.: Rethinking rank swapping to decrease disclosure risk. Knowl. Data Eng. **64**(1), 346-364 (2008b). doi:10.1016/j.datak.2007.07.006

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 **18**(4), 345-353 (2001)

Ohno-Machado, L., Silveira, P.S.P., Vinterbo, S.A.: Protecting patient privacy by quantifiable control of disclosures in disseminated databases. Int. J. Med. Inform. **73**(7-8), 599-606 (2004)

Oliveira, S.R.M., Zaïane, O.R.: A privacy-preserving clustering approach toward secure and effective data analysis for business collaboration. Computers & Security **26**(1), 81-93 (2007). doi:DOI: 10.1016/j.cose.2006.08.003

Park, H.-S., Jun, C.-H.: A simple and fast algorithm for K-medoids clustering. Expert Syst. Appl. **36**(2, Part 2), 3336-3341 (2009). doi:DOI: 10.1016/j.eswa.2008.01.039

Patwardhan, S., Pedersen, T.: Using WordNet-based Context Vectors to Estimate the Semantic Relatedness of Concepts. In: EACL 2006 Workshop Making Sense of Sense---Bringing Computational Linguistics and Psycholinguistics Together 2006, pp. 1-8

Pedersen, T., Pakhomov, S.V.S., Patwardhan, S., Chute, C.G.: Measures of semantic similarity and relatedness in the biomedical domain. J Biomed Inform **40**(3), 288-299 (2007). doi:10.1016/j.jbi.2006.06.004

Pedersen, T., Patwardhan, S., Michelizzi, J.: WordNet::Similarity: measuring the relatedness of concepts. Paper presented at the Demonstration Papers at HLT-NAACL 2004, Boston, Massachusetts, (2004)

Petrakis, G., Varelas, G., Hliaoutakis, A., Raftopoulou, R.: X-Similarity:Computing Semantic Similarity between Concepts from Different Ontologies. Journal of Digital Information Management (JDIM) **4**, 233-237 (2006)

Pirro, G., Seco, N.: Design, Implementation and Evaluation of a New Semantic Similarity Metric Combining Features and Intrinsic Information Content. Paper presented at the Proceedings of the OTM 2008 Confederated International Conferences, CoopIS, DOA, GADA, IS, and ODBASE 2008. Part II on On the Move to Meaningful Internet Systems, Monterrey, Mexico, (2008)

Porter, M.F.: An algorithm for suffix stripping. In: Readings in information retrieval. pp. 313-316. Morgan Kaufmann Publishers Inc., (1997)

Purdam, K., Elliot, M.: A case study of the impact of statistical disclosure control on data quality in the individual UK Samples of Anonymised Records. Environment and Planning **A 39**(5), 1101-1118 (2007)

Rada, R., Mili, H., Bicknell, E., Blettner, M.: Development and application of a metric on semantic nets. IEEE Trans. Syst. Man Cybern. **19**(1), 17-30 (1989)

Reiss, S.P.: Practical data-swapping: the first steps. ACM Trans. Database Syst. **9**(1), 20-37 (1984). doi:10.1145/348.349

Resnik, P.: Using information content to evaluate semantic similarity in a taxonomy. Paper presented at the Proceedings of the 14th international joint conference on Artificial intelligence - Volume 1, Montreal, Quebec, Canada, (1995)

Rodriguez, M.A., Egenhofer, M.J.: Determining Semantic Similarity among Entity Classes from Different Ontologies. IEEE Trans. on Knowl. and Data Eng. **15**(2), 442-456 (2003). doi:10.1109/tkde.2003.1185844

Rogers, J.: Publically reported breaches in EPR confidentiality. http://www.cs.man.ac.uk/~jeremy/HealthInf/Confidentiality.html (2005)

Rubenstein, H., Goodenough, J.B.: Contextual correlates of synonymy. Commun. ACM **8**(10), 627-633 (1965). doi:10.1145/365628.365657

Samarati, P., Sweeney, L.: Protecting Privacy when Disclosing Information: k-Anonymity and Its Enforcement through Generalization and Suppression. Technical Report SRI-CSL-98-04, SRI Computer Science Laboratory (1998)

Sánchez, D.: A methodology to learn ontological attributes from the Web. Knowl. Data Eng. **69**(6), 573-597 (2010). doi:http://dx.doi.org/10.1016/j.datak.2010.01.006

Sanchez, D., Batet, M.: Semantic similarity estimation in the biomedical domain: An ontology-based information-theoretic perspective. J Biomed Inform **44**(5), 749-759 (2011). doi:10.1016/j.jbi.2011.03.013

Sánchez, D., Batet, M.: A semantic similarity method based on information content exploiting multiple ontologies. Expert Syst. Appl. **40**(4), 1393-1399 (2013). doi:http://dx.doi.org/10.1016/j.eswa.2012.08.049

Sánchez, D., Batet, M., Isern, D., Valls, A.: Ontology-based semantic similarity: A new feature-based approach. Expert Syst. Appl. **39**(9), 7718-7728 (2012a). doi:10.1016/j.eswa.2012.01.082

Sánchez, D., Castellà-Roca, J., Viejo, A.: Knowledge-based scheme to create privacy-preserving but semantically-related queries for web search engines. Information Sciences **218**(0), 17-30 (2013). doi:http://dx.doi.org/10.1016/j.ins.2012.06.025

Sánchez, D., Isern, D.: Automatic extraction of acronym definitions from the Web. Applied Intelligence **34**(2), 311-327 (2011). doi:10.1007/s10489-009-0197-4

Sánchez, D., Isern, D., Millan, M.: Content annotation for the semantic web: an automatic web-based approach. Knowl. Inf. Syst., 1-26 (2010). doi:10.1007/s10115-010-0302-3

Sanchez, D., Moreno, A.: Learning non-taxonomic relationships from web documents for domain ontology construction. Knowl. Data Eng. **64**(3), 600-623 (2008a). doi:http://dx.doi.org/10.1016/j.datak.2007.10.001

Sanchez, D., Moreno, A.: Pattern-based automatic taxonomy learning from the Web. AI Commun. **21**(1), 27-48 (2008b)

Sánchez, D., Moreno, A., Del Vasto-Terrientes, L.: Learning relation axioms from text: An automatic Web-based approach. Expert Syst. Appl. **39**(5), 5792-5805 (2012b). doi:http://dx.doi.org/10.1016/j.eswa.2011.11.088

Sanchez, D., Sole-Ribalta, A., Batet, M., Serratosa, F.: Enabling semantic similarity estimation across multiple ontologies: An evaluation in the biomedical domain. J. of Biomedical Informatics **45**(1), 141-155 (2012). doi:10.1016/j.jbi.2011.10.005

Shin, H., Vaidya, J., Atluri, V.: Anonymization models for directional location based service environments. Computers & Security **29**(1), 59-73 (2010). doi:DOI: 10.1016/j.cose.2009.07.006

Shin, K., Abraham, A., Han, S.-Y.: Enhanced Centroid-Based Classification Technique by Filtering Outliers. In: Proceedings of TSD. pp. 159-163. (2006)

Singh, A.C., Yu, F., Dunteman, G.H.: MASSC: A new data mask for limiting statistical information loss and disclosure. In: Proceedings of the Joint UNECE/EUROSTAT Work Session on Statistical Data Confidentiality, Luxembourg 2003, pp. 373-394

Spackman, K.A., Campbell, K.E., Cote, R.A.: SNOMED RT: a reference terminology for health care. Proc AMIA Annu Fall Symp, 640-644 (1997)

Sweeney, L.: Achieving *k*-anonymity privacy protection using generalization and suppression. Int. J. Uncertain. Fuzziness Knowl.-Based Syst. **10**(5), 571-588 (2002a). doi:10.1142/s021848850200165x

Sweeney, L.: k-anonymity: a model for protecting privacy. Int. J. Uncertain Fuzziness Knowl-Based Syst. **10**(5), 557-570 (2002b). doi:http://dx.doi.org/10.1142/S0218488502001648

Terrovitis, M., Mamoulis, N., Kalnis, P.: Privacy-preserving anonymization of set-valued data. Proc. VLDB Endow. **1**(1), 115-125 (2008). doi:http://doi.acm.org/10.1145/1453856.1453874

Torra, V.: Microaggregation for Categorical Variables: A Median Based Approach. In: Domingo-Ferrer, J., Torra, V. (eds.) Privacy in Statistical

Databases, vol. 3050. Lecture Notes in Computer Science, pp. 518-518. Springer Berlin / Heidelberg, (2004)

Torra, V.: Towards knowledge intensive data privacy. Paper presented at the Proceedings of the 5th international Workshop on data privacy management, and 3rd international conference on Autonomous spontaneous security, Athens, Greece, (2011)

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

Torra, V., Miyamoto, S.: Evaluating Fuzzy Clustering Algorithms for Microdata Protection. In: Domingo-Ferrer, J., Torra, V. (eds.) Privacy in Statistical Databases, vol. 3050. Lecture Notes in Computer Science, pp. 519-519. Springer Berlin / Heidelberg, (2004)

Torra, V., Narukawa, Y.: Modeling Decisions. Cognitive Technologies, (2007)

Truta, T.M., Vinay, B.: Privacy Protection: p-Sensitive k-Anonymity Property. In: Data Engineering Workshops, 2006. Proceedings. 22nd International Conference on, 2006 2006, pp. 94-94

Tversky, A.: Features of similarity. Psychological Review **84**(4), 327-352 (1977)

Valls, A., Gibert, K., Sánchez, D., Batet, M.: Using ontologies for structuring organizational knowledge in Home Care assistance. Int. J. Med. Inform. **79**(5), 370-387 (2010). doi:DOI: 10.1016/j.ijmedinf.2010.01.012

Varde, A.S., Rundensteiner, E.A., Ruiz, C., Brown, D.C., Maniruzzaman, M., Sisson, R.D.: Designing semantics-preserving cluster representatives for scientific input conditions. Paper presented at the Proceedings of the 15th ACM international conference on Information and knowledge management, Arlington, Virginia, USA, (2006)

Ward, J.H.: Hierarchical Grouping to Optimize an Objective Function. J. Am. Stat. Assoc. **58**(301), 236-244 (1963). doi:citeulike-article-id:1470845

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

Willenborg, L., Waal, T.d.: Elements of Statistical Diclosure Control, vol. 155. Lecture Notes in Statistics. Springer, (2001)

Winkler, W.E.: Re-identification Methods for Masked Microdata. In: Domingo-Ferrer, J., Torra, V. (eds.) Privacy in Statistical Databases, vol. 3050. Lecture Notes in Computer Science, pp. 519-519. Springer Berlin / Heidelberg, (2004)

Wong, S.V., Hamouda, A.M.S.: The development of an online knowledge-based expert system for machinability data selection. Knowl.-Based Syst. **16**(4), 215-229 (2003). doi:10.1016/s0950-7051(02)00083-7

Wu, Z., Palmer, M.: Verbs semantics and lexical selection. Paper presented at the the 32nd annual meeting on Association for Computational Linguistics, Las Cruces, New Mexico, (1994)

Xu, J., Wang, W., Pei, J., Wang, X., Shi, B., Fu, A.W.-C.: Utility-based anonymization for privacy preservation with less information loss. SIGKDD Explor. Newsl. **8**(2), 21-30 (2006a). doi:http://doi.acm.org/10.1145/1233321.1233324

Xu, J., Wang, W., Pei, J., Wang, X., Shi, B., Fu, A.W.-C.: Utility-based anonymization using local recoding. Paper presented at the Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining, Philadelphia, PA, USA, (2006b)

Yancey, W., Winkler, W., Creecy, R.: Disclosure Risk Assessment in Perturbative Microdata Protection. In: Domingo-Ferrer, J. (ed.) Inference Control in Statistical Databases, vol. 2316. Lecture Notes in Computer Science, pp. 49-60. Springer Berlin / Heidelberg, (2002)

Yang, X.Q., Sun, N., Sun, T.L., Cao, X.Y., Zheng, S.J.: The Application of Latent Semantic Indexing and Ontology in Text Classification. International Journal of Innovative Computing, Information and Control **5**(12), 4491-4499 (2009)

Yihui, L.: Dimensionality reduction and main component extraction of mass spectrometry cancer data. Knowl.-Based Syst. **26**(0), 207-215 (2012). doi:10.1016/j.knosys.2011.08.006

Zhang, W., Yoshida, T., Tang, X., Wang, Q.: Text clustering using frequent itemsets. Knowl.-Based Syst. **23**(5), 379-388 (2010). doi:10.1016/j.knosys.2010.01.011