Semantic Perturbative Privacy-preserving Methods for Nominal Data

María Mercedes Rodríguez García

UNIVERSITAT ROVIRA i VIRGILI

Semantic Perturbative Privacy-preserving Methods for Nominal Data

María Mercedes Rodríguez García

DOCTORAL THESIS
2017

María Mercedes Rodríguez García

Semantic Perturbative Privacy-preserving Methods for Nominal Data

DOCTORAL THESIS

Advisors
Dr. Montserrat Batet Sanromà and Dr. David Sánchez Ruenes

Department of Computer Engineering and Mathematics

UNIVERSITAT ROVIRA i VIRGILI

February, 2017

I STATE that the present study, entitled “Semantic Perturbative Privacy-preserving Methods for Nominal Data”, presented by María Mercedes Rodríguez García for the award of the degree of Doctor, has been carried out under our supervision at the Department of Computer Engineering and Mathematics of this university.

Tarragona, February 15, 2017

Doctoral Thesis Supervisor/s

MONTSERRAT BATET SANROMA

Digitally signed by MONTSERRAT BATET SANROMA
DN: c=ES, ou=Vegeu https://www.catcert.cat/veridCAT (c)03, ou=Serveis Publics de Certificacio CPIXSA-2, sn=BATET SANROMA, givenName=MONTSERRAT, serialNumber=39708865D, cn=MONTSERRAT BATET SANROMA
Date: 2017.03.15 18:06:40 +01'00'

Dr. Montserrat Batet Sanromà

CPISR-1 C David Sánchez Ruenes

Digitally signed by CPISR-1 C David Sánchez Ruenes
DN: c=ES, o=Universitat Rovira i Virgili, ou=Vegeu https://www.aoc.cat/CATCert/Regulacio, sn=Sánchez Ruenes, givenName=David, serialNumber=47763566H, cn=CPISR-1 C David Sánchez Ruenes
Date: 2017.03.15 18:04:44 +01'00'

Dr. David Sánchez Ruenes

Acknowledgements

I would like to thank my advisors Dr. Montserrat Batet and Dr. David Sánchez for their guidance, support, motivation and knowledge during the development of this thesis. I am also indebted to all CRISES group members, for their hospitality during my stay in Tarragona in winter 2017. Finally, I would like to thank my husband for his motivation and continuous support during this period of time.

This work was partially funded by the European Commission (projects H2020-644024 “CLARUS” and H2020-700540 “CANVAS”), by the Spanish Government (projects TIN2015-70054-REDC “Red de excelencia Consolider ARES” and TIN2016-80250-R “Sec-MCloud”) and by the Government of Catalonia under grant 2014 SGR 537.

Contents

List of Figures

List of Tables

Abstract

The exploitation of personal microdata (such as census data, preferences or medical records) is of great interest for the data mining community. Such data often include sensitive information that can be directly or indirectly related to individuals. Therefore, privacy-preserving measures should be undertaken to minimize the risk of re-identification and, hence, of disclosing confidential information on the individuals. In the past, many privacy-preserving methods have been developed to deal with numerical data, but approaches tackling the protection of nominal values are scarce. Since the utility of this kind of data is closely related to the preservation of their semantics, in this work, we exploit several semantic technologies to enable a semantically-coherent protection of nominal data. Specifically, we use ontologies as the ground to propose a semantic framework that enables an appropriate management of nominal data in data protection tasks; such framework consists on a set of operators that characterize and transform nominal data while taking into account their semantics. Then, we use this framework to adapt perturbative privacy-preserving methods to the nominal domain. Specifically, we focus on methods based on the two main principles underlying to data protection: permutation-based approaches, i.e., *rank swapping*, and *noise addition*. The proposed methods have been extensively evaluated with real datasets. Experimental results show that a semantically-coherent management of nominal data significantly improves the semantic interpretability and the utility of the protected outcomes.

Resum

L'explotació de microdades personals (p. ex., dades censals, preferències, o registres de salut) és de gran interès per a la mineria de dades. Aquestes dades sovint contenen informació sensible que pot ser directament o indirectament relacionada amb els individus. Per tant, cal implementar mesures per a preservar la privadesa i minimitzar el risc de re-identificació i, conseqüentment, de revelació d'informació confidencial sobre els individus. Tot i que s'han desenvolupat nombroses mètodes per preservar la privadesa de dades numèriques, la protecció de valors nominals ha rebut escassa atenció. Donat que la utilitat d'aquest tipus de dades està estretament relacionada amb la preservació de la seva semàntica, en aquest treball explotem diverses tecnologies semàntiques per fe possible una protecció coherent amb el significat de les dades nominals. Específicament, fem servir ontologies com a base per a proposar un marc de treball semàntic que permeti manegar dades nominals segons en seu significat en tasques de protecció; aquest marc consta d'un conjunt d'operadors que caracteritzen i transformen dades nominals a la vegada que consideren la seva semàntica. A partir d'aquí, fer servir aquest marc per adaptar mètodes pertorbatius de protecció de la privadesa. Particularment, ens centrem en mètodes basats als dos principis subjacents a la protecció de dades: enfocaments basats en permutació, concretament, *rank swapping*, y addicció de soroll. Els mètodes proposats han estat avaluats extensament amb conjunts de dades reals. Els resultats experimentals mostren que manegar les dades nominals semànticament millora significativament la interpretabilitat i la utilitat dels resultats protegits.

Resumen

La explotación de microdatos personales (p. ej., datos del censo, preferencias, o registros de salud) es de gran interés para la minería de datos. Tales datos a menudo contienen información sensible que puede ser directa o indirectamente relacionada con los individuos. Por tanto, resulta necesario implementar medidas para preservar la privacidad y para minimizar el riesgo de re-identificación y, por consiguiente, de revelación de información confidencial sobre los individuos. Pese a que se han desarrollado numerosos métodos para preservar la privacidad de datos numéricos, la protección de valores nominales ha recibido escasa atención. Puesto que la utilidad de este tipo de datos está estrechamente relacionada con la preservación de su semántica, en este trabajo explotamos varias tecnologías semánticas para posibilitar una protección coherente con el significado de los datos nominales. Específicamente, utilizamos ontologías como base para proponer un marco de trabajo semántico que permita manejar datos nominales según su significado en tareas de protección; dicho marco consta de un conjunto de operadores que caracterizan y transforman datos nominales a la vez que tienen en consideración su semántica. A partir de aquí, utilizamos este marco para adaptar métodos perturbativos de preservación de la privacidad al dominio nominal. Particularmente, nos centramos en métodos basados en los dos principios subyacentes a la protección de los datos: enfoques basados en permutación, concretamente, *rank swapping*, and *adición de ruido*. Los métodos propuestos han sido extensamente evaluados con conjuntos de datos reales. Resultados experimentales muestran que manejar los datos nominales semánticamente mejora significativamente la interpretabilidad y la utilidad de los resultados protegidos.

Chapter 1 **Introduction**

Data of individuals arising from surveys or electronic records are of great interest for public and private organizations. The publication of this information allows conducting a variety of statistical studies, for instance, on health, education, trade preferences, living conditions or employability. Particularly, these data are crucial to improve decision-making in business [1] and healthcare [2], or to offer personalized services that enhance the online experience [1].

However, when data about individuals are made available for secondary use, special care must be taken to avoid privacy violations. Specifically, personal data usually contain personally identifiable information (PII), which may enable the re-identification of individuals, and confidential information, which may disclose sensitive information on re-identified subjects. Due to the plausibility of these threats, and because the protection of individuals' privacy is a fundamental social right, government agencies and current legislations emphasize the need of protecting personal data from disclosure. For that, individuals' detailed data (i.e., *microdata*) must be subject to an anonymization process before their release, so that subjects cannot be re-identified in the protected dataset and, thus, confidential data cannot be univocally associated to an identity. In this respect, some studies [3, 4] have shown that removing identifying attributes, such as identity numbers or names, it is not enough to anonymize data, because combinations of certain non-identifying attributes, known as *quasi-identifiers* (e.g. occupation + sex + ZIP code), may be linked with external data sources (e.g. voter registration) to enable re-identifications. Re-identification via data linkage constitutes a real and serious privacy threat that, nowadays, allows data brokers to compile and aggregate individuals' data gathered from different sources (e.g., census data, social media or on-line transactions); this information is then used to perform inferences about individuals' habits or preferences, and to construct user profiles that can be exploited to conduct marketing campaigns, but also to engage potentially discriminatory actions (e.g., in job or health insurance applications) [1].

1.1 Motivation and approach

To minimize the chance of re-identification, quasi-identifying attributes should be subjected to anonymization. In turn, because the ultimate motivation underlying to data releases is to conduct analyses on the such data, anonymization should be done in a way that the protected data still retain as much analytical utility as possible; that is, the conclusions or inferences extracted from the analysis of the anonymized dataset should be similar to those of the original dataset. With the goal of balancing these two, a priori, contradictory goals (i.e., privacy and utility preservation), different masking methods have been proposed within the disciplines of Statistical Disclosure Control (SDC) [5], Privacy-Preserving Data Publishing (PPDP) [6, 7]. The proposed methods generate a modified version of the original data by generalizing, distorting or introducing ambiguity on the quasi-identifying attributes while preserving certain statistics features. Among them, perturbative masking methods (such as those based on *noise addition*, *data aggregation* or *data permutation*) are the most widespread, because they offer a good trade-off between utility and privacy dimensions [5].

However, most perturbative masking methods have been designed to deal with continuous numerical data and, at most, with ordinal categorical data [5]. Specifically, in order to transform data while retaining some of their statistical features, perturbative methods extensively rely on arithmetical and statistical operators meant exclusively for numerical (or, at least, ordinal) data. This contrasts with the fact that most of the data that are currently being gathered and exploited on individuals are of nominal nature [1], e.g., attributes and messages posted in social media [8], healthcare outcomes stored in electronic healthcare records [9], queries performed to web search engines [10] or logs resulting from on online transactions [11]. Unlike numerical data, nominal data are finite, discrete, textual and non-ordinal. In this scenario, it is generally not possible to carry out the arithmetical transformations required by perturbative masking. Moreover, whereas the utility of numerical data depends on the preservation of their statistical features, for nominal values, which refer to concepts or instances, data utility is closely related to the preservation of *semantics* [12]; thus, data transformations carried out during the protection process require from mechanisms that consider the meaning of words.

Data semantics have been traditionally neglected (or scarcely considered) in the literature on privacy protection [12, 13]. Therefore, new proposals that are able to capture, manage and preserve the semantics underlying to nominal data during the masking process are needed.

This thesis aims at contributing to this need by studying how to capture and integrate data semantics within the context of perturbative microdata anonymization. Its originality consists on the management and transformation of nominal attributes from a semantic point of view, rather than from a symbolic way.

The semantic interpretation of nominal data for masking purposes requires the exploitation of structured knowledge sources, which allow mapping values in nominal attributes with their conceptual abstractions and, as a result, analyze the semantics underlying to them. To do so, in this work we rely in the well-known ontological paradigm. Ontologies are rigorous and exhaustive organizations of knowledge domains, modeling concepts and their interrelations [14]. Works in other fields [15] demonstrate that, by exploiting ontologies, we can design semantically grounded mechanisms are able to better interpret, analyze and manage textual resources.

1.2 Objectives

The main objective of this thesis is the design of perturbative masking methods well suited for the anonymization of nominal data from a semantic point of view. That is, we aim at obtaining an anonymized dataset that is as semantically similar to the original data as possible, while offering privacy guarantees equivalent to those of standard numerical methods. In this manner, the utility of protected nominal data, which closely depends on the preservation of their underlying semantics, can be better preserved.

To achieve this objective, the following specific goals are defined:

1. To study the privacy threats underlying data releases and survey works on data protection framed in the areas of Statistical Disclosure Control (SDC) and Privacy Preserving Data Publishing (PPDP). These methods will be characterized according to their operating principles, the types of data they are able to deal with and the data utility aspects they better

preserve. Specifically, we will compile and review the state of the art on data protection methods for nominal data.

2. To study the possibilities offered by structured knowledge sources (ontologies) to capture the semantics conveyed by nominal data. Specifically, we will rely on the notion of ontology-based semantic similarity [16], which enables a semantically-coherent management of textual data, and which we will use to guide data anonymization from a semantic perspective.
3. To design a framework for nominal data management consisting on a set of semantic operators that enables the characterization, comparison and transformation of nominal data from a semantic point of view. Ontologies and ontology-based semantic similarity will be used as the basic pillar to propose these semantically-grounded operators.
4. To apply our framework to adapt SDC and PPDP methods initially designed exclusively for numerical data so that they can deal with nominal data in a semantically-coherent way. Specifically, we will focus on methods based on the two main principles underlying to data protection [17]: permutation-based approaches (i.e., *rank swapping*) and data distortion mechanisms based on *noise addition*.
5. To implement the proposed methods and to evaluate and compare their performance on nominal data against related works w.r.t. their ability to preserve the semantic features of the data.

1.3 Document structure

The remaining of this document is organized in the following chapters:

- Chapter 2 introduces the main notions on data privacy and privacy protection, and surveys and characterizes works in SDC and PPDP.
- Chapter 3 reviews related works applying or proposing data protection mechanisms to nominal data, and highlights their limitations w.r.t. the preservation of data semantics.
- Chapter 4 introduces the notion of ontologies and surveys the literature on semantic similarity. We also present our framework that, by relying on the former, proposes a variety of semantically-grounded operators to manage nominal data.

- Chapter 5 details our adaptation to the nominal domain of a permutation-based data protection mechanism: *rank swapping*. Univariate and multivariate algorithms are proposed.
- Chapter 6 presents our notion of *semantic noise*, and details our adaptation of several noise addition mechanisms (uncorrelated and correlated noise) to nominal data.
- Chapter 7 empirically evaluates our methods with real nominal data and compares the results they provide against non-semantic mechanisms. The utility of the protected data is measured according to the preservation of several marginal and joint semantic features of the data.
- Chapter 8 summarizes the main contributions of this thesis and presents some lines of future research.

Chapter 2 **Background on privacy protection**

Data about individuals are collected by governments and companies for a variety of purposes. These data stores are valuable resources for research and market analysis and, therefore, there is a growing demand to access them. However, the dissemination of individuals' data is a controversial task. On the one side, there is a demand to access accurate data, that is, the released data should retain their analytical utility; on the other side, there is a risk of disclosing confidential information about specific individuals, that is, data should be protected before making them available for secondary use. In this chapter, we discuss such issues and review the privacy-preserving methods that exist in the literature aiming to satisfy simultaneously utility and security conditions.

2.1 Data privacy

In the current era of big data and digital societies, information collection, storage and processing capabilities have meaningfully grown. Social networks, electronic records or web browsing generate huge volumes of information about individuals which are of great interest for public and private organizations. Collection and processing (e.g., data mining) of these data allows conducting a variety of surveys, improving decision-making in business or offering personalized services to enhance the online experience. However, the dissemination of personal data may compromise the individuals' privacy, which is considered a fundamental right, and it is supported by international treaties and constitutional laws, such as the Universal Declaration of Human Rights (1948), which devotes its Article 12 to privacy.

In this scenario, governmental agencies and current legislations on data protection emphasize the need of adequately protecting Personally Identifiable Information (PII) [18] to preserve individuals' privacy. PII includes not only identifying data, such as social security numbers, but also any non-identifying data that, in combination with other non-identifying features, can be used by external entities to re-identify individuals by linking them with external data sources [3, 4]. The latter constitutes a real and serious privacy threat and, in

fact, is being currently employed by data brokers to compile and aggregate individuals' data and, from these, build user profiles that are latter used or sold to third parties for commercial and business purposes [1]. Regarding privacy violations in data dissemination [19, 20], Title 13, Chapter 1.1 of the U.S. Code states that "no individual should be re-identifiable in the released data". Against this background, reaching a tradeoff between individual's privacy protection and protected data that are still useful for analysis is the key point to guarantee individuals' rights while ensuring the continued growth of the digital society. Among the main areas where the data release takes place, we highlight the followings:

- *Official statistics*. National Statistical Institutes (NSIs) collect and publish a wide range of high quality statistical information about the population, which allow conducting a variety of statistical studies, e.g., about education, living conditions or employability.

- *Health information*. Electronic Health Records (EHRs) defined as digital collections of health information about individual patients, are especially valuable resources for clinical research and education, and a vehicle to improve quality of health care delivery and reduce medical errors [21]. The analysis of EHRs allows conducting a variety of studies about treatment models, clinical practice guidelines, prevention measures, adverse drug reactions or drug interactions [2]. Because clinical data are considered of sensitive nature by the EU Data Protection Act 1998 and the Human Rights Act, the use of the health care data for research purposes must guarantee confidentiality of the patients to which the data refer. For that, different regulations have been adopted, such as the Health Insurance Portability and Accountability Act (HIPAA) [22] in the United States or the General Data Protection Regulation (GDPR) [19] in the European Union.

- *Online services*. The extensive use of online services, whether to buy online, use social networks or make queries to web search engines, leave traces of personal information that the providers can compile and use for their purposes or sell to third parties. These services have dramatically increased the availability, variety and volume of users' data. Different commercial data sources may be used by companies to obtain a detailed profile of consumers. In a survey carried out by Federal Trade

Commission from the United States [1], the authors highlight the following sources of data acquisition: customer lists from registration websites, online advertising networks, consumers' web browsing activities and directly from their merchant and financial service company clients. This information allows companies (i) to identify groups of consumers, (ii) to predict an interest, analyze the characteristics the consumers share, and use the shared characteristic data to create a predictive model to apply to other consumers, (iii) to create or enhance products and services, and (iv) for more individualized and controversial uses, such as creating user profiles that can then be used for potentially discriminatory purposes (e.g., in jobs or health insurance). Such information is subject to strict regulation to not result in public profiling of individuals, such as [23] for regulations in the European Union.

2.2 Types of data releases

Individuals' information can be released in three main ways [3, 24]: as *macrodata*, which consist of aggregated values of groups of individuals published in *contingency tables* for the frequency distribution of the variables or *magnitude tables* for other aggregate magnitudes; as *queryable databases*, that is, interactive databases to which user can submit statistical queries to obtain aggregate information, such as counts or averages; and as *microdata*, where each record of the dataset details the attributes of a specific individual, i.e., individuals' raw data.

Unlike aggregate data, microdata confer flexibility to perform a per-individual analysis and, on the contrary to queryable databases, they do not restrict the type and number of data analyses [25]. However, the publication of microdata may lead to disclosure of confidential information related to the individuals from whom the data have been collected. In this regard, a data collector must guarantee that no sensitive information about specific individuals is disclosed. To satisfy such guarantee, it is necessary that microdata be subjected to an anonymization process before their release.

2.2.1 Microdata sets

A microdata set can be represented as a table (matrix), where each record (row) contains information about a single individual from those who took part in the data collection process, and each attribute (column) contains information regarding one of the features collected. We use X to denote the collected microdata set and assume that X contains information about n respondents and m attributes. We use x_i to refer to the record contributed by respondent i, and $X^a = (x_1^a, \ldots, x_n^a)$ to refer to the a-th attribute. The value of the a-th attribute for the respondent i is denoted by x_i^a.

The attributes in a microdata set are usually classified in the following categories according to their sensitiveness and the type of disclosure risk they cause [24]:

- *Identifiers*. An attribute is an identifier if it enables a univocal identification of the individual to whom the record refers, e.g., the social security number (SSN). To preserve individuals' privacy, identifiers are usually removed from the dataset before releasing it, so that other (confidential) attributes from the same dataset cannot be directly associated to a specific individual. Specifically, before sharing clinical datasets, the HIPAA requires that the patient data are de-identified by erasing the identifiers classified as Protected Health Information (PHI).

- *Quasi-identifiers*. A quasi-identifier attribute is a non-identifying attribute that, in combination with other non-identifying attributes from the dataset, may result identifying. We use the term *quasi-identifier* to refer to the identifying combination of non-identifying attributes in the dataset (e.g., occupation + sex + ZIP code), and *quasi-identifier attribute* to refer to the attributes that conform a quasi-identifier. A quasi-identifier may be employed by attackers to re-identify individuals by linking it with non-anonymous external data sources (e.g., voter registration). If the exploited external sources contained some identifier, attackers could determine the individuals' identity from the dataset, as demonstrated in several studies [3, 4, 26]. Nowadays, the amount of information externally available in a variety of sources (e.g., electoral rolls, census data or social media) together with the increasing amounts of computational power facilitates

conducting such re-identifications. In practice, any attribute is potentially a quasi-identifier attribute depending on the external information available for the attacker [26]. Unlike identifiers, quasi-identifier attributes must not be removed from the dataset because they provide useful information to data analysis. Therefore, to not jeopardize the individuals' privacy, the release of quasi-identifier attributes must be protected.

- *Confidential attributes*. A confidential attribute is an attribute that contains sensitive information on the individuals, e.g., health condition. Because of their sensitive nature, confidential attributes must be especially protected. This does not only mean preventing the attacker from determining the exact value that a confidential attribute takes for certain individual, but also preventing inferences on the value of that attribute, such as lower-bounding and upper-bounding it. Note that a confidential attribute also can be considered a quasi-identifier attribute.

On the other hand, according to their data type, the attributes in a microdata set can be classified as:

- *Numerical.* An attribute is numerical if it admits arithmetical operations and order relationships. In turn, a numerical attribute can be either *continuous* (e.g., income) or *discrete* (e.g., age). Some operations carried out on discrete numerical values (e.g., aggregation) may require approximating the result or using discrete arithmetical operators (e.g., the mode instead of the mean).

- *Categorical.* An attribute is categorical if it does not admit arithmetical operations. In turn, a categorical attribute can be either *ordinal* if it admits order relationships (e.g., color, where the different colors may be ranked on basis of their wave lengths) or *nominal* if it does not admit order relationships, which is the case of most textual data. Much of the information used by data brokers for categorizing individuals is of nominal type [1] (e.g., occupation, education or personal interests).

When publishing a microdata set, the data collector must guarantee that no sensitive information about specific individuals is disclosed. Disclosure can be classified into two categories [5]:

(i) *Identity disclosure* occurs when an attacker discovers the true identity of an individual in the released dataset.

(ii) *Attribute disclosure* occurs when an attacker discovers the exact value of a confidential attribute of an individual in the released dataset, or infers some information about its value, i.e., bounding it (*interval disclosure*).

2.3 Privacy-preserving methods for microdata releases

To minimize the identity disclosure and, consequently, the possibility of gaining confidential information about a specific individual (attribute disclosure), the data collector must subject the microdata set to an anonymization process prior its release. The main challenge in the anonymization process is to find out a balance between privacy and utility: the disclosure risk must be limited, but the data need to remain useful for analysis. This approach contrasts with the alternative of protecting data via encryption, which incurs no disclosure risk at all, but offers no utility.

To anonymize data, the first action to perform is subject the microdata set to a de-identification process, where the identifying attributes (identifiers) are removed. In this way, none of the remaining attributes in the dataset can be immediately and univocally associated to a specific individual. We assume in subsequent chapters that the considered microdata sets do not contain any identifier attribute, i.e., the datasets have been de-identified.

However, de-identification is not sufficient to avoid identity disclosure. De-identified datasets often contain quasi-identifier attributes whose combination may define a unique tuple and, thus, may lead to re-identification (identity disclosure). To protect quasi-identifier attributes in a microdata set and, at the same time, offer valid data for analysis, different privacy-preserving methods have been proposed within the discipline of Statistical Disclosure Control (SDC) [5], under the umbrella of National Statistical Institutes (NSIs), and within the computer science community under the name of Privacy Preserving Data Publishing (PPDP) [6, 7] and Privacy Preserving Data Mining (PPDM) [27]. Whereas both SDC and PPDP are focused on protecting microdata sets

for their release, PPDM aims at protecting the outcomes of the data mining tasks while keeping secret the original microdata set.

Privacy-preserving methods can be classified into two main categories: *masking methods* and *synthetic methods* [3, 28]. To build the protected dataset, the masking methods modify the records of the original dataset. Depending on the effect on the original data, masking methods are subdivided into *non-perturbative masking methods* and *perturbative masking methods*. By contrast, in the synthetic methods, the protected dataset consists of a set of records randomly drawn from a statistical model adjusted to the original dataset. Because the protected dataset does not directly derive from the original dataset, synthetic data seem to have the advantage of avoiding the re-identification problem. However, some authors [29, 30] state that synthetic data overfitted to original data may lead to the re-identification. On the other hand, synthetic data only preserve the statistical properties explicitly selected by the data protector, which leads to the question whether the data protector should not directly publish the statistics he wants preserved rather than a synthetic microdata set [24].

To provide a stronger protection, privacy-preserving methods can also be applied to confidential attributes. In this way, even if identity disclosure happens, there may not be attribute disclosure.

As a result of the above, data collectors publish a modified version X^* of the original microdata set X, called protected or anonymized dataset, where the identifiers have been removed and the quasi-identifiers and/or the confidential attributes (according to the policy of the statistical agency [5]) have been masked.

2.3.1 Non-perturbative masking methods

Non-perturbative masking methods protect the original dataset either by suppressing some of the data or by reducing their level of detail; in both cases, they preserve data truthfulness. We depict the main techniques below:

- *Sampling* [28]: this method is based on publishing an unmodified record sample S of the original dataset X. For example, S may be composed of the set of the even records of X. Since there is an uncertainty about whether or not a specific respondent is in the released sample, the

disclosure risk decreases. This technique is suitable for categorical data, but not for continuous data, because values in a continuous attribute are probably unique for each respondent, i.e., it is highly unlikely that two respondents will take the same value for the continuous attribute and, thus, unique matches of *S* with the external data sources could still happen.

- *Generalization* [31]: this technique, also known as global recoding, replaces the attribute values by more general values, thus reducing the detail of the original information. Generalization is usually performed at the attribute level, that is, either all or none of the records are generalized. To mask an attribute, it is necessary to represent the values of the attribute domain in a generalization hierarchy, where the most general value is at the root of the hierarchy and the most specific values correspond to the leaves. The generalization process proceeds by replacing the values represented by the leaf nodes with one of their ancestor nodes at a higher level. As an example, for a nominal attribute that details the occupation of a set of individuals, values as *hotel clerk* and *file clerk* could be replaced by the category *clerk*. For a continuous attribute, original values are replaced by numerical intervals. Top and bottom coding is a special case of generalization in which the top values (those above a certain threshold) and the bottom values (those below a certain threshold) in the attribute are respectively replaced by a value that represents the upper limit (top-code) and by a value that represents the lower limit (bottom-code). The idea behind of this technique is that values overcoming a threshold are considered unusual and, therefore, are more prone to be easily associated with specific individuals. For instance, consider a de-identified dataset that stores information about active staff; because aged people carrying on their profession are uncommon and, thus, more easily re-identifiable by an attacker, the top-code could be set to 60 whereby any age greater than this value will be replace by “>60”. By definition, this method can be used on attributes that can be ranked, that is, numerical or categorical ordinal.

- *Local suppression*: This method is based on removing certain values of an attribute that are likely to contribute significantly to the disclosure risk of the involved records. Particularly, because uncommon combinations of

quasi-identifier values (outlier records) in the dataset may lead to re-identification, certain values in those combinations are replaced by the value *missing* to subtract their uniqueness. Suppression can be performed at the record level (entire records are suppressed), or on particular attributes in some records. Obviously, deleting values may severely hamper the accuracy of the analysis. Because continuous attribute values tend to be unique for each individual and it does not make sense to systematically suppress the values of these attributes, local suppression is rather oriented to categorical variables. In this respect, [31] proposes ways to combine local suppression and generalization.

2.3.2 Perturbative masking methods

Perturbative masking methods are based on distorting the original data to produce a protected dataset. Unlike non-perturbative methods, the perturbative ones may yield non-truthful data for individual records. We depict the main perturbative techniques below.

2.3.2.1 Noise addition

Noise addition is a family of methods that consist of adding to the input data a random noise sequence, typically drawn from a probability distribution. The main approaches to noise addition are *uncorrelated noise*, for individual attributes, and *correlated noise*, for multivariate datasets, both of them formulated for continuous data [5].

Particularly, uncorrelated noise addition [32] is based on adding sequences of normally distributed random noise to individual attributes from an input dataset. In order to distort a attribute X^a with n records, each value x_i^a is replaced by a noisy version x_i^{a*}:

$$X^{a*} = X^a + \varepsilon^a, \tag{2.1}$$

where $\varepsilon^a = \{\epsilon_1^a, \ldots, \epsilon_n^a\} \sim \mathrm{N}(0, \sigma_{\varepsilon^a}^2)$ is a noise sequence randomly drawn from a normal distribution with mean zero and variance $\sigma_{\varepsilon^a}^2$. The error variance $\sigma_{\varepsilon^a}^2$ is proportional to the original attribute variance $\sigma_{X^a}^2$ as follows:

$$\sigma_{\varepsilon^a}^2 = \alpha \sigma_{X^a}^2, \quad \alpha > 0 \tag{2.2}$$

The parameter α determines the amount of noise to be added, whose value usually ranges between 0.1 and 0.5 [33]. The higher the α, the higher the distortion level. Note that if $\alpha > 0.5$, more than 50% of the variation in the distorted data is caused by the added noise and, therefore, the data values tend to become marginal.

The result of uncorrelated noise addition is a distorted attribute that preserves the mean of the input attribute and keeps the variance proportional in a factor $1+\alpha$:

$$\begin{aligned} \mu_{X^{a*}} &= \mu_{X^a} + \mu_{\varepsilon^a} = \mu_{X^a} \\ \sigma_{X^{a*}}^2 &= \sigma_{X^a}^2 + \sigma_{\varepsilon^a}^2 = (1+\alpha)\sigma_{X^a}^2 \end{aligned} \tag{2.3}$$

In order to distort multiple attributes, given the uncorrelated character of the method, the noise must be applied to each attribute independently [33, 34], without considering the noise applied to previous attributes. Consequently,

$$Cov(\varepsilon^a, \varepsilon^b) = 0, \quad \forall a \neq b \tag{2.4}$$

Because the covariance between any two noise vectors ε^a (added to an attribute X^a) and ε^b (added to an attribute X^b) is null, the correlation between noise-added attributes is not preserved. Thus, the method is suitable for statistical analyses over attributes but not over records with non-independent attributes.

In order to solve this limitation, Kim [35] proposes a method to add correlated random noise to several attributes in a dataset X with m attributes, such that:

$$X^* = X + \varepsilon, \tag{2.5}$$

where X^*, X and ε are ($n\times m$) matrices and $\varepsilon \sim N(0, \Sigma_\varepsilon)$ follows a multivariate normal distribution (MVN) with mean the m-dimensional vector 0 and covariance matrix the ($m\times m$) matrix Σ_ε,

$$\Sigma_\varepsilon = \alpha\Sigma_X, \quad \alpha > 0, \tag{2.6}$$

where Σ_X is the covariance matrix of X, a symmetric matrix whose diagonal elements are the variances of individual attributes and the off-diagonal elements are the covariances between attribute pairs.

In consequence, the method preserves the mean of each attribute, keeps the covariance matrix of the distorted data proportional to the covariance matrix of the original data in a factor $1+\alpha$ and maintains the Pearson correlation coefficient ρ between the attributes,

$$\begin{aligned} \mu_{X^*} &= \mu_X + \mu_\varepsilon = \mu_X \\ \Sigma_{X^*} &= \Sigma_X + \Sigma_\varepsilon = (1+\alpha)\Sigma_X \\ \rho_{X^{a*},X^{b*}} &= \frac{Cov(X^{a*},X^{b*})}{\sqrt{\sigma^2_{X^{a*}}\sigma^2_{X^{b*}}}} = \frac{(1+\alpha)Cov(X^a,X^b)}{(1+\alpha)\sqrt{\sigma^2_{X^a}\sigma^2_{X^b}}} = \rho_{X^a,X^b} \end{aligned} \tag{2.7}$$

2.3.2.2 Swapping

Data swapping relies on exchanging the values within of each attribute in such a way that low-order frequency counts are preserved. This technique was originally presented to protect datasets with categorical attributes [36] and was extended to continuous data in [37].

Rank swapping, a variant of data swapping proposed by Greenberg in an unpublished manuscript [38] and described by Moore in [39], improves the analytical utility of the protected dataset by limiting the scope of the swaps and maintaining each permuted value within a certain rank-distance from the original one. Firstly, the method sorts in ascending order the records of the dataset by the values of the first attribute to be protected. Then, each value of the attribute is swapped with another one randomly chosen within the interval formed by the successive k records in the ranking, such that $k=p.n/100$, being p a user-defined percent of the total number of records. In this way, the rank of

two swapped values cannot differ by more than p% of the total number of records. Large values of p lead to greater permutations in the data whereas the smaller values of p induce a higher disclosure risk. The method is independently applied on each original attribute using, as input, the permuted dataset obtained in the previous iteration. Rank swapping was initially defined for ordinal categorical data and subsequently applied to continuous data [40].

In [41] the authors propose two new versions of rank swapping where each value can be swapped by any other value of the attribute, such that closer values have a higher probability to be selected than distant ones. This approach tries to avoid the interval disclosure, i.e., an attacker trying to link records will not be able to delimit the swapping interval with full confidence. The first version, named rank swapping p-distribution, defines the swapping interval using a normal probability distribution defined by $\mu = \sigma = 0.5\ p$. To permute an attribute X^a, each ranked value x_i^a is swapped by the ranked value x_{i+r}^a, r being a random number generated from N(0.5p, 0.5p); thus, distant values have lower probability to be selected than closer values. In the second version, named rank swapping p-buckets, the ranked values of the attribute are split into p buckets. Then, each ranked value x_i^a is swapped by a randomly selected value from another bucket which has been chosen according to a probability distribution. Obviously, buckets closer to the original value are selected with higher probability than the distant buckets. Both versions provide a lower disclosure risk than the standard method, but a larger information loss.

2.3.2.3 Microaggregation

This technique reduces the variability of the attributes by replacing the original values by small aggregates or microaggregates. The masking process is carried out in two steps: first, the dataset is partitioned into sets of at least k records and, then, the original values in each set are replaced by the group representative value, typically the average value or centroid. The higher the within-group homogeneity, the lower the within-group variance, and thus, the more representative the group average value will be. To maximize within-group homogeneity, groups are formed using a criterion of maximum similarity.

Depending on whether microaggregation deals with one or several attributes at a time, this technique is classified into univariate and multivariate. The univariate method, known as individual ranking [42, 43], microaggregates each attribute independently. Firstly, the records in the dataset are sorted by the first attribute to be protected, secondly, groups of k contiguous records are formed and, finally, the attribute values within each group are replaced by the group representative value (e.g., average value). The same procedure is repeated for the rest of the attributes by using as input dataset the obtained in the previous iteration.

Individual ranking causes low information loss [40], but high disclosure risk [44]. On the other hand, multivariate methods rank the dataset through multi-dimensional sorting [42] using either the first principal component of the dataset (that is, the most highly correlated standardized attribute with most original attributes of the dataset) or the sum of z-scores (that is, the sum of the standardized attribute values in each record). After sorting the records, univariate microaggregation is applied on each attribute. An alternative that better preserves data utility is to use heuristic methods to build the set of records, such as done by the MDAV (Maximum Distance to Average Vector) algorithm [45, 46]. The method computes the average record of the dataset through the squared Euclidean distance to firstly obtain the most distant record x_r from the average record and, then, obtain the most distant record x_s from x_r. After that, a cluster is built for x_r with the k-1 nearest records, similarly for x_s. Subsequently, the two clusters are microaggregated and labeled as microaggregated. The process is iteratively repeated with the non-microaggregated records until all records in the dataset are microaggregated. In any case, multivariate microaggregation leads to higher information loss than individual ranking [40]. Microaggregation was initially defined for numerical and subsequently extended to ordinal categorical data in [47]. Specifically, the median is used to aggregate ordinal data.

2.3.2.4 Data shuffling

Data shuffling [48] is a method to protect continuous confidential attributes that combines additive noise perturbation and data swapping. The method assumes that a dataset X consists of a set C of k confidential attributes and a set

S of m-k non-confidential attributes. The masking process yields as output a dataset $X^*=\{C^*, S\}$, where only the confidential attributes have been perturbed.

In a nutshell, the masking process works as follows. Firstly, the method builds the ($m\times m$) rank-order correlation matrix R of the original dataset X. This correlation matrix measures the association between the ranked confidential attributes and the ranked non-confidential attributes. After that, it computes the ($m\times m$) product moment correlation matrix ρ of X using R. The elements $\rho^{i,j}$ of ρ are computed as follows,

$$\rho^{i,j} = 2\sin\left(\frac{\pi r^{i,j}}{6}\right), \tag{2.8}$$

where $\rho^{i,j}$ and $r^{i,j}$ are the product moment correlation and the rank-order correlation between the attributes at the positions i and j.

Secondly, the method ranks the non-confidential attributes S. With these ranks, obtain the normalized values $s_i^{\text{attr}*}$ for each non-confidential attribute by using the following transformation:

$$s_i^{\text{attr}*} = \Phi^{-1}\left(\frac{(i)-0.5}{n}\right), \tag{2.9}$$

where (i) represents the rank order of the value of the non-confidential attribute S^{attr} in the record i and Φ^{-1} represents the inverse of the standard normal distribution.

Thirdly, the method generates k perturbed variables Y^* from a multivariate normal distribution (MVN) with mean the vector $\rho_{XS}(\rho_{SS})^{-1}(S^*)^{\mathrm{T}}$ and covariance matrix $\rho_{XX} - \rho_{XS}(\rho_{SS})^{-1}\rho_{SX}$.

Once the perturbed variables Y^* have been generated, data shuffling performs a reverse mapping from the perturbed values Y^* to the original confidential values C. The reverse mapping (shuffling) is based on replacing the rank ordered perturbed values of Y^* with the rank ordered original values C. The shuffled values, designed by C^*, are the same as the original confidential values C, but reordered.

If all attributes in the dataset are confidential, an independent, multivariate normal dataset with correlation matrix ρ is generated and reverse mapping is performed on this dataset.

Because data shuffling is a patented method and, thus, cannot be implemented without an agreement from the authors, it has been scarcely evaluated in the literature.

2.4 Data masking methods w.r.t. utility preservation

Disclosure risk limitation entails some modifications of the original data, which causes information loss and decreases the utility of the protected data. In any case, the actual data utility depends on the posterior data uses: a protected dataset may be useful for some kinds of analysis, but not for others. However, because potential data uses are very diverse and it may be hard even to identify them all at the moment of data release [5], measuring data utility becomes a tough task. See [5] for a thorough review of utility metrics used for microdata releases.

To minimize the information loss is desirable to maximize the preservation of the analytic structure of the dataset, which is determined by statistical measures, both univariate (e.g., mean and variance) and multivariate (e.g., correlation), as well as by other data features of great analytical interest (e.g., outlying values or granularity of the attributes). Information loss can be measured by observing the differences between the original and masked data [5], and by considering that there is a small loss if the analytic structure of the masked dataset is very similar to the structure of the original dataset. Below, we discuss strengths and weaknesses of the main masking methods w.r.t. information loss.

Non-perturbative methods are characterized by preserving data truthfulness, which results in masked records totally consistent with the contents of the original records, but with less detail. However, these methods usually lead to a significant loss of information, either by suppressing some of the data or by reducing the level of detail. By omitting data, as done by sampling and local suppression methods, most features of the dataset are severely altered. For example, because outlier records are particularly easy to re-identify, if they are present in the input data, these are systematically eliminated by such methods. These values that are atypically distant from the rest of the data are particularly useful for researchers because they could

identify areas calling for further investigation (e.g., rare diseases). Similar consequences result from the methods that reduce the detail of the data, as do by the generalization method. Because the input values can only be replaced by a reduced set of generalizations, this entails appreciable granularity penalties. This fact is more noticeable as one moves up in the generalization hierarchy: coarse generalizations could hide certain patterns in the data analysis and, thus, result in a high loss of information. By omitting data and reducing information detail, arithmetical measures are also altered.

Unlike non-perturbative methods, perturbative ones distort the original data and lead to the publication of non-truthful data. As a consequence, it is possible that the masked dataset may contain nonsensical value combinations for individual records. For example, if the attributes of the dataset are (gender, disease) or (occupation, income), after the masking process, it may be the case that a masked record has the values (*female*, *prostate cancer*) or (*clerk*, *income of a senior executive*), which may provide wrong conclusions. To minimize the number of nonsensical combinations and, thus, obtain a masked dataset with coherent values, it is desirable that the perturbative mechanisms be parametrizable. In this way, a data collector will be able to adjust the distortion level in the masking process and graduate the dissimilarity between the original and masked values, thereby obtaining more accuracy datasets.

Despite altering data truthfulness, perturbative methods provide masked datasets of higher analytical utility than non-perturbative ones. In the case of rank swapping and data shuffling, the univariate features of the data, such as the mean, the variance, the frequency distribution, outlying values and the granularity of the attribute samples are perfectly preserved because the values in the protected attribute are exactly the same as those in the original attribute but permuted. Because microaggregation replaces the original values in each set of records by aggregates, when these microaggregates are averages, the attribute means are preserved. However, by making data more homogenous, the variability and granularity of the microaggregated dataset are reduced. Noise addition, on the other hand, is capable of preserving the mean of the original attributes and keeping the variance proportional to the level of added noise, but might alter the granularity and the outliers. Finally, a study carried out in [49] shows that microaggregation preserves better the outliers than noise addition.

Regarding the dependence among the attributes of the input dataset, as evidenced in a study that evaluates the utility of the outcomes of several

perturbative methods [50], correlated noise addition is the only method that fully preserves the correlation structure of the original data, multivariate microaggregation yields good outcomes and rank swapping alters significantly the correlation structure of the data, thus greatly distorting regression inferences. Concerning data shuffling, the authors show in [51] that the rank-order correlation obtained from the masked data is likely to be very close to that of the original data.

Table 2.1 Comparative of masking methods w.r.t. data types and preserved analytical features. X: applicable; =: fully preserved feature; ++: highly preserved feature; +: averagely preserved feature

		Non-Perturbative methods			**Perturbative methods**			
		Sampl.	Generaliz.	Suppress.	Noise addition	Rank swap.	Micro aggregation	Data shuffling
Data types	Continuous numerical		X		X	X	X	X
	Discrete numerical	X	X	X		X	X	X
	Ordinal categorical	X	X	X		X	X	X
	Nominal categorical	X	X	X				
Parametrizable perturbation					X	X	X	
Preserved analytical features	Truthfulness	=	=	=				
	Data nature	=		=	=	=	=	=
	Frequency dist.					=		=
	Outliers				+	=	+	=
	Granularity	+			+	=	+	=
	Cardinality		=		=	=	=	=
	Mean	+			=	=	=	=
	Variance	+			++	=		=
	Correlation	+			=		+	++

As summary and guide for practitioners and researchers on protected microdata releases, Table 2.1 shows the data types on which each masking method can operate and their impact on the structure of the original dataset. Note that, because the generalization method discretizes input continuous values to numerical ranges, the nature of the data changes from continuous to discrete. In contrast with non-perturbative methods, the perturbative ones maintain the continuous nature of numerical values. As we can see, most

perturbative masking methods have been designed to deal with numerical data and, in some cases, with ordinal data [5]. In this respect, because nominal values are finite, discrete, textual and non-ordinal, the arithmetic and sorting operations required in these methods do not make sense for nominal data; thus, perturbative masking methods in their standard form may seem, a priori, not applicable for such type of data.

2.5 Privacy models

Whereas data protection methods offer a posteriori privacy guarantees, privacy models establish beforehand conditions that the protected data must satisfy to guarantee a minimum level of anonymity for the respondents. These privacy guarantees can be attained for a particular dataset using one or several of the anonymization methods detailed in Section 2.3, as summarized in Table 2.2.

Below, we depict the main privacy models proposed in the literature and when these models can be combined each other to achieve more robust protection. In addition, we detail which protection methods can be used to enforce a specific privacy model, thereby offering *ex ante* privacy guarantees.

2.5.1 *k*-anonymity

k-Anonymity is a privacy model for microdata releases focused on preventing the re-identification of the individuals to whom the data refer.

Let X be a microdata set consisting of quasi-identifier attributes and confidential attributes. To prevent re-identification, the idea underlying in *k*-anonymity is to make the combination of quasi-identifier attributes non-unique by making them indistinguishable within groups of records. For that, *k*-anonymity [31] requires each combination of values of the quasi-identifier attributes in the released dataset X^* to be shared by k or more records. The set of records in X^* sharing the values for all the quasi-identifier attributes is named *equivalence class*. In this way, *k*-anonymity guarantees that, for any combination of quasi-identifier values in the released dataset X^*, there are at least k-1 records sharing the same combination. Therefore, an attacker with access to an external non-anonymous identifying dataset that contains the quasi-identifier attributes from the released dataset X^* will not be able to link a

specific individual to a specific record in X^*. In this scenario, the attacker will be able to, at most, identify the set of k records in X^* that contains the target individual. Therefore, the probability of performing the right re-identification is not greater than $1/k$.

Two main approaches can be used to generate k-anonymous datasets. The first one is based on combining two non-perturbative masking methods: generalization and local suppression [4, 31]. On the one hand, by generalizing attribute values within record groups to a common value or tuple, they are made indistinguishable. On the other hand, suppression contributes to reduce the amount of generalization required to generate the k-anonymous dataset by removing outlier records. This approach has the disadvantage of requiring a high computational cost to find an optimal recoding that minimizes the information loss [52]. A second more practical approach is based on microaggregation [46]: by construction, when microaggregation is directly applied to the quasi-identifier attributes a dataset with a minimum cluster size of k, the outcome satisfies k-anonymity.

A k-anonymous dataset X^* is protected against *identity disclosure* because its quasi-identifier attributes have been homogenized to limit the probability of record re-identification to $1/k$. However, if the confidential attributes of X^* have null or low variability within an equivalence class, an attacker can determine the exact or approximated value that a confidential attribute takes for those individuals. Specifically, if all the individuals within an equivalence class share the same value in a confidential attribute and the attacker can establish that target individual' record is within this group, then the attacker learns the confidential attribute value, even if re-identification didn't happened. Therefore, k-anonymity cannot guarantee protection against *attribute disclosure*.

Several extensions of k-anonymity have been proposed in the literature to mitigate the risk of *attribute disclosure*. On the one hand, l-diversity [53] requires the presence of at least l different well-represented values in the confidential attribute for each equivalence class. On the other hand, t-closeness [54] requires the distribution of the confidential attribute in each equivalence class to be similar to the distribution in the overall data set. To provide protection against *attribute disclosure*, and thus, offer a stronger privacy guarantee, k-anonymity must be combined with l-diversity or t-closeness. Again, generalization and suppression have been used to enforce both, specifically, by considering the additional constraints imposed by l-diversity or

t-closeness on the confidential attributes to select the groups of records to generalize [54, 55]. Very recently, microaggregation have been also adapted to enforce *t*-closeness on top of a *k*-anonymous microaggregated dataset [56].

2.5.2 Probabilistic *k*-anonymity

Probabilistic *k*-anonymity [26] is a privacy model for microdata release that offers the same protection guarantees as *k*-anonymity. Like *k*-anonymity, it limits the probability of re-identification at most 1/*k*. However, probabilistic *k*-anonymity does not require that combinations of quasi-identifier values in the released dataset X^* to be indistinguishable within groups of *k* records. By relaxing the indistinguishability requirement, it is reasonable to expect that probabilistic *k*-anonymity can be satisfied with less information loss than *k*-anonymity.

Because probabilistic *k*-anonymity is a relaxation of *k*-anonymity, if X^* satisfies *k*-anonymity, then it satisfies probabilistic *k*-anonymity. On the contrary, probabilistic *k*-anonymity does not imply *k*-anonymity.

Any approach that allows attaining the required limit in the probability of record re-identification, can be used to obtain probabilistic *k*-anonymous datasets. Obviously, generalization, suppression and microaggregation naturally support probabilistic *k*-anonymity but, because they reduce the granularity of the data to make quasi-identifier values indistinguishable, they incur in an unnecessary information loss. Additionally to *k*-anonymity, rank swapping can be used to enforce probabilistic *k*-anonymity by setting the parameter *k* in the rank swapping method; in this way, an attacker with access to the permuted attribute, whose values have been swapped in intervals encompassing at *k* records, would only be able to infer the original values with a probability at most 1/*k*.

2.5.3 ε-Differential privacy

Whereas the above privacy models are aimed at microdata releases, ε-differential privacy was proposed as a privacy guarantee for queryable databases [25], where queries (typically count queries) are submitted to a database containing the original individual records (microdata). In this query-

answer interactive environment, differential privacy states the conditions that the answers must satisfy so that disclosure risk is under control. The anonymization mechanism to attain differential privacy is called a differentially private sanitizer and sits between the user submitting queries and the database answering them.

The principle underlying differential privacy is that the presence or absence of any single individual in the database should be undetectable when analyzing the outcomes of the queries. For that, the sanitizer must limit the contribution of any single individual on the response to a query. Because differential privacy assumes that each record in the dataset refers to a different individual, comparing the outcome of a query before and after an individual has contributed her data to the dataset is equivalent to comparing the outcome of that query between datasets that differ in at most one record (neighbor datasets). Formally,

Definition 1. Let ε be a positive real number and κ a randomized function that takes a dataset X as input. An output $\kappa(X)$ to a query f is differentially private if, for all neighbor datasets X and X', and all subsets S of the domain of κ, $\kappa(X)$ holds

$$\mathrm{P}(\kappa(X) \in S) \leq e^{\varepsilon} \times \mathrm{P}(\kappa(X') \in S) \tag{2.10}$$

ε-differential privacy guarantees that the knowledge gain that can be extracted from the response to a query f is limited by a factor of e^{ε}.

The most common differentially private sanitization mechanism is noise addition. After the database computes the answer $f(X)$ to a certain user query f, the sanitizer adds random noise to $f(X)$ to mask the answer. The noisy answer $\kappa(X) = f(X) + noise$ is returned to user that submitted the query f. Different distributions can be used to generate noise according to the data type provided by $f(X)$: the Laplace distribution is used when $f(X)$ is a real value [57], and the discrete Laplace distribution [58] or the symmetric geometric distribution [59] are used when $f(X)$ is an discrete numeric value.

Among noise distributions, the Laplace distribution Lap(0, $\Delta(f)/\varepsilon$) with mean 0 and scale parameter $\Delta(f)/\varepsilon$ is the most used, where $\Delta(f)$ represents the sensitivity of f, i.e., the maximum variation in the query function between neighbor datasets, and ε represents the differential privacy parameter. The scale parameter is used to calibrate the noise such that, fixed ε, the higher the

sensitivity $\Delta(f)$, the more Laplace noise is added to mask the effect of any single individual record in the response of the query. In other words, to satisfy equation (2.10), more noise is required when the function f varies strongly between neighbor datasets. On the other hand, when ε is very small, because equation (2.10) requires that the probabilities on both sides be almost equal, $\kappa(X)$ must yield very similar values for all pairs of neighbor datasets, which is achieved by adding a lot of noise. Thus, fixed $\Delta(f)$, the smaller ε, the more noise is needed to be added.

Despite differential privacy is designed for queryable databases, different approaches have been proposed for microdata release [24, 60, 61]. In most cases, a differentially private dataset is generated from noisy histogram queries. A histogram on an attribute is constructed by partitioning its domain into mutually disjoint subsets and obtaining the frequencies in each subset. Then, to prevent the counts from disclosing information on the data and, thus, to fulfill ε-differential privacy, discrete noise is applied on the counts. In these cases, any number of analysis (queries) can be performed, but utility guarantees are only offered for a restricted class of queries (counting queries). Very recently [62, 63], an alternative approach for differentially private microdata releases have been presented. Instead of releasing noisy counts, the authors add Laplacian noise to the actual attribute values. To decrease the noise (and, thus, improve data utility), the dataset is prior microaggregated by means of univariate [62] or multivariate microaggregation [63], so that original values are replaced by averages, which have a lower sensitivity. The advantage of this method is that it does not restrict the type and number of analyses on the protected dataset.

Table 2.2 Privacy-preserving methods used to satisfy privacy models

Privacy model	Privacy-preserving method
k-anonymity	Generalization + local suppression Microaggregation
k-anonymity + *t*-closeness	Generalization + local suppression Microaggregation
k-anonymity + *l*-diversity	Generalization + local suppression
Probabilistic *k*-Anonymity	Rank swapping
ε-differential privacy	Noise addition Microaggregation + Noise addition

2.6 Conclusion

In this chapter, we have surveyed and compared privacy protection mechanisms that aim at balancing the tradeoff between the individual's privacy and the analytical utility of the protected data. Among them, the perturbative masking methods usually provide the most accurate outcomes, because they better preserve the statistical features of the data (recall Table 2.1).

However, because of their mathematical operating principle, perturbative mechanisms are seen as techniques intended for numerical data and, at most, for ordinal categorical data. To apply perturbative masking methods on nominal data and, thus, embrace their benefits regarding utility preservation, mechanisms alternative to the standard numerical algorithms are required. In the next chapter, we will discuss this issue and survey the different adaptations that have been proposed in the literature to apply perturbative masking to nominal data.

Chapter 3 State of the art on perturbative protection of nominal data

As discussed in Chapter 2, perturbative masking methods provide protected datasets that have, in general, higher analytical utility than those obtained with non-perturbative mechanisms. However, as their operating principle is based on mathematical calculations, perturbative methods, a priori, can only deal with continuous data and, in some cases, with ordinal categorical data. This contrasts with the fact that most of the personal data that are currently being gathered (e.g., from social networks, electronic healthcare records or web browsing logs) and that should be subject of anonymization are nominal. Therefore, it would be desirable to have available perturbative protection alternatives that can manage nominal data.

This constitutes a challenging goal due to the very nature of nominal data: because nominal data are finite, discrete, textual and non-ordinal, it is not possible to carry out the arithmetical data transformations required by perturbative masking. Moreover, because nominal data refer to concepts or instances (rather than numerical magnitudes), the utility of the protected nominal data is more closely related to the preservation of data semantics than of data distributions [13]; therefore, data transformations should carefully consider the *semantics* conveyed by nominal values.

Under this semantic perspective, in this chapter we survey the alternatives and adaptations proposed in the literature to protect nominal data in a perturbative way. Related works have been classified according to the perturbative principle they use to protect data: *data distortion* (which includes noise addition), *data permutation* (which includes data swapping) and *data aggregation* (which includes data microaggregation).

3.1 Related works on distortion-based methods

Nominal data have been scarcely considered in noise addition methods and, in all cases, data distortion mechanisms alternative to the standard numerical

noise have been proposed. One of the first techniques for distorting nominal data was introduced by Kooiman et al. [64]. This method, named Post Randomization Method (PRAM), changes the original values of an attribute according to a predefined probability distribution. The probability distribution is described by a Markov matrix whose entries are the probabilities associated with the transitions between each original value and any other value of the sample. However, it is generally difficult to find a suitable Markov matrix that performs changes with low loss of information [5].

From another perspective, given that nominal values lack a natural order, some authors [65] suggest breaking down the nominal attribute into ordinal sub-attributes to facilitate the operations during the distortion process. In this regard, an attribute such as *place of birth* could be turned into the numerical variables *geographical longitude* and *latitude*, but not all attributes admit an ordinal alternative. In [66], Islam and Brankovic present a noise addition framework with several probabilistic techniques to distort nominal attributes, in which the values of an attribute are replaced by other values of the same attribute according to a user defined probability.

In recent years, noise addition has also gained relevance in the context of data privacy thanks to the popularization of the ε-differential privacy model [25], whose enforcement usually relies on Laplacian noise. Under the umbrella of differential privacy, some mechanisms have been proposed to deal with discrete data, either discrete numbers or nominal values. On the one hand, the geometric mechanism [59] is capable of adding random noise from a symmetric geometric distribution to one discrete numeric value, such as an integer numeric answer to a query on a given dataset. On the other hand, McSherry and Talwar [67] propose the *exponential mechanism* to distort nominal attributes. This method probabilistically chooses the output of a discrete function according to the input dataset and a quality criterion based on a score function. The score tells us how good a noisy output is for that dataset. Since the probability associated with an output increases exponentially with its score, the distribution is biased towards outputs with high scores, thus moving the expected outcomes closer to the optimum.

All the above methods rely on the distribution of the data rather than on the actual semantics of nominal values. This makes them more suitable for discrete numerical values, rather than nominal ones. In an effort to consider the *meaning* of the values, Giggins and Brankovic [68] proposed VICUS: a noise addition technique for nominal attributes that uses a similarity measure

to capture the notion of transitive similarity between the values of an attribute. Because VICUS does not exploit the semantics modeled in ontologies, all the values of the dataset and the relationships between them must be manually represented in a graph in order to be able to use the similarity measure. From a semantic perspective, Abril et al. [69] suggest using noise addition to protect individual textual documents while preserving the semantics of the document, even though they do not specify the calculation of such noise.

From the discussion above, we can see that most distortion techniques for nominal data available in the literature neglect or poorly consider the semantics of nominal values and/or deviate from the standard notion of noise addition. Moreover, another limitation to highlight is that the available methods manage individual attributes independently and, therefore, neglect the potential correlations between attribute pairs. This crucial issue may negatively affect data analysis, which usually exploits attribute correlations to perform inferences.

3.2 Related works on permutation-based methods

Various permutation-based methods have been proposed to protect datasets while preserving certain statistical features. However, nominal data have barely been considered by these techniques. The first one, named data swapping [70], is based on swapping the values of each attribute from a dataset of t categorical attributes to yield a permuted dataset whose t-order frequency counts, or t-order statistics, are the same as those of the original dataset, i.e., a t-order equivalent dataset. Since the t-order statistics are preserved, the inferences that derive from them are not altered. To do this, the authors introduce the notion of (t-1, X^l) equivalence classes, t being the number of attributes to distort in the dataset and X^l the attribute to distort in each step. Two records $x_i = (x_i^1,\ldots,x_i^l,\ldots,x_i^t)$ and $x_j = (x_j^1,\ldots,x_j^l,\ldots,x_j^t)$ from a dataset X with t attributes belong to the same (t-1, X^l) equivalence class if $x_i^1 = x_j^1,\ldots,x_i^{l-1} = x_j^{l-1}, x_i^{l+1} = x_j^{l+1},\ldots,x_i^t = x_j^t$, i.e., the records belonging to a same class only may differ in the value associated to the attribute X^l. To distort a given attribute X^l, the method builds all (t-1, X^l) equivalence classes

for that attribute and then randomly swaps the values of X^l within each equivalence class. This process is repeated for each one of the t attributes using as input the dataset obtained in the previous permutation stage. The authors show that any data swap that preserves t-order statistics will significantly reduce the risk of disclosure. However, because of the way in which data swapping operates, this method is not suitable when most equivalence classes in the dataset are composed of one or few records, since the swaps can hardly be carried out.

Concerning the form in which the data are released, in [38, 71] the authors present two alternatives: releasing directly the permuted dataset or releasing only the t-order statistics obtained from the dataset as contingency tables. In the first case, because data are released as microdata, it is necessary to add enough uncertainty on the true values of the individuals' data to reasonably protect their privacy. However, identifying a large number of swaps that preserves the t-order statistics is computationally impractical [72, 73]. In the second case, the existence of other t-orders equivalent to that of original dataset protects individuals' privacy, because attackers cannot know whether a certain t-order statistic comes from the original dataset or a permuted one. As a feasible approach for the release of microdata, Reiss proposes in [72] a variation of data swapping where the t-order frequency counts are approximately preserved. Firstly, the method computes the relevant frequency tables from the original dataset, and then constructs a new dataset consistent with these tables. To do this, the values of an attribute are randomly selected according to the probability distribution derived from the original frequency tables; because this may produce values not appearing in the original dataset, this makes it a synthetic method, rather than a strict data swapping one.

Because the above methods do not limit the swapping range, very different values may be swapped, thereby increasing the information loss. In order to limit the scope of the swaps and maintain each permuted value within a certain rank-distance from the original one, a new permutation approach named *rank swapping* was proposed. All the methods classified as rank swapping depicted in Section 2.3.2.2 have been designed to deal with numerical or ordinal categorical data. In both cases, total orders are available to build the value ranks in which the algorithms rely. However, nominal data lack a natural total order. For such data, rank swapping has been considered either non-applicable [5, 74] or it has been suboptimally applied by defining artificial total orders, e.g., topological order of categorical labels for nominal attributes [75]. In this

latter proposal, a cumulative function of the frequency of the nominal values in the attribute defined on the topological order is used to rank the attribute. Although with this alternative non-comparables nominal values can be sorted, the utility of the protected outcomes may be hampered because of their lack of semantic coherence.

Even though rank swapping is pointed out as one of the best performer data protection mechanisms in terms of disclosure risk minimization and data utility preservation [40], because swapping is applied independently for each attribute, it may alter significantly the correlation structure of the data [50]. An alternative permutation approach that tries to overcome this shortcoming is data shuffling (Section 2.3.2.4). However, data shuffling has not been applied to nominal data because, like rank swapping, requires sorting the values of the dataset.

From the discussion above, we can see that there are no permutation techniques for nominal data in the literature that consider the semantics of the values.

3.3 Related works on aggregation-based methods

As discussed in Section 2.3.2.3, data aggregation is based on building clusters of, at least, *k* indistinguishable records and replacing the original values of each record with the representative value of the cluster to which the record belongs, typically the central element of the cluster or centroid. In order to preserve the accuracy of the data after the aggregation process (i) the records within the cluster must be as homogeneous as possible (in this way, similar records are aggregated together, thereby minimizing the information loss resulting from the replacements), and (ii) the value used to aggregate the records in the cluster must be an accurate representative of the elements in the cluster.

Several approaches have been proposed to adapt aggregation-based methods to nominal data. On the one hand, [46, 47] propose definitions for aggregation and clustering, which are the two basic operations required in microaggregation. Specifically, for aggregating nominal data the centroid is computed by using the plurality rule or mode. The clustering operation is based on the *k-modes algorithm*. In short, this algorithm obtains the optimal cluster through an iterative process. Firstly, cluster centroids are initialized at

random, and the values of the attribute are assigned to the cluster of the nearest centroid. To determine the nearest centroid of a record, a distance measure is used, which is defined as the summation of the distance between individual values. The distance is defined as 1 when nominal values are different and 0 when equal. Finally, cluster centroids are revaluated based on their newly assigned records. As both operations are performed solely according data distribution and, thus, omit the semantics of the data, this alternative impairs the utility of nominal attributes.

On the other hand, truthfully semantic centroids that consist on the concept that taxonomically subsumes all the values in the cluster [76] are negatively affected by outliers: generalizing outlying values to a common subsumer commonly produces too abstract generalizations that result in a high loss of information. In [11, 13, 77], the authors present two microaggregation techniques that yield semantically-coherent outcomes by considering both the semantics and the distribution of the data during the calculation of the centroid and the construction of the clusters. The proposed solutions, which will be detailed in the next chapter, are based on capturing and managing the semantics underlying to nominal values by exploiting the formal knowledge modeled in ontologies.

In view of the above, microaggregation is the only perturbative masking method that has been adapted to work with nominal data from a semantic perspective.

3.4 Conclusion

The above discussion evinces that most perturbation techniques for nominal data available in the literature neglect or poorly consider the semantics of nominal values. The accurate management of nominal data is not straightforward because, on the contrary to numbers, they take values from a discrete and finite list of categories, which are usually expressed by words. Since neither arithmetic operators nor a total order relation can be applied to this kind of data, simplistic approaches use equality/inequality operators, distributional statistics (e.g., mode), probabilistic techniques and artificial total orders to compare, aggregate, distort or sort them. These approaches neglect the most important dimension of nominal data: the meanings of the words, i.e.,

their semantics. Consequently, protected nominal datasets may lack semantic coherence and, thus, their utility may be severely hampered.

Microaggregation is the only perturbative masking method that has been thoughtfully adapted to yield semantically-coherent outcomes by exploiting the formal knowledge modeled in ontologies. The significant improvement of utility obtained in [13] with the semantic version against those obtained with the standard microaggregation leads to consider the use of ontologies (and the semantic techniques that rely on them) as via to interpret nominal data during the masking process.

Specifically, by exploiting ontologies to capture and manage the semantics underlying to nominal data, we could adapt to the nominal domain other perturbative methods that offer additional advantages over microaggregation. In this regard, we focus our attention in two particularly interesting techniques: noise addition, because it is the only method that is able to fully preserve the dependence relation between the attributes of the dataset, and rank swapping, because it is the only non-patented method that is able to perfectly preserve the univariate analytical features of the dataset. In fact, rank swapping is considered one of the best perturbative mechanisms w.r.t. disclosure risk minimization and data utility preservation [40]. In addition, both methods can be used to satisfy privacy models and, therefore, offer *ex ante* privacy guarantees, as discussed in Section 2.5. Specifically, rank swapping yields probabilistic *k*-anonymous datasets. On the other hand, noise addition may be used as sanitization mechanism to attain differential privacy.

In addition to the above advantages, a recent study [17] has shown that any anonymization method is functionally equivalent to a permutation plus a small amount of noise; this turns the spotlight on the permutation-based and distortion-based data transformations encompassed by the swapping and noise addition mechanisms as the essential principle underlying any data anonymization.

Chapter 4 **Semantic Operators**

As discussed in Chapter 3, most perturbative masking methods available in the literature neglect or poorly consider the semantics of nominal data. Because the utility of such data is closely related to the preservation of their semantics, data transformations employed during the protection process would require from operators (the difference, the mean, the variance, the covariance and the sorting operator) that consider the meaning of words. For such purpose, in this chapter, we first introduce the notion of *semantic domain* for nominal data, which is based on an underlying ontology, and discuss existing *ontology-based semantic similarity measures* that can be used to compare nominal values (difference operator). Then, we introduce a semantically-coherent version of the *mean*, which conveys the meaning of a sample of nominal values and present, as our contribution, semantic versions of the remaining operators needed in data protection: the *semantic variance*, the *semantic covariance* and the *semantic sorting* operator.

4.1 Ontology

An ontology is a structured knowledge source that explicitly and consensually represents the concepts and the semantic interrelations of a domain of knowledge [14]. According to the formal definition proposed in [78], an ontology O is composed of a set of concepts or classes C, and a set of relation types R. The set of concepts represents the real-world entities of the area of knowledge being modeled. For example, in a medical ontology, the concepts can be types of diseases, medical procedures or clinical findings; i.e., single units of thought with a distinct clinical meaning. R represents types of semantic relations between concepts, such as taxonomic relationships, e.g., hyponymy and hypernymy (*is-a* links), and non-taxonomic relationships, e.g., meronymy and holonymy (*part-of* links).

Taxonomic relationships define a semi-upper lattice $\leq_C$ on C with top element $root_C$. In the concept hierarchy $\leq_C$, a concept c_j is a specialization or a

subsumed concept of another concept c_i, i.e., $c_j \leq_C c_i$, if and only if every instance of c_j is also an instance of c_i, c_i being a generalization or subsumer of c_j. $c_j =_C c_i$ means that c_i and c_j are the same concept. In $\leq_C$ the concepts are linked by means of transitive taxonomic relationships, which implies that if $c_k \leq_C c_j$ and $c_j \leq_C c_i$, then $c_k \leq_C c_i$. Consequently, the more general the concept is, the upper its position in the hierarchy $\leq_C$ will be.

The most common relation types between concepts modeled in an ontology *O* include:

- Hyponym: a concept c_j is a hyponym of a concept c_i if c_j is a kind of c_i, e.g., *flu* is a hyponym of *disease*. This relation is often termed *is-a* relationship.

- Hypernym: a concept c_i is a hypernym of a concept c_j if c_j is a kind of c_i, e.g., *disease* is a hypernym of *flu*.

- Meronym: a concept c_j is a meronym of a concept c_i if c_j is a part or a member of c_i, e.g., *hand* is a meronym of *body*. This relation is often termed *part-of* relationship.

- Holonym: a concept c_i is a holonym of a concept c_j if c_j is a part or a member of c_i, e.g., *body* is a holonym of *hand*.

- Co-hyponyms: two concepts c_j and c_k are co-hyponyms of a concept c_i if (i) c_j and c_k are hyponyms of c_i (both c_j and c_k share the same hypernym c_i) and (ii) c_j and c_k are not hyponym each other, e.g., *influenza caused by influenza A virus* and *influenza caused by influenza B virus* are co-hyponyms of *flu*. Co-hyponyms are also known as coordinate terms.

4.1.1 Types of ontologies

Guarino proposes in [14] the following classification of ontologies according to their level of dependence on a particular task or point of view:

- *Top-level ontologies*: describe general concepts like entity, which are independent of a particular problem or domain, such as WordNet [79] or Cyc [80] that try to model knowledge of the world.

- *Domain-ontologies*: describe the vocabulary related to a generic domain by specializing the concepts introduced in the top-level ontology, such as SNOMED-CT [81] that models biomedicine knowledge.

- *Task ontologies*: describe the vocabulary related to a generic task or activity by especializing 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 [82, 83].

4.2 Semantic domain

As stated above, unlike numerical data, nominal data are finite, discrete and non-ordinal. Nominal domains can be expressed either as unstructured term lists or as taxonomically structured set of concepts modeled in knowledge bases such as ontologies. The former case neglects data semantics and, as a consequence, the meaning of the output data can be much more distorted than what is expected and produce outcomes with a significant loss of information. The latter case is more desirable for handling nominal data because the meaning of data is considered thanks to the formal semantics provided by the ontology.

In this work, we propose to use ontologies to capture the underlying semantics of nominal data during the masking process. In order to ensure the generality of the methods, we only consider taxonomic relations because they are available in any ontology and constitute the backbone of the knowledge structure that ontologies provide [84]. In this context, the ontology is seen as a taxonomic tree or graph in which concepts (nodes) are interrelated by means of *is-a* links (edges).

We assume that the nominal values of an attribute X^a of the original dataset have been unequivocally associated with concepts c modeled in an ontology O. This process, named *conceptual mapping*, can be carried out manually or by lexically matching the strings of nominal values and concept labels, as done in [10].

Below, we define the concept hierarchy (or taxonomy) associated with the nominal values of an attribute that have been mapped in an ontology *O*.

Definition 2. Let $S(X^a)$ be the set of subsumers of an attribute X^a mapped in an ontology *O*. The *least common subsumer* of X^a, denoted by $LCS(X^a)$, is the most specific concept in $S(X^a)$.

$$\begin{aligned} S(X^a) &= \left\{ c_i \in O \mid \forall c_j \in X^a : \; c_j \leq_C c_i \right\} \\ LCS(X^a) &= \left\{ c \in S(X^a) \mid \forall c_i \in S(X^a) : \; c \leq_C c_i \right\} \end{aligned} \tag{4.1}$$

Definition 3. The taxonomy associated with an nominal attribute X^a mapped in an ontology *O*, denoted by $\tau(X^a)$, is the concept hierarchy extracted from *O* that includes all concepts that are taxonomic specializations of $LCS(X^a)$, including itself.

$$\tau(X^a) = \left\{ c_i \in O \mid c_i \leq_C LCS(X^a) \right\} \tag{4.2}$$

Note that $LCS(X^a)$ is also the *root* concept of $\tau(X^a)$.

4.3 Semantic difference

Privacy preserving methods also require from a distance measure to detect which individuals are most similar in order to group, swap, etc. their values and minimize the loss of information resulting from the subsequent data transformation. However, when managing nominal data, the standard arithmetical operator used to measure distances does not make sense. To support the difference operator in the nominal domain, and because nominal values should be managed according to the semantics of the concepts to which they refer, we propose to use the notion of semantic distance. Due to its core importance and the need of dealing with textual inputs, semantic distance has been applied in recent years in a variety of tasks, which include natural language processing, information management and retrieval, textual data analysis and classification or privacy-protection.

Semantic distance, *sd*: $c_1 \times c_2 \rightarrow \Re$, is a function mapping a pair of concepts to a real number that quantifies the differences between the meanings of two concepts according to the semantic evidence gathered from one or several knowledge sources [16]. In general, most measures were designed to assess the *semantic similarity*. Thus, unlike semantic distance, semantic similarity, *sim*: $c_1 \times c_2 \rightarrow \Re$, quantifies the alikeness between the meanings of two concepts.

It is important to note that two different concepts, which are often confounded, can be found in the literature. On one hand, *semantic similarity* states how taxonomically near two terms are, because they share some aspects of their meaning (*e.g.*, *dogs* and *cats* are similar because they are mammals; and *bronchitis* and *flu* are similar because both are disorders of the respiratory system). On the other hand, the more general concept of *semantic relatedness* does not necessary relies on a taxonomic relation (*e.g.*, *car* and *wheel* or *pencil* and *paper*);

A plethora of measures are currently available in the literature. According to the knowledge sources exploited during the semantic assessment, they can be classified into distributional and ontology-based measures. The former measure semantic relatedness according to the relative co-occurrence in corpora of the textual terms used to refer to the concepts to be compared. Ontology-based measures, on the other hand, rely on the structured semantic relationships that link concepts modeled in an ontology. In our work, we focus on ontology-based measures because of the following reasons [16]:

(i) Since they are based on explicit semantic evidences (i.e., manually modeled semantic relationships) they usually provide more accurate assessments than distributional measures.
(ii) They are much more efficient to calculate than distributional measures, which require analyzing a large amount of textual resources to create co-occurrence matrixes.
(iii) By mapping nominal data to concepts in the ontological domain (as detailed in the previous section), we avoid the language ambiguity that may affect distributional approaches (e.g. the concepts to be compared may be refereed in corpora with different textual terms and a textual term may have several meanings).

Different ontology-based measures have been proposed in the literature [16]. Among them, we can distinguish several different approaches according to the techniques employed and the knowledge exploited to perform the assessment. Ontology-based measures can be classified into *Edge-counting measures*, *Feature-based measures* and *measures based on Information Content* as depicted in the following sections.

4.3.1 Edge counting-based measures

Edge-counting measures evaluate the number of semantic links (typically *is-a* relationships) separating the two concepts in the ontology [85-87]. In general, edge-counting measures are able to provide reasonably accurate results when a detailed and taxonomically homogenous ontology is used [87]. They have a low computational cost compared to approaches relying on textual corpora and they are easily implementable and applicable. Among the classic edge counting-based measures, there are the following ones:

- Path Length [85] is the simplest method to estimate the semantic distance between two concepts c_1 and c_2. Its calculation relies on obtaining the length of the shortest taxonomic path connecting c_1 and c_2. The taxonomic path between c_1 and c_2, denoted by $pathLinks(c_1, c_2)$, is the number of *is-a* links or edges in the taxonomy that connects both concepts. Therefore, the longest the path, the more semantically distant the concepts will be.

$$sd_{pL}(c_1, c_2) = \min_{\forall i} \left(pathLinks(c_1, c_2) \right) \tag{4.3}$$

  This measure outputs non-normalized values.
- Leacock and Chodorow [86] propose a measure to evaluate the semantic similarity in a normalized way. Its calculation relies on dividing the length of the taxonomic path between two concepts c_1 and c_2 by the double of the maximum depth of the taxonomy in a non-linear fashion. In this case, the taxonomic path, denoted by $pathNodes(c_1, c_2)$, is the number of nodes the taxonomy that connects both concepts, included themselves.

$$sim_{LC}(c_1,c_2) = -\log\left(\frac{\min_{\forall i}\left(pathNodes_i(c_1,c_2)\right)}{2\cdot depth}\right), \quad (4.4)$$

where depth is the longest defined as

$$depth = \max_{\forall i}\left(pathNodes_i(c,root)\right) \quad (4.5)$$

- Wu and Palmer [87]. A problem of path-based measures is that they rely on the notion that all links in the taxonomy represent a uniform distance. Those measures omit the fact that equally distant concept pairs belonging to an upper level of the taxonomy should be considered less similar than those belonging to a lower level because they present different degrees of generality. In an attempt to address this shortcoming, Wu and Palmer present a measure that takes into account the depth of the concepts in the hierarchy.

$$sim_{WP}(c_1,c_2) = \frac{2\times depth(LCS(c_1,c_2))}{2\times depth(LCS(c_1,c_2)) + pathLinks(c_1,LCS(c_1,c_2)) + pathLinks(c_2,LCS(c_1,c_2))}, \quad (4.6)$$

where $LCS(c_1,c_2)$ is the most specific concept in taxonomy subsuming both c_1 and c_2; $depth(LCS(c_1,c_2))$ is the number of nodes in the longest taxonomic path between the node $LCS(c_1,c_2)$ and the node *root* of the taxonomy, including both $LCS(c_1,c_2)$ and *root*; $pathLinks(c_1,LCS(c_1,c_2))$ is the number of taxonomic links in the shortest path between c_1 and $LCS(c_1,c_2)$, similarly for $pathLinks(c_2,LCS(c_1,c_2))$. The use of the *depth* normalizes and weights the similarity of concept pairs. Specifically, equally distant concepts by *path* in an upper level of a taxonomy are considered less similar than those in a deeper level because concept specializations become less semantically distinct as they are recursively specialized [87].

4.3.2 Feature-based measures

Feature-based measures, based on the Tversky's model of similarity [88], estimate the similarity between concepts as a function of their common and non-common ontological features. The exploited ontological features are mainly taxonomic and non-taxonomic relationships, sets of synonyms (*synsets*) and glosses. Common features tend to increase similarity and non-common ones tend to diminish it. Since the additional knowledge helps to better differentiate concept pairs, they tend to be more accurate than edge-based measures [89]. Below, we depict three of the most representative feature-based measures:

- Tversky [88] proposes a similarity measure based on similarities between synsets.

$$sim_t(c_1, c_2) = \frac{|A \cap B|}{|A \cap B| \times \gamma(c_1, c_2) |A \setminus B| + (1 - \gamma(c_1, c_2)) |B \setminus A|}, \qquad (4.7)$$

  where A and B are the synsets for the concepts c_1 and c_2, respectively; $A \setminus B$ is the set of terms in A but not in B, i.e., non-common terms of A; $B \setminus A$ the set of terms in B but not in A, i.e., non-common terms of B; and $\gamma(c_1, c_2)$ is computed as a function of the depth of c_1 and c_2 in the taxonomy, such that

$$\gamma(c_1, c_2) = \begin{cases} \dfrac{depth(c_1)}{depth(c_1) + depth(c_2)}, & depth(c_1) \leq depth(c_2) \\ 1 - \dfrac{depth(c_1)}{depth(c_1) + depth(c_2)}, & depth(c_1) > depth(c_2) \end{cases} \qquad (4.8)$$

- Rodriguez and Egenhofer [90] consider that two concepts are similar if the synsets and glosses of their concepts and those of the concepts in their neighborhood are lexically similar. For that, they propose to evaluate the semantic similarity as the weighted sum of similarities between synsets, distinguishing features (i.e., meronyms, functions and attributes) and

semantic neighborhood of the evaluated concepts. A semantic neighborhood set of a concept c_i is the concept set whose distance to c_i is less than or equal to a non-negative integer named the radius of the semantic neighborhood.

$$sim_{RE}(c_1,c_2) = w \cdot S_{synsets}(c_1,c_2) + u \cdot S_{features}(c_1,c_2) + v \cdot S_{neighborhoods}(c_1,c_2), \quad w,u,v \geq 0 \tag{4.9}$$

where w, u and v are the parameters that weight the contribution of each component and S represents the overlapping between the different features.

Despite exploiting more semantic knowledge, these measures are not able to significantly outperform the accuracy of edge-counting measures, as evidenced by a study performed in [89]. The study attributes this fact to that some concept features, such as non-taxonomic relationships, are a form of knowledge partially modeled in most ontologies [84]. Thus, those measures limit their applicability to ontologies in which this information is available. To overcome this limitation, a new feature-based approach was proposed in [89]:

- Sánchez and coworkers [89] suggest to measure the semantic distance between two concepts by using only their taxonomic features. Specifically, this measure is defined as the ratio between the number of non-common taxonomic ancestors and the total number of taxonomic ancestors from the compared terms.

$$sd_{\log SC}(c_1,c_2) = \log_2\left(1 + \frac{\left|S(c_1) \cup S(c_2)\right| - \left|S(c_1) \cap S(c_2)\right|}{\left|S(c_1) \cup S(c_2)\right|}\right), \tag{4.10}$$

where $S(c_1)$ is the set of taxonomic subsumers of the concept c_1, 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.

4.3.3 Information content-based measures

Information content-based measures combine the taxonomic features of the evaluated concepts with their probability of occurrence in a given text corpus. Specifically, the occurrence frequency is used to estimate concept specificity, i.e., infrequent concepts are considered more specific because they convey more information and, therefore, semantics. These measures rely on quantifying the amount of information, i.e., Information Content (IC), that concepts have in common [91-93]. The IC of a concept states the amount of information provided by the concept when appearing in a context. On the one hand, the commonality between the concepts to compare is assessed from the taxonomic ancestors they have in common, which is referred as the least common subsumer, *LCS* (equation (4.1)). On the other hand, the informativeness of concepts is computed either extrinsically from the concept occurrences in a corpus [91-93] or intrinsically, according to the number of taxonomical descendants and/or ancestors modeled in the ontology [94-96]. In classical approaches [91-93] IC of a concept *c* is computed extrinsically as the inverse of the probability P(*c*) of occurrence of *c* in a given corpus, such that, infrequent concepts obtain a higher IC [91].

$$IC(c) = -\log(\mathrm{P}(c)) \tag{4.11}$$

Below, we detail three classical IC-based similarity measures:

- According to Resnik [91], semantic similarity depends on the amount of shared information between two terms, a dimension which is represented by *their LCS* in an ontology. The more specific the subsumer is (higher IC), the more similar the terms are, as they share *more information*. Similarity is computed as the IC of the LCS.

$$sim_{res}(c_1, c_2) = IC(LCS(c_1, c_2)) \tag{4.12}$$

  One of the problems of Resnik's proposal is that any pair of terms having the same LCS results in exactly the same semantic similarity. Both Lin [92] and Jiang and Conrath [93] extended Resnik's work by also considering the IC of each of the evaluated terms.

- Lin [92] proposes similarity measure that depends on the relation between the information content of the *LCS* of the evaluated concepts c_1 and c_2 and the sum of the information content of the individual concepts,

$$sim_{lin}(c_1, c_2) = \frac{2 \cdot IC(LCS(c_1, c_2))}{IC(c_1) + IC(c_2)} \tag{4.13}$$

- The measure proposed by Jiang and Conrath [93] is based on quantifying the length of the taxonomical links as the difference between the IC of a concept and its subsumer. When comparing term pairs, they compute their distance by subtracting the sum of the IC of each term alone from the IC of their LCS.

$$dis_{j\&c}(c_1, c_2) = (IC(c_1) + IC(c_2)) - 2 \times sim_{res}(c_1, c_2) \tag{4.14}$$

However, due to their dependence on corpora, these measures present some issues: accuracy depending on the size and adequacy of the corpus, high computational cost and language ambiguity problems. To overcome the disadvantages that present the corpus-based IC approaches, new methods propose to intrinsically compute the IC of a concept according to the number of hyponyms in the taxonomy [94, 95]. In comparison to corpora-based IC computation models, intrinsic IC computation models consider that abstract ontological concepts with many hyponyms are more likely to appear in a corpus because they can be implicitly referred in text by means of all their specializations. In consequence, concepts located at a higher level in the taxonomy with many hyponyms or leaves (i.e. specializations) under their taxonomic branches would have less IC than highly specialized concepts (with many hypernyms or subsumers) located on the leaves of the hierarchy. Three of the main methods to compute the intrinsic-IC of a concept are:

- Seco and coworkers [94] propose to compute the intrinsic-IC of a concept *c* as follows:

$$IC_{\text{seco}}(c) = 1 - \frac{\log(hypo(c) + 1)}{\log(max\_nodes)}, \tag{4.15}$$

where *hypo*(*c*) is the number of hyponyms of the concept *c* and *max_nodes* is the number of hyponyms of the *root* node of the taxonomy. The denominator ensures that IC values are normalized in the range [0..1]. This approach only considers hyponyms of a given concept in the taxonomy; so, concepts with the same number of hyponyms but different degrees of generality appear to be equally similar.

- Zhou [95]. In order to overcome the shortcoming of IC_{seco}, Zhou and coworkers proposed to complement hyponym-based IC computation with the relative depth of each concept in the taxonomy.

$$IC_{zhou}(c) = k\left(1 - \frac{\log(hypo(c)+1)}{\log(max\_nodes)}\right) + (1-k)\left(\frac{\log(depth(c)}{\log(max\_depth)}\right), \quad (4.16)$$

where *depth*(*c*) represent the depth of the concept *c* in the taxonomy and *max_depth* the maximum depth of the taxonomy.

- In [96], the *p*(*c*) is estimated as the ratio between the number of leaves in the taxonomical hierarchy under the concept *c* (as a measure of *c*'s generality) and the number of taxonomical subsumers above *c* including itself (as a measure of *c*'s concreteness). This ratio is normalized by the least informative concept (i.e. the root of the taxonomy), for which the number of leaves is the total amount of leaves in the taxonomy (*max_leaves*) and the number of subsumers including itself is 1. To produce values in the range [0..1] (i.e., in the same range as the original probability) and avoid *log*(0) values, 1 is added to the numerator and denominator.

$$IC(c) = -\log p(c) \cong -\log\left(\frac{\frac{|leaves(c)|}{|subsumers(c)|}+1}{max\_leaves+1}\right) \quad (4.17)$$

Intrinsic IC-based approaches overcome most of the problems observed for corpus-based IC approaches (specifically, the need of corpus processing and data-sparseness). Moreover, they achieve a similar, or even better accuracy than corpus-based IC calculation when applied over detailed and fine grained

ontologies [96]. However, for small or very specialized ontologies with a limited taxonomical depth and low branching factor, the resulting IC values could be too homogenous to enable a proper differentiation of concepts' meanings [96].

4.4 Semantic mean

An averaging operator, such as the mean, is usually employed to select a prototypical value from a sample, which acts as the central point of such sample. For the calculation of the mean, we use the notion of *centroid* of a sample of discrete values [97]. The centroid of a set of values is the least distant element of the domain to all the values in the sample. As discussed in Chapter 3, centroids are used in many clustering and data analysis algorithms to construct representative values of clusters or to find out the central or average value of a discrete dataset.

Among the different alternatives to calculate the centroid depicted in Chapter 3, we propose to use the technique in [97], which yields semantically-coherent outcomes by considering both the semantics and the distribution of the data during the calculation of the centroid. By applying this technique to the semantic domain defined in Section 4.2, and by using the semantic distance discussed in Section 4.3, we formulate the mean of a nominal attribute as follows:

Definition 4. The *semantic mean* of a nominal attribute X^a, denoted by $sMean(X^a)$, is the concept c from associated taxonomy $\tau(X^a)$ that minimizes the sum of the semantic distances with respect to all x_i^a in X^a.

$$sMean(X^a) = \underset{c \in \tau(X^a)}{\arg\min} \left(\sum_{x_i^a \in X^a} sd(c, x_i^a) \right) \tag{4.18}$$

With this definition, any concept in $\tau(X^a)$ can be the mean of the attribute, regardless whether it was present in X^a or not. In this manner, we expand the set of mean candidates to obtain a more accurate discretization. When more

than one candidate minimizes the distance, all of them are equally representative, and any of them can be selected as the mean of the attribute.

4.5 Semantic variance

The variance is used to measure the dispersion of the data in a sample. In data distortion mechanisms, the variance of a sample of values is relevant to configure the magnitude of the noise to be added to the input data. From a semantic perspective, [98] presents a measure that quantifies the semantic dispersion of an ontology, which we adapt here to measure the semantic variance of a sample.

By strictly following the mathematical notion of the arithmetic variance, the *semantic variance* of a nominal attribute should take into account the semantic differences between each value of the attribute and its *semantic mean*. Again, these semantic differences can be computed from the semantic distances between the values of the attribute and the mean, which we use to measure the semantic dispersion of a nominal attribute, as follows.

Definition 5. The *semantic variance* of a nominal attribute X^a, denoted by $sVar(X^a)$, is the average of squared semantic distances between each concept x_i^a in X^a and the semantic mean $sMean(X^a)$.

$$sVar(X^a) = \frac{\sum_{x_i^a \in X^a} \left(sd(x_i^a, sMean(X^a)) \right)^2}{n}, \tag{4.19}$$

where n is the number of values in X^a. Numerically, when the semantic variance is equal to 0, it indicates that all values are identical, whereas a high dispersion indicates that concepts are very spread out from the reference center (*sMean*) and from each other, thus being well differentiated.

4.6 Semantic distance covariance and correlation

The standard covariance and the correlation coefficient, which is the normalized version of the covariance, are used to measure the dependence between two numerical attributes. In the numerical domain, the calculation of the covariance and the correlation relies on the ability to define a total order over the variables to compare. Specifically, when the greater values of one variable mainly correspond to the greater values of the other variable and the same holds for the smaller values, the covariance is positive because the variables show a similar behavior. In the opposite case, when the greater values of one variable mainly correspond to the smaller values of the other, the variables tend to show opposite behaviors and the covariance is negative. Therefore, the covariance shows the tendency towards linear relationships between variables. To be able to capture this relationship, a total order over the domains of the two variables must exist, so that we can differentiate "large values" and "small values". However, most semantic domains lack a total order; that is, nominal values can be ordered in as many different ways as reference points. Hence, we cannot identify "large values" and "small values", but just pairwise distances between concepts. For this reason, it is not possible to carry out a direct adaptation of the numerical covariance to the semantic domain, as we did for the variance.

To address this issue, we opted for alternative measures of statistical dependence that rely on distances between values rather than a total order: the *distance covariance* and the *distance correlation*. These measures were recently introduced by Székely [99] and use the distance between value pairs as the fundamental part of its calculation. Essentially, these measures quantify up to which point the two variables are directly or independently dispersed, where dispersion is measured according to the pairwise distances between all pairs of values of each variable. Unlike the Pearson correlation coefficient, equation (2.7), the distance correlation is capable of detecting a wider variety of dependence relationships: whereas the Pearson correlation coefficient only recognizes linear dependencies, the distance correlation recognizes linear and nonlinear dependencies. Moreover, because these measures compare dispersions rather than actual values, they can be employed on pairs of variables of different cardinality. Even though being new, these measures have been applied in a variety of scenarios [100, 101]; however, as far as we know,

our work is the first incorporating semantics into the definition of the distance covariance and correlation measures in order to measure the semantic dependence between nominal attributes.

Let X^a and X^b be two nominal attributes of a dataset X. If their samples have n records, we obtain the following set of value pairs $(X^a, X^b) = \{(x_i^a, x_i^b) : 1, \ldots, n\}$ where the pair (x_i^a, x_i^b) represents the value of the attributes X^a and X^b for the record i. For example, in a dataset X of n patients from a hospital, where the attribute X^a stores diagnoses and the attribute X^b stores medical procedures, the pair (x_i^a, x_i^b) represents the diagnosis and the medical procedure of the patient i.

According to Székely, the first step to compute the distance covariance is to obtain a distance matrix for each attribute, which captures the dissimilarity of the values of an attribute. Subsequently, the distance matrices are used to compute double centered distance matrices. In the semantic domain, we propose using semantic distances to measure the dissimilarity between the values of a nominal attribute. In this way, we define SD_{X^a} as the $(n \times n)$ *semantic distance matrix* of the attribute $X^a = (x_1^a, \ldots, x_n^a)$ and SD_{X^b} as the matrix of the attribute $X^b = (x_1^b, \ldots, x_n^b)$, such that

$$SD_{X^a} = \left(sd_{ij}^{X^a} \right)_{i,j=1}^{n}, \quad SD_{X^b} = \left(sd_{ij}^{X^b} \right)_{i,j=1}^{n}, \tag{4.20}$$

where elements $sd_{ij}^{X^a}$ and $sd_{ij}^{X^b}$ are semantic distances. Thereby, $sd_{ij}^{X^a} = sd(x_i^a, x_j^a)$ is the semantic distance between the values of the attribute X^a in positions i and j. In line with the previous example, $sd_{ij}^{X^a}$ would express the semantic distance between the main diagnoses of patients i and j. Analogously, $sd_{ij}^{X^b}$ represents the semantic distance between the values of the attribute X^b in positions i and j, which, in our example, expresses the semantic distance between the medical procedures of patients i and j. Therefore, to build a semantic distance matrix it is necessary to compute all the pairwise semantic distances between the values of the corresponding attribute. Note that both SD_{X^a} and SD_{X^b} have a zero diagonal because the semantic distance between two identical concepts is zero.

By means of the semantic distance matrices, we can compute the *double centered semantic distance matrices*. In short, these matrices are semantic distance matrices with the row and column means subtracted and the grand mean added. Formally, let Δ_{X^a} and Δ_{X^b} be two ($n \times n$) double centered semantic distance matrices whose elements $\delta_{ij}^{X^a}$ and $\delta_{ij}^{X^b}$ are computed from their respective matrices SD_{X^a} and SD_{X^b} as follows

$$\begin{aligned} \Delta_{X^a} &= \left(\delta_{ij}^{X^a}\right)_{i,j=1}^{n} = \left(sd_{ij}^{X^a} - \overline{sd}_{i.}^{X^a} - \overline{sd}_{.j}^{X^a} + \overline{sd}_{..}^{X^a} \right)_{i,j=1}^{n} \\ \Delta_{X^b} &= \left(\delta_{ij}^{X^b}\right)_{i,j=1}^{n} = \left(sd_{ij}^{X^b} - \overline{sd}_{i.}^{X^b} - \overline{sd}_{.j}^{X^b} + \overline{sd}_{..}^{X^b} \right)_{i,j=1}^{n} \end{aligned} \tag{4.21}$$

where $\overline{sd}_{i.}^{X^a}$ is the mean of i-th row from matrix SD_{X^a}, $\overline{sd}_{.j}^{X^a}$ is the mean of j-th column from matrix SD_{X^a} and $\overline{sd}_{..}^{X^a}$ is the mean of all values from matrix SD_{X^a}:

$$\overline{sd}_{i.}^{X^a} = \frac{1}{n}\sum_{j=1}^{n} sd_{ij}^{X^a}, \qquad \overline{sd}_{.j}^{X^a} = \frac{1}{n}\sum_{i=1}^{n} sd_{ij}^{X^a}, \qquad \overline{sd}_{..}^{X^a} = \frac{1}{n^2}\sum_{ij=1}^{n} sd_{ij}^{X^a} \tag{4.22}$$

Note that, when i is equal to j, $\overline{sd}_{i.}^{X^a}$ is equal to $\overline{sd}_{.j}^{X^b}$ by the commutative property of the semantic distance measure. Analogously, $\overline{sd}_{i.}^{X^b}$ is the mean of i-th row from matrix SD_{X^b}, $\overline{sd}_{.j}^{X^b}$ is the mean of j-th column from matrix SD_{X^b} and $\overline{sd}_{..}^{X^b}$ is the mean of all values from matrix SD_{X^b}.

With all the above elements computed in the semantic domain, we propose measuring the semantic dependency of two nominal attributes by means of the following definitions:

Definition 6. The semantic distance covariance between two nominal attributes X^a and X^b, denoted by $sdCov(X^a, X^b)$, is the square root of the arithmetic mean of the product $\delta_{ij}^{X^a}\delta_{ij}^{X^b}$.

$$sdCov\left(X^a, X^b\right) = \frac{1}{n}\sqrt{\sum_{i,j=1}^{n} \delta_{ij}^{X^a} \delta_{ij}^{X^b}} \tag{4.23}$$

According to [99], the distance covariance satisfies $sdCov(X^a, X^b) \geq 0$. Further, $sdCov(X^a, X^b) = 0$ if and only if X^a and X^b are independent. This property is a consequence of dealing with centered distances and allows measuring nonlinear associations.

Definition 7. The *semantic distance correlation* between two nominal attributes X^a and X^b, denoted by $sdCov(X^a, X^b)$, is the nonnegative number obtained by dividing the distance covariance by the product of the distance standard deviations of the attributes.

$$sdCor(X^a, X^b) = \begin{cases} \dfrac{sdCov(X^a, X^b)}{\sqrt{sdVar(X^a) \times sdVar(X^b)}}, & sdVar(X^a) \times sdVar(X^b) > 0 \\ 0, & sdVar(X^a) \times sdVar(X^b) = 0 \end{cases}, \tag{4.24}$$

In the above equation $sdVar(X^a)$ and $sdVar(X^b)$ are the *semantic distance variances* of X^a and X^b. The distance variance is a particular case of distance covariance where the two attributes are identical; therefore, the *semantic distance variance* $sdVar(X^a)$ of X^a is the nonnegative number defined by $sdCov(X^a, X^b)$, similarly for the attribute X^b.

$$\begin{aligned} sdVar\left(X^a\right) &= sdCov\left(X^a, X^a\right) = \frac{1}{n}\sqrt{\sum_{i,j=1}^{n} \delta_{ij}^{X^a} \delta_{ij}^{X^a}} \\ sdVar\left(X^b\right) &= sdCov\left(X^b, X^b\right) = \frac{1}{n}\sqrt{\sum_{i,j=1}^{n} \delta_{ij}^{X^b} \delta_{ij}^{X^b}} \end{aligned} \tag{4.25}$$

The semantic *distance variance* of an attribute is equal to zero if and only if all its values are identical. As in the numerical domain, the *semantic distance standard deviation* is the square root of the distance variance.

The *semantic distance correlation* satisfies $0 \leq sdCor(X^a, X^b) \leq 1$, and $sdCor(X^a, X^b) = 0$ if and only if X^a and X^b are semantically independent. Values close to zero of *sdCor* indicate very weak association between the meanings of X^a and X^b. Greater values of *sdCor* suggest a stronger semantic association. If $sdCor(X^a, X^b) = 1$ then there is a linear relationship between X^a and X^b and exists a vector *v*, a non-zero real number *c* and an orthogonal matrix *R* such that $B = v + cAR$.

4.7 Semantic sorting operator

Some privacy-preserving methods (e.g., rank swapping) require from an order relation on the values of the input attribute in order to sort them. To enforce this operation on nominal data, it is necessary to define a binary relation that allows sorting all nominal values of the attribute. However, because nominal data are non-ordinal, a priori, it is not possible to carry out such operation. Below, we discuss this issue and propose a solution.

An order relation describes the criterion whereby a collection of values is organized in a sequence following statements such as "*x* is less than or equal to *y*". In natural numbers, we say that a number *x* is less than or equal to a number *y*, i.e., $x \leq y$, if there exists another natural number *z* such that $x + z = y$. According to this criterion, the position of natural numbers in the order sequence is determined by the quantity they represent. In the domain of nominal data, this order relation cannot be applied directly because the meanings of nominal values (i.e., the concepts they refer to) do not denote quantities; e.g., in the following sample of the *disease* attribute $X^a = \{$*coma*, *hepatic coma*, *disorder of nervous system*$\}$ it does not make sense to say that *coma* is less or greater than *hepatic coma*. If nominal data could be sorted according to a magnitude, those data should be considered ordinal categorical data. An example of ordinal categorical attribute may be *color*, where the different categories may be sorted on basis of their wave lengths.

Nonetheless, beyond artificial orders such as the alphabetical order, nominal data may be sorted while considering their semantics by applying statements such as "*x* is a *y*" or "*x* is part of *y*", as formalized in a background ontology. By applying the statement "*x* is a *y*", we can determine if a concept

specializes another one. For example, if we consider the fragment of the medical ontology SNOMED-CT shown in Figure 4-1, whose edges represent *is-a* relationships, we can say that *coma* is a *disorder of nervous system*. Formally, the binary relation "*x* is a *y*" (and, similarly, for "*x* is part of *y*") applied on a nominal attribute X^a mapped on an ontology *O*, denoted by $\subseteq^{X^a}$, is an order relation that holds the following properties for all x_i^a, x_j^a and x_l^a in X^a:

- Reflexivity: $x_i^a \subseteq^{X^a} x_i^a$.
- Antisymmetry: if $x_i^a \subseteq^{X^a} x_j^a$ and $x_j^a \subseteq^{X^a} x_i^a$ then $x_i^a = x_j^a$.
- Transitivity: if $x_i^a \subseteq^{X^a} x_j^a$ and $x_j^a \subseteq^{X^a} x_l^a$ then $x_i^a \subseteq^{X^a} x_l^a$.

Now, thanks to $\subseteq^{X^a}$, the above sample can be sorted considering the semantics of its values as follows: *hepatic coma* $\subseteq$ *coma* $\subseteq$ *disorder of nervous system*. However, $\subseteq^{X^a}$ lacks a feature that is key to be able to sort all values of an attribute: the totality property. This property gives rise to a total order on a set that satisfies reflexivity, antisymmetry and transitivity since each element can be compared to any other element; e.g., in natural numbers, any pair of numbers is comparable under $\leq$, i.e., $x \leq y$ or $y \leq x$. However, as we can see in the following sample X^a = {*hypoglycemic coma*, *coma*, *hepatic coma*, *disorder of nervous system*}, there are nominal values that are incomparable under $\subseteq^{X^a}$, e.g., neither *hypoglycemic coma* $\subseteq^{X^a}$ *hepatic coma* nor *hepatic coma* $\subseteq^{X^a}$ *hypoglycemic coma*.

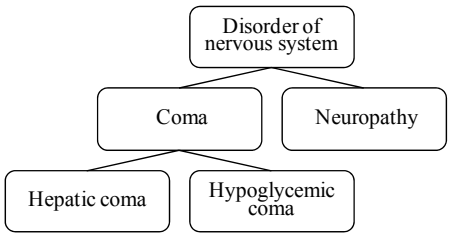


Figure 4-1 Example of taxonomy associated to the domain *Disease*, extracted from the SNOMED-CT medical ontology

Therefore, to be able to sort all values of an attribute, we require a binary relation that fulfills the totality property as an alternative to the partial order $\subseteq^{X^a}$. In an attempt to construct a totally ordered set from a partial order that permits to sort the non-comparable data, Torra [75] proposes defining a total topological order consistent with the partial order. Subsequently, a cumulative function of the frequency of the nominal values in the attribute defined on this topological order is used to rank the attribute. However, if this order relation is applied on the partial order $\subseteq^{X^a}$, the sorted attribute would lack semantic coherence because the result is partially determined by the attribute frequency distribution, rather than the semantics underlying to the values; that is, the fact that two nominal attributes are equally frequent in a sample, does not imply that they have equal or even similar meanings.

To obtain a semantically-coherent result, we propose an order relation based on the notion of closeness to a reference point. Given a reference value x^a_{ref} in X^a, a value x^a_i is less than or equal to a value x^a_j, if x^a_i is closer of x^a_{ref} than x^a_j. Because the closeness of a nominal value to another of reference is determined by the difference between the semantics they convey, we can use the notion of *semantic distance* for this purpose. For example, if we want to sort the nominal values {*disorder of nervous system*, *coma*, *neuropathy*, *hepatic coma*, *hypoglycemic coma*} mapped into the taxonomy of Figure 4-1, firstly, we must select a value in the set as reference point for the order relation. Secondly, we must calculate the semantic distance between each value of the set and that reference point. With this, the values of the set can be coherently sorted according to the computed distances. In the example, if the reference point was the concept *coma*, the sorted set would follow the sequence shown in Figure 4-2 (the semantic distances have been computed with the Wu and Palmer measure). As we can see, if we change the reference point, we obtain a different sort sequence, as shown in Figure 4-3. Therefore, this order relation generates as many rank sequences as different values has the dataset.

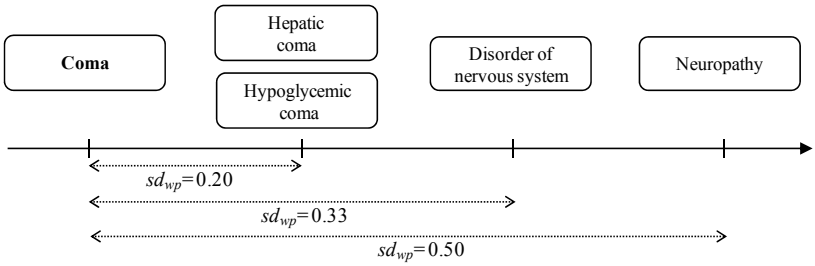


Figure 4-2 Example of ascending sorted sequence when the reference point is *Coma*.

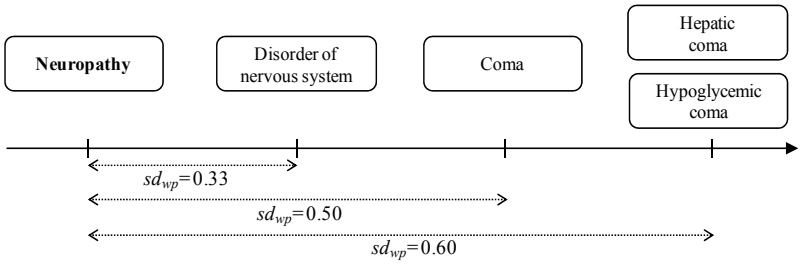


Figure 4-3 Example of ascending sorted sequence when the reference point is *Neuropathy*.

A formal definition of the order relation on nominal data based on the closeness to a reference point is provided below.

Definition 8. The *order relation* on a nominal attribute X^a, given a reference value x^a_{ref} in X^a, denoted by $\leq_{x^a_{ref}}$, is defined as a binary relation where a value x^a_i is less than or equal to a value x^a_j, i.e., $x^a_i \leq_{x^a_{ref}} x^a_j$, if and only if the *semantic distance , sd*$(\cdot,\cdot)$, between x^a_i and x^a_{ref} is less than or equal to the *semantic distance* between x^a_j and x^a_{ref}.

$$\leq_{x^a_{ref}} = \left\{ x_i, x_j \in X^a : \quad x_i \leq_{x^a_{ref}} x_j \mid sd(x_i, x_{ref}) \leq sd(x_j, x_{ref}) \right\} \qquad (4.26)$$

The relation $\leq_{x^a_{ref}}$ holds the following properties for any x^a_i, x^a_j and x^a_l in X^a:

- Reflexivity: $x^a_i \leq_{x^a_{ref}} x^a_i$.
- Transitivity: if $x^a_i \leq_{x^a_{ref}} x^a_j$ and $x^a_j \leq_{x^a_{ref}} x^a_l$ then $x^a_i \leq_{x^a_{ref}} x^a_l$.
- Totality: $x^a_i \leq_{x^a_{ref}} x^a_j$ or $x^a_j \leq_{x^a_{ref}} x^a_i$, i.e., any pair of values in X^a is comparable under the relation $\leq_{x^a_{ref}}$.

Note that, $\leq_{x^a_{ref}}$ is a total preorder relation (or weak order relation), but not a total order relation because, despite fulfilling the totality property, it does not satisfy the antisymmetric property.

The antisymmetry holds if $x^a_i \leq_{x^a_{ref}} x^a_j$ and $x^a_j \leq_{x^a_{ref}} x^a_i$ then $x^a_i = x^a_j$; but, as shown in Figure 4-1, this condition does not fulfill in all cases, e.g., *Hepatic coma* $\leq_{Coma}$ *Hypoglycemic coma* and *Hypoglycemic coma* $\leq_{Coma}$ *Hepatic coma*, but *Hepatic coma* ≠ *Hypoglycemic coma*. The fact that $\leq_{x^a_{ref}}$ lacks antisymmetric property implies that there may be different values tied in semantic distance w.r.t. x^a_{ref}.

The sequence of values of X^a sorted in ascending order according to $\leq_{x^a_{ref}}$ is represented by $\overrightarrow{X^a} = \langle x^a_{(1)}, \ldots, x^a_{(n)} \rangle$. Note that x^a_i and $x^a_{(i)}$ represent the i^{th}-unordered and i^{th}-ordered value of X^a, respectively, i.e., $rank(x^a_{(i)}) = i$. Obviously, the first value in the ranking is x^a_{ref}, i.e., $x^a_{(1)} = x^a_{ref}$. If there are tied values w.r.t. x^a_{ref}, these will place in contiguous positions in the ranking.

4.8 Conclusion

The perturbative techniques depicted in Chapter 2 base their masking process on arithmetical operations. As shown in Table 4.1, the numerical noise addition mechanism requires calculating the mean and the variance of the

input attribute to generate a noise sequence that reflects the degree of dispersion of the original values. The correlated noise addition mechanism additionally requires computing the covariance between attribute pairs to generate the noise sequences that reflect the degree of correlation between the attributes. On the other hand, the standard rank swapping mechanism requires sorting the values of the input attribute to be able to perform rank-distance swaps during the permutation process, thereby restricting and controlling the information loss associated to each swap.

Table 4.1 Semantic operators required in the semantic noise addition and semantic rank swapping methods.

Semantic Operator	Noise Addition	Rank Swapping
Difference	X	X
Mean	X	
Variance	X	
Covariance	X	
Correlation		
Sorting		X

Numerical data can be directly and easily manipulated and compared by means of classical mathematical operators. However, the handling of nominal data entails a greater difficulty because they are of a finite, discrete, textual and non-ordinal nature. Moreover, because nominal data utility is closely related to the preservation of data semantics, any data transformation or calculation performed to anonymize data should consider the *meaning* of the values. For such reason, to properly deal with nominal data, perturbative methods require adapting the arithmetical operators involved in the masking process to the semantic domain.

The main hypothesis of our work, which will be exploited and evaluated in next chapters, is that the use of ontologies allows a better interpretation of nominal data during the masking process, thus producing anonymized data of higher quality. For such purpose, and by exploiting the formal knowledge offered by ontologies, we have defined semantic versions of the difference, the mean, the variance, the covariance and the sorting operators. These can not only guide the masking process, but they can be employed to measure the utility of the protected outcomes in a semantically coherent way.

Chapter 5 **Semantic Rank Swapping**

In this chapter we present a rank swapping method capable of protecting nominal categorical data from a semantic perspective by exploiting the formal semantics provided by an ontology and using the semantics operators defined in Chapter 4. Our objective is to provide mechanisms to control the degree of permutation, in order to enforce a certain level of protection while preserving, as much as possible, the semantic features, and thus, the analytical utility, of the data. In particular, we propose semantically-grounded rank swapping solutions to perturb individual nominal attributes and multivariate nominal datasets. The latter is especially relevant because it is capable of protecting multivariate nominal datasets while reasonably preserving the correlation among non-independent attributes (e.g., among symptoms, diagnosis and treatments), which is of outmost importance for research.

5.1 Introduction

As explained in Chapter 2, *rank swapping*, which is based on the idea of proximity swapping [39], ranks the values of each attribute in ascending order for later swapping each value with another one randomly chosen within a restricted size range. Thus, the higher the range size, the higher the ambiguity in the re-identification inferences and the lower the disclosure risk; but also, the lower the data utility, because swapped values would tend to be less similar. Concerning data utility, and on the contrary to other data protection mechanisms [5], rank swapping perfectly preserves univariate statistics, such as the mean, the variance and the frequency distribution, because the values in the protected attribute are the same as those in the original attribute but permuted. For this same reason, rank swapping also preserves other very useful features for data analysis, such as data granularity or outlying values.

5.2 Semantic management of nominal data in rank swapping

Due to rank swapping relies on the ability to sort the values of the input attribute, as discussed in Chapter 2, it only can deal with numerical and ordinal categorical data [5]. To address this issue, we propose to use the semantic sorting operator defined in Section 4.7, which enables us to sort all the nominal values of an attribute coherently with their underlying semantics through a total preorder relation. By using this semantic sorting operator, we can adapt rank swapping to the semantic domain of nominal data.

To carry out the sorting operation, we can use any semantic distance measure defined in Section 4.3 on the taxonomy defined in Section 4.2.

5.3 Semantic univariate rank swapping method

In this section, we propose a semantically-grounded univariate rank swapping method for individual nominal attributes that pursues a twofold objective:

1. To control and bind the swapping process according to a configurable level of permutation.
2. To maximize data utility by (i) preserving the semantic mean and variance of each nominal attribute and (ii) obtaining an information loss (error) proportional to the desired level of permutation.

Because our mechanism is general, it accommodates several variations depending on how the swapping ranges are built. The first one intuitively adapts the idea of "swapping interval" of the standard rank swapping method to nominal data. In a second approach, we propose to dynamically build the swapping ranges to minimize the information loss associated to each swap.

To generate the permuted version X^* of the original dataset X, the semantic univariate rank swapping method must be independently applied on each nominal attribute. Following the notation of Section 2.2.1, let $X^a = (x_1^a, \ldots, x_n^a)$ be a nominal attribute in X. Like the numerical method, the records of X must be ranked in ascending order by values x_i^a of the attribute

X^a. To do this, we use the total preorder relation $\leq_{x^a_{ref}}$ defined in equation (4.26). As reference point x^a_{ref} of the order relation, we consider the boundary of the attribute, i.e. the most semantically-distant value from X^a. To find the most distant value from a set of nominal data in a semantically-coherent way, we use the notion of *marginality* [102]. Specifically, the marginality of a nominal value of a sample shows how outlying is that element with respect to the remaining values of the sample according to the aggregation of semantics distances. On this basis, we propose the following definition of the *most semantically-distant value of a nominal attribute*:

Definition 9. The *most semantically-distant value of a nominal attribute* X^a, denoted by *MostDistantValue*(X^a), is the value x^a from X^a that maximizes the sum of the semantic distances with respect to all x^a_i in X^a.

$$MostDistantValue\left(X^a\right) = \arg\max_{x^a \in X^a}\left(\sum_{x^a_i \in X^a} sd\left(x^a, x^a_i\right)\right) \tag{5.1}$$

On the other hand, in order to offer a configurable level of permutation, and thus, to satisfy the first objective of the method, we should allow the user defining the length of the swapping interval or range. Similar to numerical rank swapping method, this is accomplished through an input parameter *k*, which represents the length of the swapping interval in number of records. Notice that related works use *p* as input parameter, which specifies the percentage of the records in *X* in the interval, i.e., $k=p.n/100$. In our case, by setting *k*, the rank of two swapped values cannot differ by more than *k* records, which provides a clearer privacy guarantee than the use of percentages. Specifically, with our approach, we guarantee that an attacker with access to the permuted attribute would only be able to infer the original values with a probability at most $1/k$. This guarantee against the re-identification fulfills *probabilistic k-anonymity* [26], which is a privacy model that provides the same protection level as *k-anonymity* [4, 31]; but, whereas *k-anonymity* requires each record to be indistinguishable from at least *k*-1 other records, *probabilistic k-anonymity* only constraints the re-identification probability. Notice that when *k* increases, the dissimilarity or semantic distance between the original value (*before-swap value*) and the permuted value (*after-swap*

value) tends to increase; this prevents re-identifications, but deteriorates the permuted data quality because the information loss associated to each swap tends to increase.

The semantic univariated rank swapping method is formalized in SRS-Algorithm1. Firstly, the taxonomy $\tau(X^a)$ associated to the attribute X^a is obtained from the ontology O by following the procedure detailed in Section 4.2. Secondly, in line 2, the reference point x^a_{ref} used to rank X^a is computed by using equation (5.1); this value is the most semantically-distant value of X^a. According to x^a_{ref}, the values in X^a are ranked in ascending order in line 3 by using the total preorder relation defined in equation (4.26). Then, in line 4, the values of the ranked attribute $\overrightarrow{X^a} = \langle x^a_{(1)}, \ldots, x^a_{(n)} \rangle$ are labeled as *unswapped*. In lines 5-12, each unswapped value $x^a_{(i)}$ is permuted by another unswapped value randomly chosen within a restricted range through the procedure *swap_value*. This swapping range is composed of the k values following to $x^a_{(i)}$ in the ranking $\langle x^a_{(1)}, \ldots, x^a_{(n)} \rangle$, i.e., the interval $\left[x^a_{(i+1)}, x^a_{(i+k)} \right]$. The size of the range is kept in k values, except when the index $i+k$ is greater than n, i.e., is greater than the size of the attribute. For this reason, the upper limit of the interval is the lower value of $\{i+k, n\}$, i.e., $\left[x^a_{(i+1)}, x^a_{(\min\{i+k,n\})} \right]$. After each swap, in lines 22 and 24, the processed values are labeled as *swapped*.

SRS-Algorithm1. *Semantic univariate rank swapping* method

Input :

X^a : nominal attribute with *n* records

O: ontology

k: length of the swapping interval in number of records

Output :

X^{a*} : permuted nominal attribute

1: $\tau(X^a) \leftarrow$ obtain_taxonomy(X^a, O)

2: $x^a_{ref} \leftarrow$ obtain_MostDistantValue(X^a) $\quad // x^a_{ref} = \arg\max_{x^a \in X^a} \left(\sum_{x^a_i \in X^a} sd\left(x^a, x^a_i\right) \right)$

3: $\overrightarrow{X^a} \leftarrow$ ascendingRank_attribute(X^a, x^a_{ref})

$// \overrightarrow{X^a} = \langle x^a_{(1)}, \ldots, x^a_{(n)} \rangle$ such that $x^a_i \leq_{x^a_{ref}} x^a_j$ if $sd(x^a_i, x^a_{ref}) \leq sd(x^a_j, x^a_{ref})$

4: label_unswapped($\overrightarrow{X^a}$)

5: **for all** $x^a_{(i)}$ in $\overrightarrow{X^a}$ **do**

6: **if** $x^a_{(i)}$ is unswapped **then**

7: $lower\_limit \leftarrow i+1$

8: $upper\_limit \leftarrow \min\{i+k, n\}$

9: $interval \leftarrow$ obtain_SwappingInterval($\overrightarrow{X^a}, lower\_limit, upper\_limit$)

$// interval = \left[x^a_{(i+1)}, x^a_{(\min\{i+k,n\})} \right]$

10: swap_value($interval, x^a_{(i)}$)

11: **end if**

12: **end for**

13: **return** X^{a*}

14 : **procedure** swap_value ($swappingRange, x_{ref}^{attr}$)

15 : $\overline{swappingRange} \leftarrow$ obtain_UnswappedValues($swappingRange$)

//$\overline{swappingRange}$ is the set of the unswapped values in *sawppingRange*

16 : **if** $\overline{swappingRange}$ is not empty **then**

17 : $x_{swa}^{attr} \leftarrow$ select_SwappingValue($\overline{swappingRange}$) //random selection

18 : $i \leftarrow$ obtain_Index(X^{attr}, x_{ref}^{attr})

19 : $j \leftarrow$ obtain_Index(X^{attr}, x_{swa}^{attr})

20: $X^{attr*}[i] \leftarrow x_{swa}^{attr}$

21: $X^{attr*}[j] \leftarrow x_{ref}^{attr}$

22 : label_swapped(X^{attr}, x_{swa}^{attr})

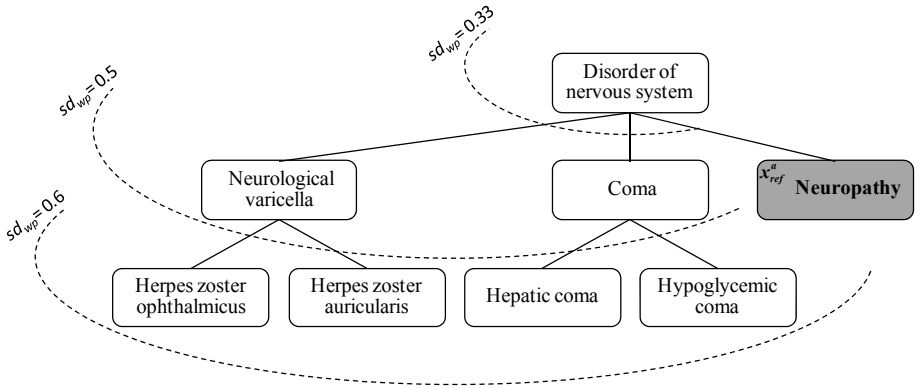
23 : **end if**

24 : label_swapped(X^{attr}, x_{ref}^{attr})

25 : **end procedure**

In SRS-Algorithm1, the attribute is only ranked once at the beginning of the process. As discussed in Section 4.7, the total preorder relation $\leq_{x_{ref}^a}$ yields as many order sequences as different values the attribute has. This means that the ranking obtained at the beginning of the process is suitable (w.r.t. bounding the maximum semantic permutation resulting from each swap) to build the swapping interval of the first treated value, but not to build the intervals of the remaining values. Because the attribute is not well-ordered for those remaining values, very different values may be swapped from the second swap and upwards, thus incurring in an information loss much higher than that expected from the permutation level *k*. This issue is illustrated in Figure 5-1 for a sample of the *disease* attribute X^a={*Disorder of nervous system*, *Neurological varicella*, *Coma*, *Neuropathy*, *Herpes zoster ophthalmicus*, *Herpes zoster auricularis*, *Hepatic coma, Hypoglycemic coma*} and k =2. The nominal values of X^a have been unequivocally associated with concepts modeled in the SNOMED-CT medical ontology. After ranking the values of X^a w.r.t. the most semantically-distant value of X^a, *Neuropathy*, we can see that the established

ranking is suitable to build the swapping interval of $x^a_{(1)} = Neuropathy$ because the resulting interval is composed of the two most semantically-similar values to *Neuropathy* (i.e, *Disorder of nervous system* and *Coma*). However, for $x^a_{(6)} = Hepatic\ coma$, the swapping interval includes values that are much more semantically distant that expected from the value of *k* (e.g, *Herpes zoster ophthalmicus*, whose semantic distance w.r.t. *Hepatic coma* is sd_{wp}= 0.67 by using equation (6.1) and (4.6)).



$\overrightarrow{X^a} = \langle x^a_{(1)} = \text{Neuropathy},\ x^a_{(2)} = \text{Disorder of nervous system},\ x^a_{(3)} = \text{Coma},$
$x^a_{(4)} = \text{Neurological varicella},\ x^a_{(5)} = \text{Hypoglycemic coma},$
$x^a_{(6)} = \text{Hepatic coma},\ x^a_{(7)} = \text{Herpes zoster auricularis},\ x^a_{(8)} = \text{Herpes zoster ophthalmicus}\rangle$

Figure 5-1. Example of swapping intervals in an ascending ranked attribute w.r.t. MostDistantValue(X^a) = *Neuropathy*

To solve this issue, we propose a variation of SRS-Algorithm1 that better preserves data semantics by swapping the original values with others within a semantic distance coherent with the input parameter *k*, thereby satisfying the objective 2(ii). The difference of this approach with respect to the previous one lies in the way of generating the swapping intervals during the permutation process. In this new approach, we propose to re-rank the attribute for each value x^a_i to swap by using x^a_i as reference point, i.e., $x^a_{ref} = x^a_i$. In this way, once the attribute has been ascending ranked w.r.t. x^a_i, the swapping range

will be the set of the *k* semantically-closest values to x_i^a. This range, which we name *swapping cluster*, is formally defined bellow.

Definition 10. The *swapping cluster* associated to a reference point x_{ref}^a from a nominal attribute X^a, denoted by $C_{x_{ref}^a}^{X^a}$, is defined as the set of the *k* semantically-closest values to x_{ref}^a in X^a, i.e. the *k* first values in the ranked attribute $\overrightarrow{X^a} = \langle x_{(1)}^a, \ldots, x_{(n)}^a \rangle$, excluding $x_{(1)}^a$.

$$C_{x_{ref}^a}^{X^a} = \left\{ x_{(2)}^a, \ldots, x_{(1+k)}^a \right\}, \quad 1 \leq k < n \tag{5.2}$$

Because a new cluster $C_{x_{ref}^a}^{X^a}$ needs to be generated every time a new value x_i^a must be swapped, we call this process *dynamic clustering*.

To prioritize the permutation of the most marginal values, which are those whose swaps entail more information loss, we propose the strategy of *clustering at opposite ends*, which is inspired by clustering methods such as MDAV [46]: each time a new cluster must be generated, it must be as far as possible from the previously generated cluster. In other words, the reference points of two consecutive clusters must *maximize* their semantic distance. To do this, the method starts by building the *swapping cluster* at the boundary of X^a; that is, the first element to swap and, thus, the reference point $x_{ref}^{a'}$ for the first cluster, will be the *most distant value* of X^a (equation (5.1)). Then, to select the second element to swap, and thus, the second reference point $x_{ref}^{a''}$, we look for the value in X^a that is semantically-farthest from $x_{ref}^{a'}$, as formalized in Definition 4. Note that, $x_{ref}^{a''}$ must be selected among the still unswapped values in X^a.

Definition 11. The most semantically-distant value to a reference value x_{ref}^a in a nominal attribute X^a, denoted as *MostDistantValue*(X^a, x_{ref}^a), is the value x^a from X^a that maximizes the semantic distance with x_{ref}^a.

$$MostDistantValue\left(X^a, x^a_{ref}\right) = \arg\max_{x^a \in X^a}\left(sd\left(x^a, x^a_{ref}\right)\right) \qquad (5.3)$$

If there were several values at the same maximum semantic distance, the algorithm selects one at random. Note that MostDistantValue(X^a, x^a_{ref}) is the last value in the ranking of x^a_{ref}.

Once a reference point x^a_{ref} has been selected (by using equation (5.3)), the corresponding swapping cluster is built according to equation (5.2). Finally, x^a_{ref} is swapped with a value from the cluster randomly chosen among those that still have not been swapped. In this way, like SRS-Algorithm1, the rank of two swapped values cannot differ by more than *k* records.

SRS-Algorithm2 formalizes the above-described procedure. As stated above, the first element to swap and, thus, the first reference point, is the most semantically-distant value of X^a, selected by using equation (5.1). Then, in line 5, the swapping cluster is built around the reference point by using equation (5.2). In line 6, through the procedure *swap_value* of SRS-Algorithm1, the value x^a_{ref} is swapped with another unswapped value randomly chosen within the cluster. Finally, in line 7, the next reference point and, thus, the next value to swap, is chosen by applying the strategy of *clustering at opposite ends* (equation (5.3)). This process is repeated until all values in X^a are swapped.

SRS-Algorithm2. *Semantic univariate rank swapping* method based on *dynamic clustering at opposite ends*.

Input :

X^a : nominal attribute with *n* records

O: ontology

k: length of the swapping cluster in number of records

Output :

X^{a*} : permuted nominal attribute

1 : $\tau(X^a) \leftarrow$ obtain_taxonomy(X^a, O)

2 : $\overline{X^a} \leftarrow X^a$ // $\overline{X^a}$ is the set of unswapped values in X^a

3 : $x^a_{ref} \leftarrow$ obtain_MostDistantValue$(\overline{X^a})$

$$//x^a_{ref} = \arg\max_{x^a \in \overline{X^a}} \left(\sum_{x^a_i \in \overline{X^a}} sd(x^a, x^a_i) \right)$$

4 : **while** X^a has unswapped values **do**

5: $C^{X^a}_{x^a_{ref}} \leftarrow$ obtain_SwappingCluster(X^a, x^a_{ref}, k)

$//C^{X^a}_{x^a_{ref}}$ is the set of the k semantically-closest values to x^a_{ref} in X

6 : swap_value$(C^{X^a}_{x^a_{ref}}, x^a_{ref})$

7 : $x^a_{ref} \leftarrow$ obtain_MostDistantValue$(\overline{X^a}, x^a_{ref})$

$$//x^a_{ref} = \arg\max_{x^a \in \overline{X^a}} \left(sd(x^a, x^a_{ref}) \right)$$

8 : **end while**

9 : **return** X^{a*}

Regarding the objective 2(i), because the values in the permuted attribute are the same as those in the original attribute but swapped, by definition, the semantic mean and variance are perfectly preserved for each individual attribute.

5.4 Semantic multivariate rank swapping method

The rank swapping method presented in the previous section preserves, by construction, the semantic mean and the variance of the attributes and incurs in an information loss (error) proportional to the desired level of permutation. However, because the method is independently applied to each attribute of the dataset, the potential correlation among attributes is likely to be significantly hampered, as discussed in Section 2.4. To solve this issue, we propose a semantic rank swapping method that, in addition to fulfilling the objectives of the univariate version, is also capable of reasonably preserving the semantic correlation among non-independent attributes.

Let X be a dataset with m nominal attributes and n records or tuples, such that $X = (X^1, \ldots, X^m) = \{t_i = (x_i^1, \ldots, x_i^m) : \ i = 1, \ldots, n\}$, where the tuple $t_i = (x_i^1, \ldots, x_i^m)$ represents the value of the m attributes for individual i.

In order to preserve, as much as possible, the correlation among the m non-independent attributes of X during the permutation process, we propose considering each tuple as a unit, which thus conveys the relationship between attribute values. In this way, a *swapping cluster* will be composed of the k semantically-closest tuples to a reference tuple. Like SRS-Algorithm2, the clusters will be dynamically built at opposite ends to minimize the information loss, but now they will encompass records rather than individual attributes. After obtaining the swapping cluster of a reference tuple, each value in the reference tuple is swapped with another value of the same attribute, randomly chosen among those in the cluster that still have not been swapped. Because the swap is independently carried out for each attribute value of the reference tuple, the resulting records will be different from those in the original dataset, thus preventing re-identification. Nonetheless, because the swapping range is delimited by semantically similar tuples, attribute values within the swapping range will be both semantically similar within each attribute (which minimizes information loss) and semantically interrelated with the values of the other attributes (which contributes to preserve the attribute correlation).

Below, definitions 8-11 are adapted to work with nominal tuples rather than individual attributes.

Definition 12. The order relation on a dataset X of m nominal attributes, given a reference tuple t_{ref} in X, denoted by $\leq_{t_{ref}}$, is defined as a binary relation where a tuple t_i is less than or equal to a tuple t_j, i.e. , $t_i \leq_{t_{ref}} t_j$, if and only if the semantic distance between t_i and t_{ref} is less than or equal to the semantic distance between t_j and t_{ref}, for all t_i and t_j belonging to X.

$$\leq_{t_{ref}} = \left\{ t_{ref}, t_i, t_j \in X : \quad t_i \leq_{t_{ref}} t_j \mid sd(t_i, t_{ref}) \leq sd(t_j, t_{ref}) \right\}, \tag{5.4}$$

where the *semantic distance* between a pair of tuples t_i and t_j is computed as the average of pairwise semantic distances between attribute values:

$$sd\left(t_i, t_j\right) = \frac{1}{m} \sum_{attr=1}^{m} sd\left(x_i^{attr}, x_j^{attr}\right) \tag{5.5}$$

Definition 13. The most semantically-distant tuple in a dataset X of m nominal attributes, denoted by *MostDistantTuple*(X), is the tuple t from X that maximizes the sum of the semantic distances respect to all t_i in X.

$$MostDistantTuple\left(X\right) = \arg\max_{t \in X} \left(\sum_{t_i \in X} sd(t, t_i) \right) \tag{5.6}$$

Definition 14. The swapping cluster associated to a reference tuple t_{ref} in a dataset X of m nominal attributes, denoted by $C_{t_{ref}}^{X}$, is defined as the set of the k semantically-closest tuples to t_{ref} in X, i.e. the k first tuples in the ranked set $\left\langle t_{(1)}, \ldots, t_{(n)} \right\rangle$, excluding $t_{(1)}$.

$$C_{t_{ref}}^{X} = \left\{ t_{(2)}, \ldots, t_{(k+1)} \right\}, \quad 1 \leq k < n \tag{5.7}$$

Definition 15. The most semantically-distant tuple from a reference tuple t_{ref} in a dataset X of m nominal attributes, denoted by *MostDistantTuple*(X,t_{ref}), is the tuple t from X, that maximizes the semantic distance with t_{ref}.

$$MostDistantTuple\left(X, t_{ref}\right) = \arg\max_{t \in X}\left(sd(t, t_{ref})\right) \quad (5.8)$$

The method for *m* attributes is formalized in SRS-Algorithm3. First, in line 4, all tuples in the input dataset *X* are labeled as unswapped in $\overline{X}$. In line 5, the most distant tuple from *X* is obtained by using the equation (5.6). Similar to the SRS-Algorithm2, this tuple is the first to swap and, thus, the first reference point $t_{ref} = (x_{ref}^1, \ldots, x_{ref}^m)$. Then, in line 7, the cluster is built around the reference tuple t_{ref} by using the equation (5.7). In lines 8-10, through the procedure *swap_value* of SRS-Algorithm1, the swaps are independently undertaken for each attribute. Each value x_{ref}^{attr} in t_{ref} is swapped by another unswapped value belonging to the same attribute randomly chosen within the cluster. Finally, in line 11, the next reference point is chosen by applying the idea of *clustering at opposite ends* (equation (5.8)). This process is repeated until all tuples in *X* are swapped. A tuple is considered swapped when all its values have been swapped.

SRS-Algorithm3. *Semantic multivariate rank swapping* method based on *dynamic clustering at opposite ends*.

Input :

X : dataset with m nominal attributes and n records (tuples)

$// X=(X^1,\ldots,X^m)=\left\{t_i=(x_i^1,\ldots,x_i^m):\ i\ =\ 1,\ldots,\ n\right\}$

O: ontology

k: length of the swapping cluster in number of records

Output :

X^* : permuted nominal dataset $\quad //X^*=(X^{1*},\ldots,X^{m*})$

1 : **for each** X^{attr} in X **do**

2 : $\tau(X^{attr})\leftarrow$ obtain_taxonomy$(X^{attr},O)\quad //attr\ =\ 1,\ldots,\ m$

3 : **end for**

4 : $\overline{X}\leftarrow X$

$// \overline{X}=\overline{(X^1,\ldots,X^m)}$ is the set of unswapped tuples

a tuple is considered swapped when all its values have been swapped

5 : $t_{ref}\leftarrow$ obtain_MostDistantTuple$(\overline{X})$

$$//t_{ref}=(x_{ref}^1,\ldots,x_{ref}^m)=\arg\max_{t\in\overline{X}}\left(\sum_{t_i\in\overline{X}}sd(t,t_i)\right)$$

6 : **while** X has unswapped tuples **do**

7: $C_{t_{ref}}^X\leftarrow$ obtain_SwappingCluster(X,t_{ref},k)

$//C_{t_{ref}}^X$ is the set of the k semantically-closest tuples to t_{ref} in X

8 : **for each** X^{attr} in X **do**

9 : swap_value$(C_{t_{ref}}^X[attr],x_{ref}^{attr})$

10 : **end for**

11 : $t_{ref}\leftarrow$ obtain_MostDistantTuple$(\overline{X},t_{ref})$

$$//t_{ref}=\arg\max_{t\in\overline{X}}\left(sd(t,t_{ref})\right)$$

12 : **end while**

13 : **return** X^*

5.5 Conclusion

In this chapter, we have presented a semantically-grounded alternative to the standard rank swapping mechanism that is capable of protecting nominal data while preserving their semantic features. Our proposal relies on the ability to define a total preorder relation on the nominal values according to their semantic similarity; in this way, we limit the information loss resulting from each data permutation and, thus, we better retain the utility of the outcomes.

We have proposed solutions to protect individual nominal attributes and multivariate datasets. The latter is especially innovative (because standard rank swapping algorithms are univariate) and of great interest for data analysis (because it retains the dependence relationship between nominal attributes).

Chapter 6 **Semantic Noise Addition**

In this chapter, we present a noise addition framework capable of distorting nominal data from a semantic perspective. Our objective is twofold: (i) to semantically manage data during the noise-addition process by exploiting the formal knowledge modeled in ontologies, and (ii) to provide mechanisms to tune noise addition while preserving the semantic features of the data as much as possible. In particular, we propose semantically-grounded noise addition solutions to distort individual nominal attributes (uncorrelated noise) and multivariate nominal datasets (correlated noise). The multivariate case is especially relevant because, unlike the related works focusing on nominal data discussed in Chapter 3, our proposal is able to distort multivariate nominal datasets while reasonably preserving the semantic correlation among attributes. To guide the noise addition process, we use the semantic versions of the difference, mean, variance and covariance measures defined in Chapter 4, which are able to capture the meaning conveyed by nominal values.

6.1 Introduction

Noise addition, which distort original values by adding random noise, is characterized by its relatively high utility and low disclosure risk [40, 50]. Moreover, noise addition has a remarkable advantage over the other perturbative methods: it is the only method that is able to fully preserve the dependence relation between the attributes, i.e., the correlation structure of the dataset.

On the other hand, unlike aggregation-based and permutation-based methods, noise addition is able to deal with records individually, which is a very useful feature in scenarios such as in the online anonymization of transactional data [27, 103]. While traditional static datasets must be protected off-line in a homogenous and monolithic way, transactional data streams (dynamic and continuous) need to be protected on the fly. To correlate the

noise with the stream behavior, [104] proposes to use the correlations in different time series while deciding the noise to be added to any particular value. A representative example of private transactional textual data is a user performing queries to a Web Search Engine (WSE), which profiles her according to such queries to provide personalized search services. In this scenario, the user desires to protect the privacy of her profile w.r.t. the WSE while not impairing the WSE functionalities (e.g., query disambiguation [105], query suggestion and refinement [106]). Because the generated profiles may fully characterize the personal features of the users [107, 108], it is desirable to add some uncertainty to the user's queries. In this regard, noise addition could create fake but plausible and semantically related queries from the original ones in a controlled way. This would help to hide the real user details while preserving, as much as possible, the WSE functionalities.

6.2 Semantic management of nominal data in noise addition

As discussed in Chapter 2, noise addition is commonly seen as a method exclusively intended for numerical data because of its mathematical operating principle. On the one hand, many of the operations carried out to manage and transform data in numerical domain require comparing two values, for example, for assessing how far the noisy value must be from the original one according to the noise magnitude to be added. This noise magnitude thus represents the numerical distance or arithmetical difference between the original value and masked one. To work in the nominal domain, Chapter 4 shows that we can replace the arithmetical difference operator by the notion of *semantic distance* on an underlying ontology O. In this way, nominal values can be managed according to the semantics of the concepts to which they refer. On the other hand, as shown Chapter 2, the numerical noise addition mechanism needs to calculate either the variance of the input attribute to generate a random noise sequence that reflects the degree of dispersion of the original values (uncorrelated method) or the covariance matrix of the input dataset to generate the noise sequences that reflect the degree of correlation among the attributes of the dataset (correlated method). To carry out these

operations, we can use the semantic versions of the variance and covariance measures defined in Chapter 4.

Because during the noise addition process on an input attribute X^a, we may obtain noisy values from O different from the original ones, it is necessary to determine what concepts of O are candidates to participate in the distortion process. To ensure the semantic coherence of the results and, in turn, maximize the application range of noise in O, it must be held that:

(i) any value belonging to the *attribute domain* may be a noisy value in the distorted dataset, *attribute domain* $D(X^a)$ being the universe of values that can take the attribute.

(ii) the concepts in O used as noisy replacements must be limited to the concept hierarchy $\leq_C$ (taxonomy) that represents the *attribute domain*; for example, if the attribute domain represents diseases, the taxonomy used during the noise addition should be limited to the set of all diseases in O, because the noisy values must also be diseases.

By applying (i) and (ii) to the semantic domain, the taxonomy $\tau(X^a)$ defined in Section 4.2 should be extended to $\tau(D(X^a))$, that is, the taxonomy that encompasses all the concepts in the attribute domain, rather than the just attribute sample. Formally, $\tau(D(X^a))$ is the concept hierarchy extracted from O that includes all concepts that are taxonomic specializations of $LCS(D(X^a))$.

Note that $\tau(D(X^a))$ delimits the noise application range in O and, therefore, determines the set of concepts in O that are candidates to replace the original values. By using $\tau(D(X^a))$ instead of $\tau(X^a)$, we constrain the replacement of values within the domain of the attribute while making it independent on the attribute sample, as done in the numerical domain.

Figure 6-1 shows an example of the taxonomy associated with the domain of an attribute on a fragment of the medical ontology SNOMED-CT [81]. Let $X^a = \{gastritis, hematoma, gingivitis\}$ be a nominal attribute that stores the diseases of a set of 3 patients. As $D(X^a)$ is the set of all diseases in SNOMED-CT, the $LCS(D(X^a))$ is the concept *Disease* of SNOMED-CT and $\tau(D(X^a))$ is

the taxonomy consisting of the concepts that are taxonomic specializations of $LCS(D(X^a))$, which are shown in gray in Figure 6-1.

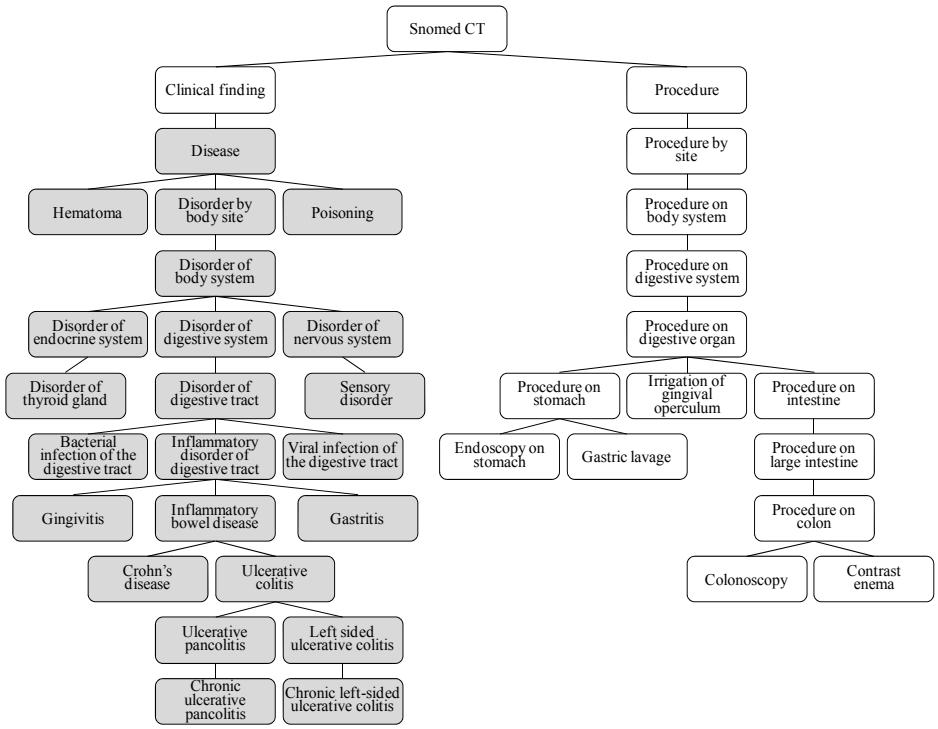


Figure 6-1 Example of taxonomy associated with the domain of the attribute *Disease* (gray-shaded concepts), extracted from the SNOMED-CT medical ontology.

On the other hand, a suitable semantic distance to be applied in a noise addition scenario should (i) be computationally efficient, due to the number of distance calculations that are needed during the noise addition process, (ii) provide values normalized in the range [0..1], where 0 represents the minimum distance, i.e., both concepts are the same, and 1 represents the maximum distance of concepts in the finite domain of the attribute within the ontology, and (iii) perform the calculation of the semantic distance consistent with the noise distribution.

For noise sequences that are normally distributed, $sd(.,.)$ should perform a non-logarithmic and non-exponential calculation, so that distances are well spread through the range [0..1]. In this way, it would be more likely to find

appropriate replacement concepts during the noise-addition process, i.e., concepts that are as semantically distant as defined by the noise magnitude. However, for other types of noise distributions, such as Laplace, which follows a symmetric exponential distribution, non-linear semantic distance measures that concentrate the distance mass either in the high or low output ranges would be more appropriate [63].

To determine the most appropriate distance measure to guide the masking process of noise addition, in the following, we analyze the different semantic similarity measures depicted in Section 4.3. Among the edge-counting measures, only [87] provide normalized results and performs a non-logarithmic and non-exponential assessment. On the other hand, feature-based measures [88, 90] estimate the similarity between concepts as a function of their common and non-common ontological features, such as taxonomic and non-taxonomic relationships. As evidenced in [89], many of these latter measures use non-taxonomic relationships, a form of knowledge partially modeled in most ontologies [84]. In this regard, [89] proposes a measure based on taxonomic features alone, but it is logarithmic. Finally, information content-based measures combine the taxonomic features of the evaluated concepts with their probability of occurrence in a given text corpus [91, 92]. However, due to their dependence on corpora, these measures present some issues: accuracy depending on the size and adequacy of the corpus, high computational cost and language ambiguity problems. Even though there are methods [94, 95] that intrinsically compute the concept specificity according to the number of hyponyms in the taxonomy, their calculation is logarithmic and counting hyponyms in large ontologies is costly.

According to the discussion above, for noise sequences following a normal distribution, we propose using a normalized edge-counting measure that is neither logarithmic nor exponential. Specifically, we use the well-known semantic similarity measure proposed by Wu and Palmer, sim_{wp}, [87] because it fulfills the above requirements and reasonably mimic human judgments on semantic similarity by estimating the specificity of concepts from their taxonomic depth [15, 89]. Because sim_{wp} (equation (4.6)) evaluates the similarity between concepts, we formulate sd_{wp} to compute the semantic distance, as the opposite of sim_{wp}:

$$sd_{wp}(c_1, c_2) = 1 - sim_{wp}(c_1, c_2) \tag{6.1}$$

Because the taxonomy considered to measure semantic similarities in a noise addition scenario is limited to $\tau(D(X^a))$, the concepts c_1, c_2, $LCS(c_1, c_2)$ of the equation (4.6) belong to $\tau(D(X^a))$ and the concept *root* of the taxonomy is the top node of $\tau(D(X^a))$. As an example, we show the calculation of the semantic distance sd_{wp} for two cases: when the concepts are different (*gingivitis* and *gastritis*) and when the concepts are the same (*gastritis* and *gastritis*). The distances have been calculated on the taxonomy associated with the domain *Disease* illustrated in Figure 6-1.

$$sd_{wp}(gingivitis, gastritis) = 1 - \frac{2 \times 6}{2 \times 6 + 1 + 1} = 0.14$$

$$sd_{wp}(gastritis, gastritis) = 1 - \frac{2 \times 7}{2 \times 7 + 0 + 0} = 0$$

6.3 Semantic uncorrelated noise addition method

As discussed in Section 6.2, uncorrelated noise addition needs (i) to calculate the variance of the input attribute to generate a noise sequence that reflects the degree of dispersion of the original values, and (ii) to compare values for assessing how far the noisy value must be from the original one according to the random magnitude of noise defined by the user. In the nominal domain, we propose using the semantic versions of the variance (Section 4.5) and the difference (with the measure selected in Section 6.2). Specifically, the distance measure discussed in Section 6.2 enables us to semantically compare nominal values while being consistent with a normal noise distribution. In order to control data distortion and maximize data utility, we define the following objectives for our method:

1. To provide a parameterized noise level.
2. To replace original values by noisy ones within a semantic distance consistent with the desired noise level.
3. To preserve the semantic mean of individual attributes as much as possible.

4. To obtain a dispersion proportional to the semantic variance of the original data and the desired noise level.

The steps to add noise to an attribute X^a through the uncorrelated noise addition method are shown in Figure 6-2. Firstly, the taxonomy $\tau(D(X^a))$ associated with the domain of attribute X^a is obtained from the ontology O, as described in Section 6.2. Thereby, we delimit the set of concepts from O that are candidates to replace the original values during the noise-addition process.

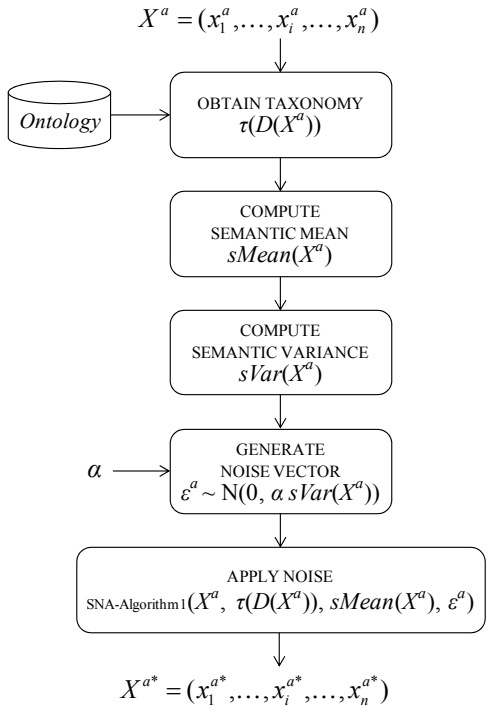


Figure 6-2. *Semantic uncorrelated noise addition* method for a nominal attribute X^a.

Secondly, in order to provide a user-settable noise and, therefore, to satisfy objective (1), it is necessary that the noise sequence added to original data has a configurable dispersion. Such as in the numerical uncorrelated noise addition method, this is defined by the parameter α, which allows customizing the error variance such that $Var(\varepsilon^a) = \alpha\ sVar(X^a)$. To compute the semantic variance

$sVar(X^a)$ of the attribute X^a, we use equation (4.19), which requires the calculation of the semantic mean $sMean(X^a)$ by equation (4.18) adapted to $\tau(D(X^a))$. After that, the noise sequence consisting of a vector of $n = |X^a|$ random numbers $\varepsilon^a = \{\epsilon_1^a, \ldots, \epsilon_n^a\}$, which follows a normal distribution $\varepsilon^a \sim N(0, \alpha\, sVar(X^a))$ with mean 0 and variance $\alpha\, sVar(X^a)$, is generated. Finally, after obtaining ε^a, the error values ϵ_i^a are applied to the original values x_i^a of attribute X^a. To apply the error values, it is necessary to provide an interpretation of the error magnitude and its sign, which helps achieving objectives (2) and (3) of the method, and therefore (4). In the following we describe in detail this interpretation.

To replace the original values by semantically-coherent noisy ones and, therefore, satisfy objective (2), it is necessary to interpret the error magnitude within the semantic domain. In the numerical domain, the noise represents the magnitude to be added to or subtracted from the input values. Therefore, the error values define the numerical distances between original values, x_i^a, and their noisy versions, x_i^{a*}. In the same way, in the semantic domain, error values should correspond to semantic distances. These distances are used to replace the original values by other concepts in the taxonomy associated with the domain that are as semantically distant as defined by the error magnitude, i.e., $sd(x_i^a, x_i^{a*}) = \left|\epsilon_i^a\right|$. However, because the semantic domain is discrete, it may happen that there is not a concept at the exact required distance. In such case, to fulfill the desired noise level, we propose selecting the concept that exceeds and best approximates the error magnitude.

$$x_i^{a*} = \underset{c \in \tau(D(X^a))}{\arg\min} \left\{ sd(c, x_i^a) \mid sd(c, x_i^a) \geq \left|\epsilon_i^a\right| \right\} \tag{6.2}$$

Regarding the preservation of the semantic mean required in objective (3), we must examine how the previous additions or subtractions influence this feature. In the numerical domain, if a positive error greater than the mean is added to an original value, the new value will be further away from the mean at the same magnitude. Otherwise, if the error is negative, the new value will be closer to the mean at the same magnitude. Because the noise sequence is normally distributed around zero, the magnitude of the accumulated additions

and subtractions with respect to the mean will compensate each other. Therefore, the mean of the noise-added values will be the same as the mean of the original values. In the semantic domain, it will be necessary to balance the number of movements towards and away from the *mean* concept. However, as discussed in Section 4.7, the semantic domain lacks a total order; that is, if we move away a certain distance from a concept, we cannot guarantee getting closer to or away from the mean concept at the same distance. Therefore, if we use the original values as reference points to apply the error values, but we do this uncontrollably, we will fulfill the expected absolute errors w.r.t. the original values, but we cannot ensure that the *semantic mean* will be preserved.

This problem can be solved by using the error sign to guide the replacement of values towards the preservation of the semantic mean. To do so, we propose balancing the number of movements towards and away from the mean by following a specific strategy:

- If the error ϵ_i^a is positive, the concept *c* in $\tau(D(X^a))$ that will replace the original value x_i^a must be farther from $sMean(X^a)$ than x_i^a, i.e., $sd(c, sMean(X^a)) > sd(x_i^a, sMean(X^a))$. For example, by applying this condition to the original nominal value x_i^a =*Ulcerative colitis* in Figure 6-3, we obtain several possible concepts for replacement, which are all further from the *mean* concept *Disorder of digestive tract* than *Ulcerative colitis*. These concepts constitute the set of replacement candidates of x_i^a.

- If the error ϵ_i^a is negative, the concept *c* in $\tau(D(X^a))$ that will replace the original value x_i^a must be closer to $sMean(X^a)$ than x_i^a, i.e., $sd(c, sMean(X^a)) < sd(x_i^a, sMean(X^a))$. By applying this condition to the previous example, the set of replacement candidates would comprise concepts that are closer to the *mean* concept *Disorder of digestive tract* than *Ulcerative colitis*, as shown Figure 6-4.

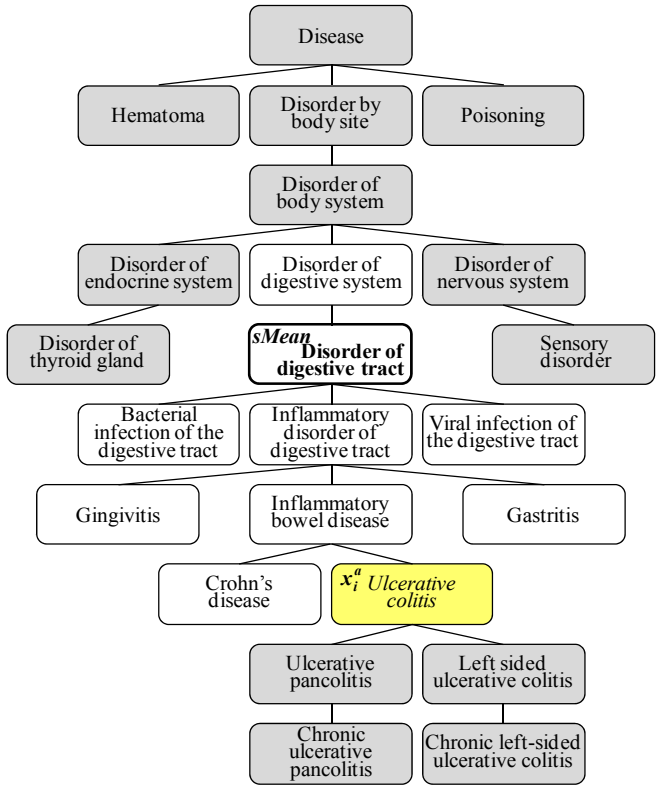


Figure 6-3. Example of replacement candidates (gray-shaded concepts) for an original value x_i^a (*Ulcerative colitis*) when the error sign is positive.

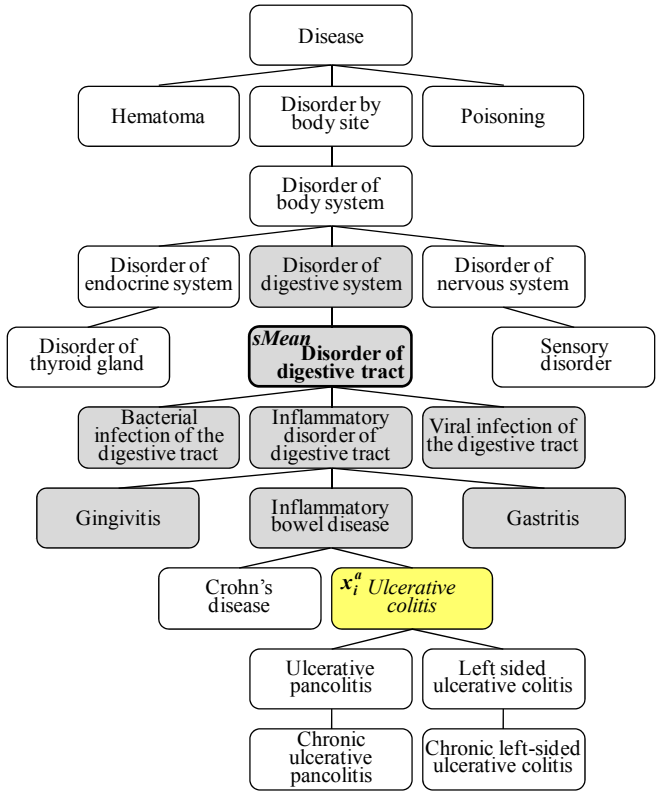


Figure 6-4. Example of replacement candidates (gray-shaded concepts) for an original value x_i^a (*Ulcerative colitis*) when the error sign is negative.

Finally, our method will select as noise-added value x_i^{a*} the candidate concept *c* that best approximates the error magnitude $|\epsilon_i^a|$ according to equation (6.2). Because the accumulated mass of positive and negative errors in the normal noise sequence should be equivalent, this strategy will tend to preserve the semantic mean.

Formally, the procedure that applies the noise vector to the original values of the attribute X^a is shown in SNA-Algorithm1. Together with X^a, $\tau(D(X^a))$ and the noise vector ε^a, $sMean(X^a)$ is passed as input parameter to the algorithm because it is necessary to balance value replacements with respect to the semantic mean of X^a. In order to select the noise-added values x_i^{a*}, we

apply the noise magnitude $\left|\epsilon_i^a\right|$ to each original value x_i^a by replacing it by a concept c in $\tau(D(X^a))$ that ideally matches the error magnitude or that, while exceeding the error magnitude, minimizes its distance from x_i^a (lines 7 and 9). At this step, the interpretation of the error sign proposed above is used: line 7 when ϵ_i^a is positive and line 9 when ϵ_i^a is negative; if ϵ_i^a is zero, the noise-added value x_i^{a*} is exactly x_i^a (line 3). Finally, when x_i^a matches $sMean(X^a)$, the noise-added value x_i^{a*} will simply be the concept c in $\tau(D(X^a))$ that ideally matches the error magnitude or that, while exceeding the error magnitude, minimizes its distance from x_i^a (line 5). In any case, if no concept c in $\tau(D(X^a))$ with $sd(c, x_i^a) \geq \left|\epsilon_i^a\right|$ exists, i.e., we cannot get further enough within $\tau(D(X^a))$, we select the concept that best approximates the condition. Because of this truncation and due to the need to discretize error values, the accuracy of the noise-added data will be limited by the size and granularity of the underlying ontology.

SNA-Algorithm1. Method to apply the noise vector to an attribute X^a by using the *mean* concept as reference point in the replacements.

Input :

X^a : nominal attribute with *n* records

$\tau(D(X^a))$: taxonomy associated with the domain of X^a

$sMean(X^a)$: semantic mean of X^a

ε^a: noise vector

Output :

X^{a*} : noise-added nominal attribute

1 : **for all** x_i^a in X^a **do**

2: **if** $\epsilon_i^a = 0$ **then**

3: $x_i^{a*} \leftarrow x_i^a$

4 : **else if** $x_i^a = sMean(X^a)$ **then**

5: $x_i^{a*} \leftarrow \underset{c \in \tau(D(X^a))}{\arg\min} \left\{ sd(c, x_i^a) \mid sd(c, x_i^a) \geq \left| \epsilon_i^a \right| \right\}$

6 : **else if** ϵ_i^a is positive **then**

7: $x_i^{a*} \leftarrow \underset{c \in \tau(D(X^a))}{\arg\min} \left\{ sd(c, x_i^a) \mid sd(c, x_i^a) \geq \left| \epsilon_i^a \right| \wedge sd\left(c, sMean(X^a)\right) > sd\left(x_i^a, sMean(X^a)\right) \right\}$

8 : **else if** ϵ_i^a is negative **then**

9: $x_i^{a*} \leftarrow \underset{c \in \tau(D(X^a))}{\arg\min} \left\{ sd(c, x_i^a) \mid sd(c, x_i^a) \geq \left| \epsilon_i^a \right| \wedge sd\left(c, sMean(X^a)\right) < sd\left(x_i^a, sMean(X^a)\right) \right\}$

10 : **end if**

11 : **end for**

12 : **return** X^{a*}

The computational cost of this algorithm for a single attribute with *n* records is O(*n*×*m*), where *m* is the number of concepts in the semantic domain.

It should be pointed out that, as it was stated in Section 2.3.2.1, to add noise to a multivariate dataset with *m* nominal attributes through uncorrelated noise, SNA-Algorithm1 must be applied to each attribute independently. Therefore, correlation among attributes will not be preserved.

6.4 Semantic correlated noise addition method

The correlated noise addition mechanism requires computing the covariance matrix to generate the noise sequences that reflect the degree of correlation between the attributes. By relying on the semantically-grounded versions of the distance covariance and distance variance measures proposed in Section 4.6, we can build the semantic covariance matrix of the input nominal dataset and adapt the numerical correlated noise addition method detailed in Section 2.3.2.1 to the semantic domain defined in Section 6.2. Because attribute covariances are now considered during the noise addition process, we will be able to preserve the semantic relationships between nominal attributes better than with uncorrelated noise.

In addition to the objectives of the uncorrelated method depicted in the previous section, our semantic correlated noise addition method has the following ones:

1. To obtain a data dispersion proportional to the covariance matrix of the original data and the noise magnitude.
2. To preserve the correlation between the attributes as much as possible.

For clarity, in the following description of the method, we assume that the dataset X has two nominal attributes X^a and X^b with n records.

As shown in Figure 6-5, the fundamental difference between this method and the uncorrelated one is the procedure employed to generate the noise sequence for each attribute. As in the previous method, it is necessary to generate noise sequences with a configurable dispersion level. However, the noise sequences have to reflect the degree of correlation between attributes. Only in this way can this method preserve the association between the attributes. For this reason, and according to Section 2.3.2.1, the generated noise consists of a ($n \times 2$) matrix of random numbers $\varepsilon^{a,b} = \{(\epsilon_1^a, \epsilon_1^b), ..., (\epsilon_n^a, \epsilon_n^a)\}$ that follows a multivariate normal distribution $\varepsilon^{a,b} \sim N(0, \alpha\Sigma_{X^a,X^b})$ with mean the vector 0 and covariance matrix $\alpha\Sigma_{X^a,X^b}$, where the parameter α determines the desired degree of semantic noise and Σ_{X^a,X^b} represents the semantic covariance matrix of the attributes. In the

semantic domain, Σ_{X^a,X^b} is a (2×2) matrix where the diagonal elements are the *semantic distance variances* of the attributes, and the off-diagonal elements are the *semantic distance covariances* between the attributes; both measures are obtained by using the equations (4.25) and (4.23) respectively.

$$\Sigma_{X^a,X^b} = \begin{pmatrix} sdVar(X^a) & sdCov(X^a,X^b) \\ sdCov(X^b,X^a) & sdVar(X^b) \end{pmatrix} \tag{6.3}$$

Finally, as shown in the last step depicted in Figure 6-5, the noise vectors ε^a and ε^b from $\varepsilon^{a,b}$ are applied to the attributes X^a and X^b, respectively. To do this, we propose three noise addition strategies, which are detailed below.

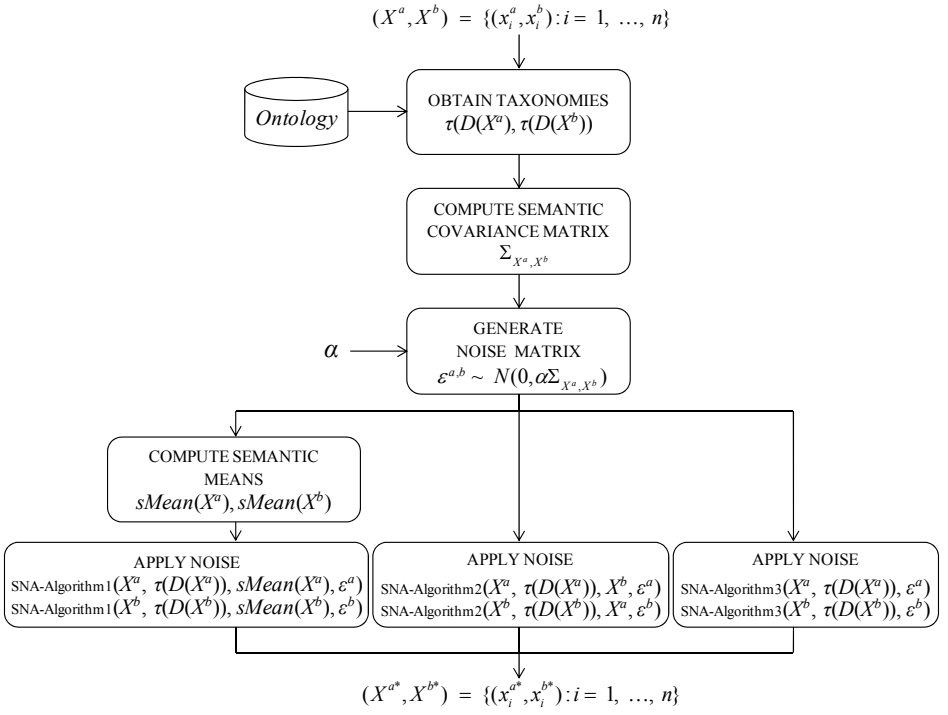


Figure 6-5. *Semantic correlated noise addition* method for two nominal attributes X^a and X^b.

The first approach follows the strategy detailed in SNA-Algorithm1: to balance value replacements with respect to the *semantic mean* of the attribute. As an alternative, to preserve the correlation between attributes regardless of the mean, we propose SNA-Algorithm2. The difference of this approach from the previous one lies in the reference point used to select the replacements in the noise addition process. In this regard, we must examine how, in the numerical domain, the additions and subtractions of error values on attribute X^a influence attribute X^b, and vice versa. Specifically, to preserve the correlation of the data (x_i^a, x_i^b), if a positive error is added to an original numerical value $x_i^a > x_i^b$, the noisy value will be further away from x_i^b at the same magnitude; on the other hand, if the error is negative, the new value will get closer to x_i^b. Because the noise sequences must reflect the degree of correlation of the attributes, the magnitude of the accumulated additions and subtractions between value pairs will compensate each other. Therefore, the correlation of the noise-added values will be the same as the correlation of the original values. Once more, because of the lack of a total order in the semantic domain, it will be necessary to balance the number of movements between value pairs. For this reason, we propose a new strategy that uses as reference point the value x_i^b corresponding to the value x_i^a that is being replaced, and vice versa.

Formally, as shown in SNA-Algorithm2, if the error ϵ_i^a is positive, the concept c in $\tau(D(X^a))$ that will replace the original value x_i^a must be farther from x_i^b than x_i^a, i.e., $sd(c, x_i^b) > sd(x_i^a, x_i^b)$, and vice versa. Otherwise, if the error ϵ_i^a is negative, the concept c must be closer to x_i^b than x_i^a, i.e., $sd(c, x_i^b) < sd(x_i^a, x_i^b)$, and vice versa. Understandably, both attributes must belong to the same semantic domain, i.e., $\tau(D(X^a)) = \tau(D(X^b))$. For each attribute SNA-Algorithm2 will be called instead of SNA-Algorithm1 in the last step depicted in Figure 6-5.

SNA-Algorithm2. Method to apply the noise vector to an attribute X^a by using the values of the attribute X^b as reference points in the replacements.

Input :

X^a, X^b : nominal attributes with *n* records

$\tau(D(X^a))$: taxonomy associated with the domain of X^a

ε^a: noise vector

Output :

X^{a*} : noise-added nominal attribute

1: **for all** x_i^a in X^a **do**

2: **if** $\epsilon_i^a = 0$ **then**

3: $x_i^{a*} \leftarrow x_i^a$

4: **else if** ϵ_i^a is positive **then**

5: $x_i^{a*} \leftarrow \underset{c \in \tau(D(X^a))}{\arg\min} \left\{ sd(c, x_i^a) \mid sd(c, x_i^a) \geq \left| \epsilon_i^a \right| \wedge sd(c, x_i^b) > sd(x_i^a, x_i^b) \right\}$

6: **else if** ϵ_i^a is negative **then**

7: $x_i^{a*} \leftarrow \underset{c \in \tau(D(X^a))}{\arg\min} \left\{ sd(c, x_i^a) \mid sd(c, x_i^a) \geq \left| \epsilon_i^a \right| \wedge sd(c, x_i^b) < sd(x_i^a, x_i^b) \right\}$

8: **end if**

9: **end for**

10: **return** X^{a*}

A solution for attributes belonging to different semantic domains, i.e., $\tau(D(X^a)) \neq \tau(D(X^b))$, for example, $\tau(D(X^a))$ = {*diseases*} and $\tau(D(X^b))$ = {*medical procedures*}, is to consider as reference point the most generic concept of the taxonomy, i.e., the *root* concept. In this sense, the *root* concept is seen as the gateway to other domains. This process is formally shown in SNA-Algorithm3.

SNA-Algorithm3. Method to apply the noise vector to an attribute X^a using the *root* concept of $\tau(D(X^a))$ as reference point in the replacements.

Input :

X^a : nominal attribute with *n* records

$\tau(D(X^a))$: taxonomy associated with the domain of X^a

ε^a: noise vector

Output :

X^{a*} : noise-added nominal attribute

1 : **for all** x_i^a in X^a **do**

2 : **if** $\epsilon_i^a = 0$ **then**

3: $x_i^{a*} \leftarrow x_i^a$

4 : **else if** ϵ_i^a is positive **then**

5: $x_i^{a*} \leftarrow \underset{c \in \tau(D(X^a))}{\arg\min} \left\{ sd(c, x_i^a) \mid sd(c, x_i^a) \geq \left| \epsilon_i^a \right| \wedge sd(c, root) > sd(x_i^a, root) \right\}$

6 : **else if** ϵ_i^a is negative **then**

7: $x_i^{a*} \leftarrow \underset{c \in \tau(D(X^a))}{\arg\min} \left\{ sd(c, x_i^a) \mid sd(c, x_i^a) \geq \left| \epsilon_i^a \right| \wedge sd(c, root) < sd(x_i^a, root) \right\}$

8 : **end if**

9 : **end for**

10 : **return** X^{a*}

Further, if the semantic domains of the attributes are modeled in different ontologies, we may need to adjust the semantic distance calculation whereby distance values obtained from different ontologies with different sizes and granularities can be fairly compared. In this respect, some authors [109-111] have recently proposed methods to consistently compute the semantic similarity across multiple ontologies.

Concerning the multivariate character of the correlated noise-addition method, it should be noted that the algorithms that use the *mean* or the *root* concept as reference points (SNA-Algorithm1 and SNA-Algorithm3) do not constrain the number of attributes they support. This is because the selection of replacements of an attribute in the noise addition process does not require taking into account the values of the remaining attributes: once the correlated noise sequences have been generated, they are applied to each attribute separately. On the other hand, SNA-Algorithm2 must be employed on disjoint

pairs of attributes because it uses the values of the second attribute as reference points to select the replacements of the first. For example, let $X = (X^a, X^b, X^c, X^d) = \{(x_i^a, x_i^b, x_i^c, x_i^d) : i = 1, \ldots, n\}$ be a dataset with four nominal attributes; there are three options to apply SNA-Algorithm2: pairs (X^a, X^b) and (X^c, X^d); pairs (X^a, X^c) and (X^b, X^d), and pairs (X^a, X^d) and (X^b, X^c). As a consequence, the correlation of the pairs would be preserved, but we cannot guarantee the same for the overall correlation of the dataset.

Also, notice that due to the discretizations and truncations of noise magnitudes inherent to the semantic domain, the accuracy of the noise-added outcomes of the three variations of the correlated method depends on the size and granularity of the underlying taxonomy, as it also happens for the uncorrelated method.

The computational cost of the three algorithms for two attributes with n records is O($n \times m$), where m is the number of concepts in the semantic domain.

6.5 Conclusion

We have presented in this chapter semantic solutions to noise addition with individual attributes (uncorrelated noise) and to multivariate datasets (correlated noise). In order to be able to deal with nominal data, and thereby distort data consistently with their semantics, we have used the semantic versions of operators used in the standard noise addition mechanism, which we defined in Chapter 4. In addition, several strategies have been proposed to guide the replacement of values during the noise addition process towards the preservation of either the semantic mean or the semantic distance correlation.

As a summary and guide for practitioners and researchers, Table 6.1 shows which of our methods is best suited to distort nominal data according to the type of dataset and the analytical utility requirements, that is, the semantic feature whose preservation should be optimized.

Table 6.1. Best suited methods according to the type of dataset and semantic feature to be optimized.

Dataset	**Optimized feature**	**Suggested method**
One attribute	*sMean*	Uncorrelated- SNA-Algorithm1
Two attributes with the same taxonomy	*sMean*	Correlated- SNA-Algorithm1
	sdCor	Correlated- SNA-Algorithm2
Two attributes with different taxonomies	*sMean*	Correlated- SNA-Algorithm1
	sdCor	Correlated- SNA-Algorithm3
More than two attributes	*sMean*	Correlated- SNA-Algorithm1
	sdCor	Correlated- SNA-Algorithm3

Chapter 7 **Empirical study**

In this section, we evaluate the semantic methods we propose in Chapter 5 and Chapter 6 with several nominal datasets and w.r.t. different evaluation metrics. The metrics we use quantify the data utility preserved in the outcomes from a semantic perspective. As baselines, we compare the results provided with our methods with those obtained by non-semantic data perturbation methods that rely on the distribution of the data.

7.1 Evaluation data

As evaluation data, we used a structured database containing patient discharge data provided by the California Office of Statewide Health Planning and Development (OSHPD), which were collected from licensed hospitals in California in 2009 [1]. Each record of the database details the healthcare discharge of a patient and, among others, it contains several nominal attributes stating the *principal diagnosis*, the *secondary diagnosis* and the *medical procedure* applied to the patient, which we selected to evaluate our methods.

As discussed in Chapter 2, discharge patient data are of highly sensitive nature and different regulations, such as the Health Insurance Portability and Accountability Act (HIPAA) [22] in the United States or the General Data Protection Regulation (GDPR) [19] in the European Union, state the need to protect them. Consequently, appropriate data protection measures should be undertaken by the data controller before making these data available for secondary use. In this context, the database provided by the OSHPD is especially suitable to illustrate the need for semantic privacy-preserving methods because most patient discharge data compiled are nominal. In addition, because correlations between medical attributes, such as *principal diagnosis* and *medical procedure*, are crucial for research, the OSHPD

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

database allows us to evaluate how well the semantic relationship between the attributes is preserved after the masking process.

To evaluate our methods, we have created several datasets from multiple samples of the OSHPD database with different correlation degrees, which we detail in the following sections.

7.2 Underlying ontology used in the study

Diagnoses and procedure codes in the OSHPD database have been mapped to healthcare concepts in the SNOMED-CT[2] medical ontology [81], which is especially well-suited to assist semantic similarity assessments of medical-related data because of its large size and fine grained taxonomic detail.

SNOMED-CT (Systematized Nomenclature of Medicine, Clinical Terms) is a domain ontology developed for medical purposes. It 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, computerized physician order entry, disease surveillance, image indexing and consumer health information services.

The medical terms (e.g., *flu*) in SNOMED-CT have unique meanings, in contrast with other ontologies of polysemic character (e.g., WordNet [79]). These terms are grouped into sets of synonyms (*synset*). A synset is thus a set of terms that are interchangeable in some context, because they share a commonly-agreed upon meaning without variation (e.g., {*flu*, *influenza*, *grippe*}). Each synset represents a distinct concept in SNOMED-CT which is identified by a code SCTID (e.g., the synset {*flu*, *influenza*, *grippe*} is identified by the SCTID 6142004).

SNOMED-CT contains more than 311,000 concepts with formal logic-based definitions organized 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 in SNOMED-CT may belong to one or more of these hierarchies by multiple

[2] https://www.nlm.nih.gov/healthit/snomedct/index.html

inheritance 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 [15, 112].

7.3 Evaluation metrics

To quantify the data utility preserved in the outcomes from a semantic perspective, we have considered the following semantic features and evaluation metrics:

1. The *semantic mean* of the original attribute X^a, $sMean(X^a)$, and of the permuted attribute X^{a^*}, $sMean(X^{a^*})$ defined in Section 4.4 and the *semantic distance* between both, $sd(sMean(X^{a^*}), sMean(X^a))$. A distance value of 0 indicates that the mean has been perfectly preserved after the swapping process.
2. To evaluate the semantic dispersion of permuted attributes, we use the *semantic distance variance* defined in Section 4.6, and the absolute difference between the variances of the original and permuted attributes, $sdVar(X^a)$ and $sdVar(X^{a^*})$. A difference of 0 indicates that the variance has been perfectly preserved after the swapping process. On the other hand, to evaluate the semantic dispersion of noise-added attributes and according to Section 2.3.2.1, we use the *semantic variance* defined in Section 4.5 and the absolute difference between the actual *semantic variance* of the noise-added attribute values and the expected *semantic variance* after adding noise with a noise parameter α, i.e., $|sVar(X^{a^*}) - (1+ \alpha)\ sVar(X^a)|$. Differences near 0 indicate that the variance of the noise-added results has been well-controlled.
3. To measure the overall loss of information in terms of semantics of the masked attributes, we use the *root mean square error* (*RMSE*). To evaluate loss of semantics in permuted attributes, we measure the root average square semantic distance between original and permuted value pairs, $RMSE(X^a, X^{a^*})$. To evaluate loss of semantics in noise-added attributes, we measure the RMSE between original and noise-added value pairs, $RMSE_{Actual}(X^a, X^{a^*})$, w.r.t. the target error defined by the

desired magnitude of the noise to be added, $RMSE_{Target} = \varepsilon^a$. In any case, small values indicate low information loss, and thus, perturbed data with better quality.

4. The *semantic distance correlation* of original and perturbed attribute pairs, $sdCor(X^a, X^b)$ and $sdCor(X^{a*}, X^{b*})$, by using equation (4.24), and the absolute difference between the actual *semantic distance correlation* of pairs of perturbed attributes and original attributes, i.e., $|sdCor(X^{a*}, X^{b*}) - sdCor(X^a, X^b)|$. Difference=0 indicates that the correlation has been perfectly preserved after the masking process.

7.4 Evaluation of semantic rank swapping

To evaluate our semantic rank swapping methods defined in Chapter 5, we have used two samples from the OSHPD database with different correlation degree.

The first experiment has been carried out with a sample of 1,172 patients and two moderately correlated attributes (X^a=*principal diagnosis* and X^b=*medical procedure*) belonging to different semantic domains: attribute X^a belongs to the taxonomy of *diseases* and X^b belongs to the taxonomy of *procedures*, both from SNOMED-CT. The sample of the attribute X^a contains 783 different categories with an average of 1.5 records per category and the sample of the attribute X^b contains 430 with an average of 2.7. The semantic features of this sample, which we name *Dataset1*, are depicted Table 7.1.

Table 7.1. Semantic features of *Dataset1*: 1,172 patients with two moderately correlated attributes, X^a = *principal diagnosis,* X^b =*medical procedure*

Semantic feature	Value
$sMean(X^a)$	*Acute appendicitis with peritoneal abscess*
$sMean(X^b)$	*Endoscopic division of adhesions of peritoneum*
$sdVar(X^a)$	0.1148
$sdVar(X^b)$	0.1240
$sdCor(X^a, X^b)$	0.4595

We have evaluated the two versions of the univariate method (SRS-Algorithm1 and SRS-Algorithm2, Section 5.3) and the multivariate method (SRS-Algorithm3, Section 5.4). As semantic distance $sd(\cdot,\cdot)$, we have used the inverse of the well-known Wu and Palmer semantic similarity measure depicted in Section 4.3.

Table 7.2, Table 7.3 and Table 7.4 depict the semantic features and evaluation metrics of the results provided by SRS-Algorithm1, SRS-Algorithm2 and SRS-Algorithm3, respectively, for several values of the input parameter $k=\{2, 5,10, 20, 50, 100\}$.

Table 7.2. *Dataset1*: evaluation metrics of rank-swapped attributes values (X^a = *principal diagnosis,* X^b = *medical procedure*) with the univariate method: SRS-Algorithm1.

Metric	***k*=2**	***k*=5**	***k*=10**	***k*=20**	***k*=50**	***k*=100**
$sMean(X^a)$ | $sd(sMean(X^{a^*}), sMean(X^a))$	*Acute appendicitis with peritoneal abscess* | 0					
$sMean(X^b)$ | $sd(sMean(X^{b^*}), sMean(X^b))$	*Endoscopic division of adhesions of peritoneum* | 0					
$sdVar(X^a)$ | $\lvert sdVar(X^{a^*}) - sdVar(X^a)\rvert$	0.1148 | 0					
$sdVar(X^b)$ | $\lvert sdVar(X^{b^*}) - sdVar(X^b)\rvert$	0.1240 | 0					
$RMSE(X^a, X^{a^*})$	0.4558	0.5423	0.5560	0.5823	0.6018	0.6083
$RMSE(X^b, X^{b^*})$	0.3562	0.3712	0.4118	0.4166	0.4650	0.5016
$sdCor(X^{a^*}, X^{b^*})$	0.2705	0.2689	0.2649	0.2647	0.2628	0.2520
$\lvert sdCor(X^{a^*}, X^{b^*}) - sdCor(X^a, X^b)\rvert$	0.1890	0.1906	0.1946	0.1948	0.1967	0.2075

Table 7.3. *Dataset1*: evaluation metrics of rank-swapped attributes values (X^a = *principal diagnosis,* X^b = *medical procedure*) with the univariate method: SRS-Algorithm2.

Metric	***k*=2**	***k*=5**	***k*=10**	***k*=20**	***k*=50**	***k*=100**
$sMean(X^a)$ | $sd(sMean(X^{a^*}), sMean(X^a))$	*Acute appendicitis with peritoneal abscess* | 0					
$sMean(X^b)$ | $sd(sMean(X^{b^*}), sMean(X^b))$	*Endoscopic division of adhesions of peritoneum* | 0					
$sdVar(X^a)$ | $\lvert sdVar(X^{a^*}) - sdVar(X^a)\rvert$	0.1148 | 0					
$sdVar(X^b)$ | $\lvert sdVar(X^{b^*}) - sdVar(X^b)\rvert$	0.1240 | 0					
$RMSE(X^a, X^{a^*})$	0.1439	0.1887	0.2300	0.2782	0.3513	0.4062
$RMSE(X^b, X^{b^*})$	0.0544	0.0776	0.1014	0.1413	0.2153	0.2941
$sdCor(X^{a^*}, X^{b^*})$	0.4191	0.4092	0.4039	0.3648	0.3235	0.2800
$\lvert sdCor(X^{a^*}, X^{b^*}) - sdCor(X^a, X^b)\rvert$	0.0404	0.0503	0.0556	0.0947	0.1360	0.1795

Table 7.4. *Dataset1*: evaluation metrics of rank-swapped attributes values (X^a = *principal diagnosis,* X^b = *medical procedure*) with the multivariate method: SRS-Algorithm3.

Metric	*k*=2	*k*=5	*k*=10	*k*=20	*k*=50	*k*=100
$sMean(X^a) \mid sd(sMean(X^{a^*}), sMean(X^a))$	*Acute appendicitis with peritoneal abscess* | 0					
$sMean(X^b) \mid sd(sMean(X^{b^*}), sMean(X^b))$	*Endoscopic division of adhesions of peritoneum* | 0					
$sdVar(X^a) \mid \lvert sdVar(X^{a^*}) - sdVar(X^a)\rvert$	0.1148 | 0					
$sdVar(X^b) \mid \lvert sdVar(X^{b^*}) - sdVar(X^b)\rvert$	0.1240 | 0					
$RMSE\ (X^a,\ X^{a^*})$	0.1966	0.2707	0.3130	0.3659	0.4145	0.4656
$RMSE\ (X^b,\ X^{b^*})$	0.0920	0.1433	0.1613	0.2083	0.2697	0.3745
$sdCor(X^{a^*}, X^{b^*})$	0.4567	0.4410	0.4363	0.4160	0.3826	0.3145
$\lvert sdCor(X^{a^*}, X^{b^*}) - sdCor(X^a, X^b)\rvert$	0.0028	0.0185	0.0232	0.0435	0.0769	0.1450

First, we can see the semantic mean and the semantic variance are perfectly preserved for all the attributes. Because attribute values in the permuted outcome are the same as those in the original dataset but swapped, by definition, marginal statistics (e.g., mean, variance, min/max values) are perfectly preserved for each individual attribute. This contrast with other data protection mechanisms discussed in Chapter 2, such as data microaggregation or generalization, which protect data by making them more homogenous and, thus, reduce the variance and/or granularity of the original data.

Second, because the *k* parameter determines the swapping range, the larger the *k*, the larger the permutation and, thus, the larger the RMSE. Specifically, RMSEs show that our methods are able to proportionally distort the outcomes according to the desired level of protection. In addition, SRS-Algorithm2 and SRS-Algorithm3 provide significantly better values for the RMSE than SRS-Algorithm1. The differences observed between the former and the later measure the positive influence of the *dynamic clustering at opposite ends* strategy detailed in Section 5.3, which contributes to minimize the information loss resulting from the permutation process by i) limiting the swapping range of the original value/tuple to the *k* semantically-closest values/tuples in the dataset, and ii) prioritizing the permutation of those values whose swaps entail more information loss. SRS-Algorithm2 provides slightly better RMSEs for individual attributes than SRS-Algorithm3 because the former is able to optimize the swapping ranges for individual attributes, whereas the latter does it for complete tuples, which is likely suboptimal for individual attributes.

Finally, as expected, correlations between attributes are preserved by the multivariate method significantly better than by any of the univariate

algorithms, because the former constraints value swapping towards maintaining the dependence between the attributes. Regarding univariate methods, we can see that SRS-Algorithm2 reasonably preserves correlations because, for two correlated attributes, it is reasonable to assume that, if a value of the first attribute is closely related to another value of the second attribute, values semantically similar to the former one (resulting from the swap) will also be related to values semantically similar to the latter. This behavior also explains why SRS-Algorithm1 provides such poor correlation results, especially for low values of *k*: correlation differences are proportional to the also large RMSEs. So, we can conclude that SRS-Algorithm2 is still valid when maintaining attribute correlations is not priority or when dealing with non-dependent attributes, because it is able to minimize per-attribute errors better than SRS-Algorithm3.

To contextualize the results of our methods against those of related works, in Figures Figure 7-1, Figure 7-2 and Figure 7-3, we compare the outcomes of our algorithms with those provided by a non-semantic *data swapping* mechanism, which, due to its impossibility of using a total order on the nominal attributes to define the swapping ranges, randomly and unconstrainedly swaps values within attribute domains.

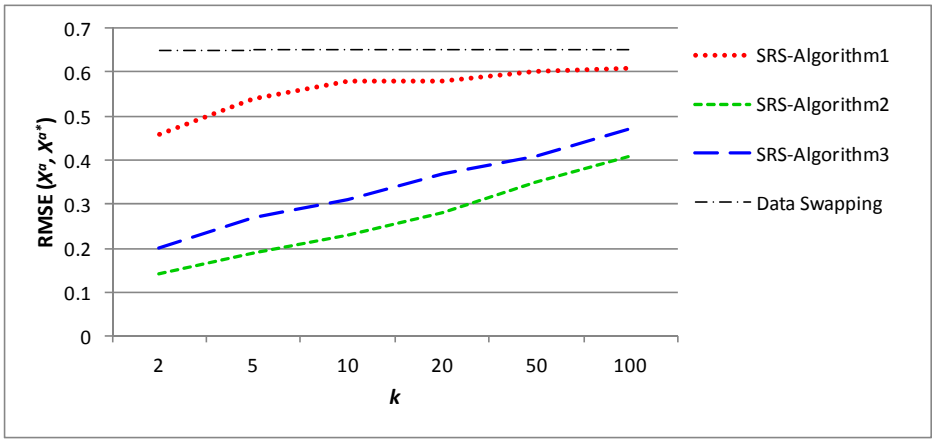


Figure 7-1. RMSE of attribute X^a with *data swapping* and *semantic rank swapping* (SRS-Algorithm1, SRS-Algorithm2 and SRS-Algorithm3) with *Dataset1*.

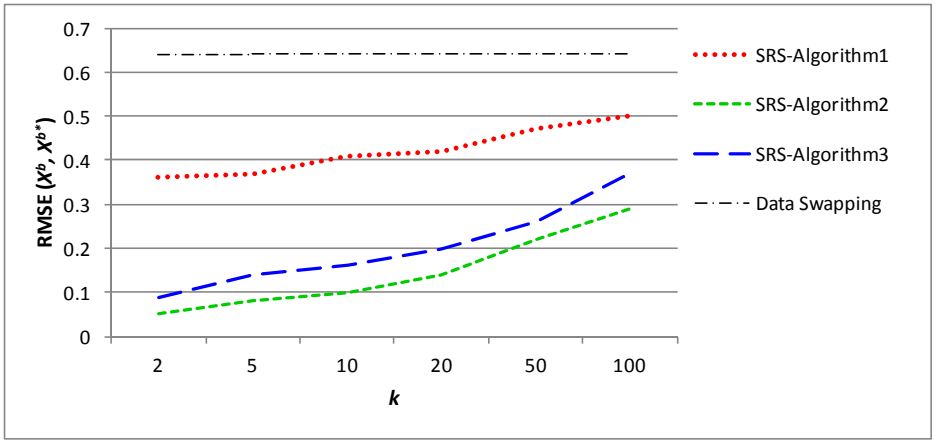


Figure 7-2. RMSE of attribute X^b with *data swapping* and *semantic rank swapping* (SRS-Algorithm1, SRS-Algorithm2 and SRS-Algorithm3) with *Dataset1*.

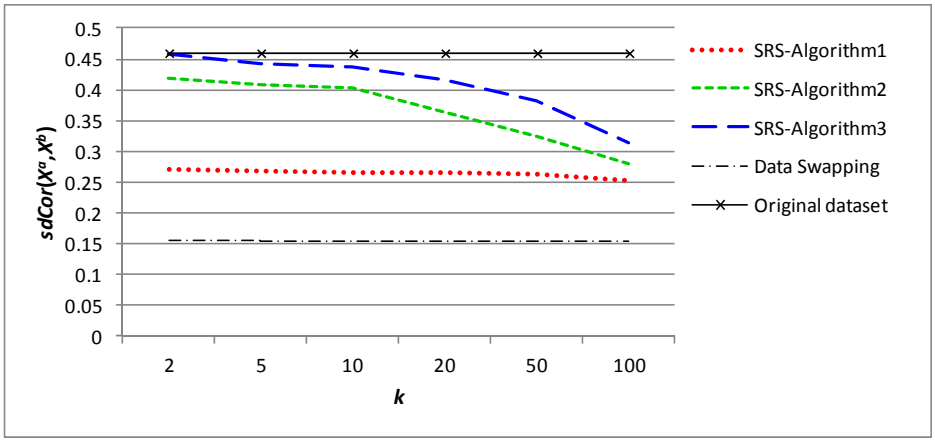


Figure 7-3. Semantic distance correlation for *data swapping* and *semantic rank swapping* (SRS-Algorithm1, SRS-Algorithm2 and SRS-Algorithm3) with *Dataset1*.

In comparison to the non-semantic *data swapping*, we can see that our semantic methods, especially SRS-Algorithm2 and SRS-Algorithm3, drastically improve the RMSEs and attribute correlations. Specifically, *data swapping* produces a significantly larger permutation that is also non-configurable and, on the other hand, largely breaks the correlation between attributes.

To test the generality of our multivariate method (SRS-Algorithm3), a second experiment has been carried out with another dataset of 1,012 patients, named *Dataset2*. In this case, the attributes X^a =*principal diagnosis* and X^b=*procedure* present a stronger correlation than in *Dataset1*, $sdCor(X^a, X^b)$ =0.6392. This stronger correlation implies that records are more homogenous and frequencies of attribute categories are higher: X^a = 55 different categories (average of 18.4 records per category) and X^b = 94 different categories (average of 10.8 records per category). Table 7.5 depicts the RMSEs and semantic correlation metrics of the results provided by SRS-Algorithm3.

Table 7.5. *Dataset2*: evaluation metrics of rank-swapped attributes values (X^a = *principal diagnosis,* X^b = *medical procedure*) with the SRS-Algorithm3.

Metric	***k*=2**	***k*=5**	***k*=10**	***k*=20**	***k*=50**	***k*=100**
$RMSE\ (X^a,\ X^{a*})$	0	0.0129	0.0336	0.0749	0.1965	0.3750
$RMSE\ (X^b,\ X^{b*})$	0	0.0288	0.0810	0.1486	0.2300	0.3218
$sdCor(X^{a*},X^{b*})$	0.6392	0.6383	0.6371	0.6238	0.5765	0.4658
$\lvert sdCor(X^{a*},X^{b*}) - sdCor(X^a,X^b)\rvert$	0	0.0009	0.0021	0.0154	0.0627	0.1734

As we can see, for k=2, the RMSE of both attributes is zero, which means that the swaps of attribute values have not resulted in values different from the original ones. This behavior is consistent with the frequency distribution of both attributes in *Dataset2:* because the cardinality of attribute values, and also of tuples of the two attributes, is larger than 2, values within the swapping range are equal. From the perspective of k-anonymity, this means that the original dataset was already indistinguishable for sets of k=2 records; that is, it is already *probabilistically-2-anonymous* and, also, *2-anonymous*. In this case, for k=2, the original dataset does not need to be modified to achieve the desired level of protection, and so does our algorithm, which enforces *probabilistic k-anonymity*; otherwise, unnecessary information loss would occur.

For k=5 or 10, which are still below the average frequency of attribute categories, we obtain very small (albeit not null) errors, whereas for $k \geq 20$, which exceed the average frequencies, differences are more noticeable. In all

cases, the semantic features of the data (in particular the strong attribute correlation) are preserved proportionally to the desired permutation level.

7.5 Evaluation of semantic noise addition

In this section, we evaluate the semantic noise addition methods we have proposed in Chapter 6. For that, we use two samples from the OSHPD database with different correlation degrees.

The first experiment was carried out with a pair of strongly correlated attributes X^a =*principal diagnosis* and X^b=*secondary diagnosis*, both with the same associated taxonomy, that is, the hierarchy of *diseases* of SNOMED-CT. Specifically, we have taken a sample of 1,350 patients, named *Dataset1*, whose semantic features are depicted in Table 7.6.

Table 7.6. Semantic features of *Dataset1*: 1,350 patients with two strongly correlated attributes X^a =*principal diagnosis* and X^b =*secondary diagnosis*, both with the same associated taxonomy

Semantic feature	Value
$sMean(X^a)$	Furuncle of chest wall
$sMean(X^b)$	Viral hepatitis with hepatic coma
$sVar(X^a)$	0.22
$sVar(X^b)$	0.24
$sdCor(X^a, X^b)$	0.94

The fact that the attributes are strongly correlated, $sdCor(X^a, X^b)$ = 0.94, allows us to study the behavior of our methods in the most challenging scenario: when a strong correlation should be preserved. Specifically, we have tested the uncorrelated method discussed in Section 6.3 with the noise-addition strategy defined in SNA-Algorithm1 (Uncorrelated-SNA-Algorithm1), the correlated method discussed in Section 6.4 with the noise-addition strategy defined in SNA-Algorithm1 (Correlated-SNA-Algorithm1) and the correlated method with the noise-addition strategy designed to optimize the preservation of the correlation between attributes defined in SNA-Algorithm2 (Correlated-SNA-Algorithm2), since both attributes are drawn from the same taxonomy.

Table 7.7, 7.8 and 7.9 collect the evaluation metrics of the results provided by these methods for several values of the noise parameter α={0.1, 0.3, 0.5, 1}.

Table 7.7. Evaluation metrics of the noise-added dataset obtained with Uncorrelated-SNA-Algorithm1 for *Dataset1* (X^a = *principal diagnosis* and X^b =*secondary diagnosis*).

Metric	**α=0.1**	**α=0.3**	**α=0.5**	**α=1**
$sMean(X^{a^*})$	Blister of axilla with infection	Axillary hydra-denitis suppurativa	Blister of axilla with infection	Blister of axilla with infection
$sd(sMean(X^{a^*}), sMean(X^a))$	0.20	0.20	0.20	0.20
$sVar(X^{a^*}) \mid \lvert sVar(X^{a^*}) - (1+\alpha)\, sVar(X^a)\rvert$	0.24 | 0	0.24 | 0.05	0.25 | 0.08	0.27 | 0.17
$RMSE_{Actual}(X^a, X^{a^*}) \mid RMSE_{Target} = \varepsilon^a$	0.23 | 0.15	0.31 | 0.25	0.38 | 0.34	0.49 | 0.48
$sMean(X^{b^*})$	Viral hepatitis with hepatic coma	Viral hepatitis with hepatic coma	Viral hepatitis with hepatic coma	Inflammatory disease of liver
$sd(sMean(X^{b^*}), sMean(X^b))$	0	0	0	0.18
$sVar(X^{b^*}) \mid \lvert sVar(X^{b^*}) - (1+\alpha)\, sVar(X^b)\rvert$	0.23 | 0.03	0.24 | 0.07	0.26 | 0.1	0.30 | 0.18
$RMSE_{Actual}(X^b, X^{b^*}) \mid RMSE_{Target} = \varepsilon^b$	0.19 | 0.15	0.30 | 0.27	0.37 | 0.35	0.50 | 0.51
$sdCor(X^{a^*}, X^{b^*}) \mid \lvert sdCor(X^{a^*}, X^{b^*}) - sdCor(X^a, X^b)\rvert$	0.78 | 0.16	0.65 | 0.29	0.57 | 0.37	0.38 | 0.56

Table 7.8. Evaluation metrics of the noise-added dataset obtained with Correlated-SNA-Algorithm1 for *Dataset1* (X^a = *principal diagnosis* and X^b =*secondary diagnosis*).

Metric	**α=0.1**	**α=0.3**	**α=0.5**	**α=1**
$sMean(X^{a^*})$	Blister of axilla with infection	Blister of axilla with infection	Blister of axilla with infection	Blister of axilla with infection
$sd(sMean(X^{a^*}), sMean(X^a))$	0.20	0.20	0.20	0.20
$sVar(X^{a^*}) \mid \lvert sVar(X^{a^*}) - (1+\alpha)\, sVar(X^a)\rvert$	0.24 | 0	0.24 | 0.05	0.26 | 0.07	0.29 | 0.15
$RMSE_{Actual}(X^a, X^{a^*}) \mid RMSE_{Target} = \varepsilon^a$	0.24 | 0.17	0.33 | 0.28	0.40 | 0.37	0.51 | 0.52
$sMean(X^{b^*})$	Viral hepatitis with hepatic coma	Inflammatory disease of liver	Mouth-gen. ulcers inflam. cartil. synd.	Mouth-gen. ulcers inflam. cartil. synd.
$sd(sMean(X^{b^*}), sMean(X^b))$	0	0.18	0.45	0.45
$sVar(X^{b^*}) \mid \lvert sVar(X^{b^*}) - (1+\alpha)\, sVar(X^b)\rvert$	0.22 | 0.04	0.23 | 0.08	0.24 | 0.12	0.28 | 0.2
$RMSE_{Actual}(X^b, X^{b^*}) \mid RMSE_{Target} = \varepsilon^b$	0.20 | 0.17	0.30 | 0.28	0.38 | 0.37	0.49 | 0.52
$sdCor(X^{a^*}, X^{b^*}) \mid \lvert sdCor(X^{a^*}, X^{b^*}) - sdCor(X^a, X^b)\rvert$	0.82 | 0.12	0.73 | 0.21	0.69 | 0.25	0.59 | 0.35

Table 7.9. Evaluation metrics of the noise-added dataset obtained with Correlated-SNA-Algorithm2 for *Dataset1* (X^a = *principal diagnosis* and X^b =*secondary diagnosis*).

Metric	α=0.1	α=0.3	α=0.5	α=1
$sMean(X^{a^*})$	Blister of axilla with infection	Blister of axilla with infection	Blister of axilla with infection	Granuloma inguinale
$sd(sMean(X^{a^*}), sMean(X^a))$	0.20	0.20	0.20	0.20
$sVar(X^{a^*}) \mid \lvert sVar(X^{a^*}) - (1+\alpha)\, sVar(X^a)\rvert$	0.24 | 0	0.26 | 0.03	0.28 | 0.05	0.32 | 0.12
$RMSE_{Actual}(X^a, X^{a^*}) \mid RMSE_{Target} = \varepsilon^a$	0.25 | 0.17	0.35 | 0.28	0.41 | 0.37	0.52 | 0.52
$sMean(X^{b^*})$	Inflammatory disease of liver	Inflammatory disease of liver	Inflammatory disease of liver	Mouth-gen. ulcers inflam. cartil. synd.
$sd(sMean(X^{b^*}), sMean(X^b))$	0.18	0.18	0.18	0.45
$sVar(X^{b^*}) \mid \lvert sVar(X^{b^*}) - (1+\alpha)\, sVar(X^b)\rvert$	0.23 | 0.03	0.25 | 0.06	0.27 | 0.09	0.31 | 0.17
$RMSE_{Actual}(X^b, X^{b^*}) \mid RMSE_{Target} = \varepsilon^b$	0.22 | 0.17	0.32 | 0.28	0.40 | 0.37	0.51 | 0.52
$sdCor(X^{a^*}, X^{b^*}) \mid \lvert sdCor(X^{a^*}, X^{b^*}) - sdCor(X^a, X^b)\rvert$	0.84 | 0.10	0.76 | 0.18	0.71 | 0.23	0.59 | 0.35

Evaluation metrics show that, since α determines the amount of applied noise, the greater the α, the greater the RMSE and, therefore, the distortion applied to the data. The actual RMSEs show that our methods are able to appropriately adapt the data distortion to the desired magnitude of noise; that is, the actual RMSE is greater than or equal to the target RMSE in all cases except for those with a very large noise parameter (α =1). Actual and target errors are not expected to be equal with nominal data due to the need to discretize error values, and because of the limited scope offered by the underlying taxonomy. In the first case, the small differences between actual and target RMSEs are caused by the need to discretize noise-added values to concepts in the taxonomy; this difference tends to be greater for small values of α because, when the error components ϵ_{a_i} and ϵ_{b_i} are small, the relative effect of the discretization is more noticeable over the absolute magnitude. In the second case, when the error components ϵ_{a_i} and ϵ_{b_i} are very large (α =1), the number of truncated error values increases due to the limited scope of the taxonomy, i.e., cases in which there is no concept in the taxonomy that meets or exceeds the error magnitude. When this happens, the actual RMSE may be smaller than the target RMSE.

We can also see that the *sMean* of the noise-added datasets is largely preserved, particularly if the noise level is small. This shows the effectiveness of the strategies we propose to guide the replacement process, which tends to

balance the distances of the replaced values with respect to the mean concept. On the other hand, the difference between the variance of the noise-added attribute and the expected variance is maintained below 25% of the parameter α. Such as for the RMSE, for nominal data it would be difficult to achieve a null difference because of the discretizations and truncated noisy values.

Finally, as expected, the correlation between attributes is better preserved by the correlated methods, especially for large values of α. Therefore, the uncorrelated method is well-suited when preserving the correlation is not crucial. Regarding correlated methods, we can see that, despite using the same noise sequences, Correlated-SNA-Algorithm1 provides a slightly better mean than Correlated-SNA-Algorithm2, because Correlated-SNA-Algorithm1 has been designed to optimize this feature. On the contrary, the correlation is better preserved by Correlated-SNA-Algorithm2 because it guides the replacement process towards optimizing the dependence between the attributes.

To compare our results against baseline methods representative of the data distortion strategies of related works discussed in Section 3.1, in Figure 7-4, Figure 7-5 and Figure 7-6, we compare the accuracy of our algorithms w.r.t. two non-semantic data distortion methods based on distributions:

- A *naïve distortion*, in which original values are randomly replaced by other values of the same dataset.

- A *probabilistic distortion*, in which the probability of selecting a value as replacement corresponds to the occurrence frequency of that value in the input dataset. Because the distribution of the data is considered during the distortion process, the outcome will preserve the statistical features of the data better than with the naïve method.

In contrast to distortion methods based on the distribution of the data, we can see that our methods dramatically improve two evaluation metrics: RMSEs and correlations. On the one hand, the former methods tend to add a significantly greater amount of noise, which is also non-configurable and, on the other hand, they totally break the correlation between attributes. Our methods provide better results even when the error magnitude is set to the maximum reasonable value (α=1), that is, an α value that tries to match the degree of distortion added by the methods based on the distribution of the data.

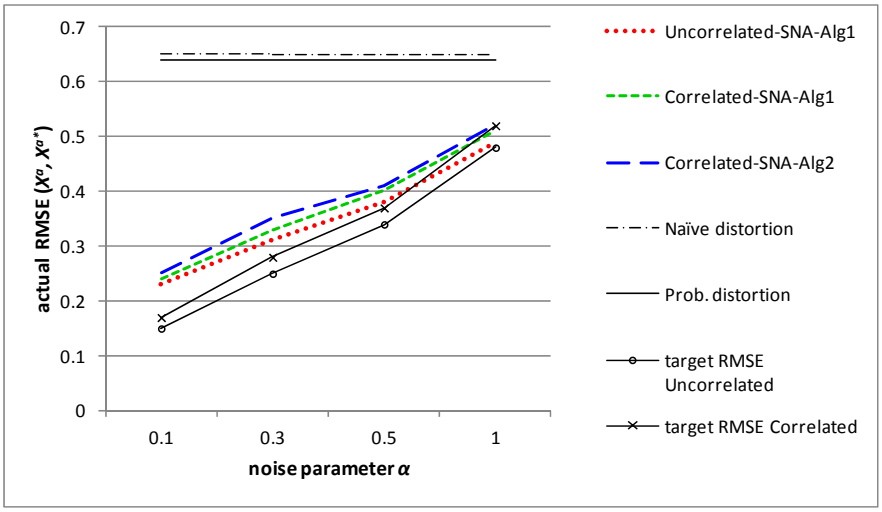


Figure 7-4. Evaluation of the actual RMSE of attribute X^a for the *naïve distortion*, *probabilistic distortion* and our semantic methods (Uncorrelated-SNA-Algorithm1, Correlated-SNA-Algorithm1 and Correlated-SNA-Algorithm2) in *Dataset1*.

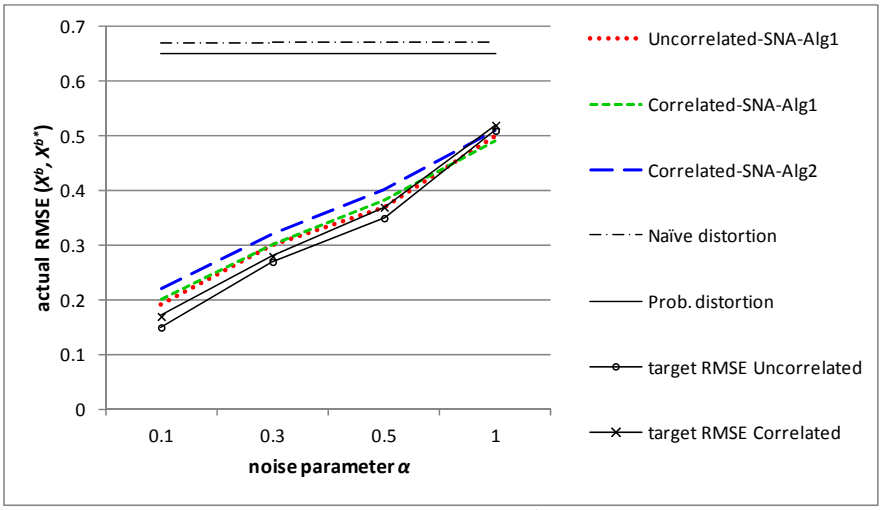


Figure 7-5. Evaluation of the actual RMSE of attribute X^b for the *naïve distortion*, *probabilistic distortion* and our semantic methods (Uncorrelated-SNA-Algorithm1, Correlated-SNA-Algorithm1 and Correlated-SNA-Algorithm2) in *Dataset1*.

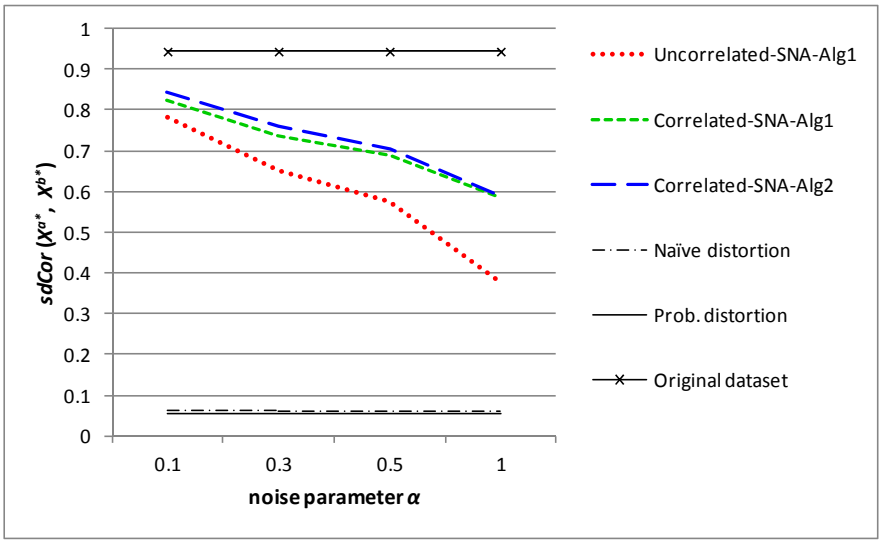


Figure 7-6. Evaluation of the *semantic distance correlation* for the *naïve distortion*, *probabilistic distortion* and our semantic methods (Uncorrelated-SNA-Algorithm1, Correlated-SNA-Algorithm1 and Correlated-SNA-Algorithm2) in *Dataset1*.

However, distribution-based data distortions are able to preserve the mean better in some cases, as shown Table 7.10. This is due to the small spectrum of different categories in the dataset, that is, only 17 different diseases for attribute X^a and 19 for X^b, and the large and even balance of repetitions among the categories, which configure a favorable scenario for methods based on the distribution of the data. It is expected that distribution-based methods produce worse results with fine grained datasets with many different categories and uneven distributions.

Table 7.10. Evaluation of the *semantic mean* for the *naïve distortion*, *probabilistic distortion* and our semantic methods (Uncorrelated-SNA-Algorithm1, Correlated-SNA-Algorithm1 and Correlated-SNA-Algorithm2) in *Dataset1*.

Metric	Naïve distortion	Probabilistic distortion	Uncorrelated SNA-1 (α=1)	Correlated SNA-1 (α=1)	Correlated SNA-2 (α=1)
$sd(sMean(X^{a*}), sMean(X^{a}))$	0.20	0.04	0.20	0.20	0.20
$sd(sMean(X^{b*}), sMean(X^{b}))$	0.27	0	0.18	0.45	0.45

To test the generality of our methods, in a second experiment, we configured a dataset named *Dataset2,* with 1,316 patients and two strongly correlated attributes belonging to different taxonomies: X^a =*principal diagnosis*, which is associated with the taxonomy of *diseases* and X^b =*medical procedure*, which is associated with the taxonomy of *procedures,* both from SNOMED-CT. This allows us to compare Correlated-SNA-Algorithm1 with the version designed to optimize the preservation of the correlation between attributes with domains in different taxonomies, Correlated-SNA-Algorithm3. Table 7.11 depicts the semantic features of *Dataset2*, which also shows a strong correlation of 0.87; Tables 7.12 and 7.13 depict the evaluation metrics for *Dataset2*.

Table 7.11. Semantic features of *Dataset2*: 1,316 patients with two strongly correlated attributes, X^a = *principal diagnosis,* X^b = *medical procedure* with different associated taxonomies.

Semantic feature	Value
$sMean(X^a)$	Malignant neoplasm of costovertebral joint
$sMean(X^b)$	Arthrectomy of hip
$sVar(X^a)$	0.15
$sVar(X^b)$	0.07
$sdCor(X^a, X^b)$	0.88

Table 7.12. Evaluation metrics of the noise-added dataset provided by Correlated-SNA-Algorithm1 for *Dataset2* (X^a =*principal diagnosis* and X^b =*medical procedure*).

Metric	α=0.1	α=0.3	α=0.5	α=1
$sMean(X^{a^*})$	Second. malignant neoplasm of lumbar vertebral column	Malignant neo-plasm of costo-vertebral joint	Malignant neo-plasm of costo-vertebral joint	Malignant neo-plasm of costo-vertebral joint
$sd(sMean(X^{a^*}), sMean(X^a))$	0.33	0	0	0
$sVar(X^{a^*}) \mid \lvert sVar(X^{a^*}) - (1+\alpha)\, sVar(X^a)\rvert$	0.18 | 0.02	0.19 | 0.01	0.2 | 0.03	0.23 | 0.07
$RMSE_{Actual}(X^a, X^{a^*}) \mid RMSE_{Target} = \varepsilon^a$	0.23 | 0.16	0.32 | 0.28	0.40 | 0.37	0.50 | 0.52
$sMean(X^{b^*})$	Arthrectomy of hip	Arthrectomy of hip	Arthrectomy of hip	Arthrectomy of hip
$sd(sMean(X^{b^*}), sMean(X^b))$	0	0	0	0
$sVar(X^{b^*}) \mid \lvert sVar(X^{b^*}) - (1+\alpha)\, sVar(X^b)\rvert$	0.08 | 0	0.09 | 0	0.1 | 0.01	0.14 | 0
$RMSE_{Actual}(X^b, X^{b^*}) \mid RMSE_{Target} = \varepsilon^b$	0.15 | 0.13	0.24 | 0.23	0.30 | 0.29	0.41 | 0.40
$sdCor(X^{a^*}, X^{b^*}) \mid \lvert sdCor(X^{a^*}, X^{b^*}) - sdCor(X^a, X^b)\rvert$	0.68 | 0.20	0.56 | 0.32	0.48 | 0.40	0.42 | 0.46

Table 7.13. Evaluation metrics of the noise-added dataset provided by Correlated-SNA-Algorithm3 for *Dataset2* (X^a =*principal diagnosis* and X^b =*medical procedure*).

Metric	α=0.1	α=0.3	α=0.5	α=1
$sMean(X^{a^*})$	Fracture of prox. end of femur	Fracture of prox. end of femur	Recurrent dislocation of joint	Recurrent dislocation of joint
$sd(sMean(X^{a^*}), sMean(X^a))$	0.23	0.23	0.23	0.23
$sVar(X^{a^*}) \mid \lvert sVar(X^{a^*}) - (1+\alpha)\, sVar(X^a)\rvert$	0.2 | 0.04	0.22 | 0.03	0.23 | 0.01	0.26 | 0.04
$RMSE_{Actual}(X^a, X^{a^*}) \mid RMSE_{Target} = \varepsilon^a$	0.27 | 0.16	0.34 | 0.28	0.41 | 0.37	0.51 | 0.52
$sMean(X^{b^*})$	Arthrectomy of hip	Arthrectomy of hip	Arthrectomy of hip	Arthrectomy of hip
$sd(sMean(X^{b^*}), sMean(X^b))$	0	0	0	0
$sVar(X^{b^*}) \mid \lvert sVar(X^{b^*}) - (1+\alpha)\, sVar(X^b)\rvert$	0.08 | 0	0.11 | 0.02	0.13 | 0.03	0.2 | 0.06
$RMSE_{Actual}(X^b, X^{b^*}) \mid RMSE_{Target} = \varepsilon^b$	0.16 | 0.13	0.25 | 0.23	0.30 | 0.29	0.41 | 0.40
$sdCor(X^{a^*}, X^{b^*}) \mid \lvert sdCor(X^{a^*}, X^{b^*}) - sdCor(X^a, X^b)\rvert$	0.70 | 0.18	0.60 | 0.28	0.54 | 0.34	0.47 | 0.41

Such as the preceding case, and for the same reasons, Correlated-SNA-Algorithm1 better preserves the mean, while Correlated-SNA-Algorithm3 provides a better-preserved correlation, despite using the same noise sequences in both methods.

7.6 Conclusion

The empirical study carried on non-independent nominal attributes has shown that our methods are capable of permuting (with semantic rank swapping) or distorting (with semantic noise addition) values consistently with the desired level of perturbation, and incurring in an information loss much lower than non-semantic perturbation methods based on the distribution of the data. Another strength of our proposal is that both the semantic multivariate rank swapping solution and the semantic correlated noise addition solution are able to largely preserve the semantic correlation between attributes, while this is totally broken by non-semantic methods.

Our semantic proposals also provide a configurable level of perturbation, which is achieved through k in the semantic rank swapping methods and α in the semantic noise addition methods. Concerning the preservation of univariate statistical features (mean and variance), the permutation-based

approach provided better outcomes than the distortion-based approach because the former is capable, by construction, of perfectly preserving them.

These benefits, together with the privacy models that both methods can satisfy, make our methods yield protected data that are significantly useful for analysis while offering robust privacy guarantees. Specifically, as discussed in Section 2.5, rank swapping yields probabilistic *k*-anonymous datasets, and noise addition may be used as sanitization mechanism to attain differential privacy.

Chapter 8 **Conclusions and future work**

This thesis has dealt with privacy-preserving methods in microdata releases. Among the available methods, we have focused on perturbative mechanisms based on data permutation and noise addition, which constitute the basic principles underlying data protection [17], and which outstand due to their ability to preserve certain features of the data better than other protection mechanisms [5]. The focus has been placed on improving the utility of protected nominal data that, nowadays, constitutes most of the personal data on individuals that is compiled, aggregated and exploited by third parties, and which are of outmost importance for research [2]. By exploiting the structured knowledge modeled in ontologies and the notion of semantic similarity, we have proposed an arsenal of semantically-grounded operators; these enabled us to adapt to the nominal domain perturbative data protection mechanisms that, in principle, were restricted to numerical data. The empirical experiments we carried out on real nominal data supported our starting hypothesis: by relying on sound semantic tools, nominal data can be protected while retaining their utility significantly better than with methods that neglect data semantics at all.

8.1 Contributions

The contributions of our work are the following:

- The accurate management of nominal data is not straightforward because, on the contrary to numbers, they take values from a discrete and finite list of non-ordinal categories, which are usually expressed by words. In this scenario, it is not possible to carry out neither the arithmetical data operations nor the data ranking needed by most data protection methods. To address this issue, in Chapter 4 we have formalized a set of operators (the difference, the mean, the variance, the covariance, the correlation and the sort operator) that, by relying on formal knowledge structures, enable a semantically-coherent interpretation of nominal data without neglecting

their distributional features. In particular, our work is the first that incorporates semantics into the definition of the distance covariance and correlation measures, in order to assess the semantic dependence between nominal attributes. Finally, as sort operator, we have provided a total preorder relation that allows semantically sorting all nominal values of an attribute.

- In Chapter 5 we have presented a semantically-grounded permutation-based mechanism to protect nominal data alternative to the standard rank swapping method. To capture and manage the semantics underlying to nominal values, our algorithms exploit the formal knowledge modeled in ontologies. By using the semantic difference and sort operators defined in Chapter 4, data permutation can be done consistently with data semantics. In particular, we have proposed solutions to protect individual nominal attributes and multivariate datasets which are capable of protecting nominal data while preserving their semantic features. The latter is of great interest for data analysis, because it is able to protect multivariate datasets while reasonably preserving the semantic relationship among attributes. In this way, the inferences extracted from the semantic analysis of non-independent attributes protected with our method will be similar to those drawn from the original data. Our mechanism is also capable of limiting the scope of the permutation, which allows controlling the perturbation level and, thus, the information loss incurred by data protection.
- In Chapter 6 we have presented the notion and practical enforcement of *semantic noise*, a semantically-grounded version of the standard numerical noise addition mechanism that is capable of distorting nominal data while preserving their semantics. Like the semantic permutation-based mechanism presented in Chapter 5, semantic noise captures and manages the underlying semantics of the nominal values to be distorted by exploiting ontologies and by using several of the semantic operators defined in Chapter 4 (the difference, the mean, the variance and the covariance). Particularly, we have proposed solutions for the two main families of noise addition methods: uncorrelated noise for individual attributes and correlated noise for multivariate datasets. The correlated solution is able to protect multivariate datasets while reasonably preserving their correlation structure. In addition, several strategies have been proposed to guide the replacement of values during the noise

addition process towards the preservation of either the semantic mean or the semantic distance correlation. On the other hand, unlike other distortion-based methods discussed in Chapter 3, our algorithms provide a configurable distortion level, thus being able to control the information loss of the noised-added data.

- In Chapter 7 we have carried out a comprehensive empirical study of our proposals on several nominal multivariate datasets containing real patient discharge data provided by the California Office of Statewide Health Planning and Development (OSHPD) and by exploiting the medical structured knowledge provided SNOMED-CT. Regarding our permutation-based data protection proposal, the outcomes show that is capable of permuting values consistently with the desired level of protection, and incurring in an information loss much lower than non-semantic swapping methods depicted in Chapter 3. Especially, the experiments evidences that our swapping methods yield protected data that are significantly useful for analysis because (i) perfectly preserve all univariate features of the dataset, such as the mean, variance, frequency distribution, outlying values, granularity and cardinality and (ii) largely preserve the semantic correlation between attributes, while this is totally broken by non-semantic swapping. Regarding data protection, our proposal provides the *ex-ante* privacy guarantees of *probabilistic k-anonymity*, which offers the same practical protection against disclosure than *k-anonymity*, but imposing less constraints on the way data should be transformed [26]. Regarding our noise-addition mechanisms, the empirical study evidences that our methods are capable of replacing original values by noisy ones within a semantic distance consistent with the desired distortion level significantly better than non-semantic distortion methods based on the distribution of the data. Another strength of our methods is that they are able to largely preserve the correlation between attributes for typical noise levels, while this is totally broken by the methods based on the distribution of the data. These benefits, together with the preservation of other statistical features such as the mean, ensure our methods yield protected data that are still useful for statistical analysis.

8.2 Publications

The following publications support the contributions described in this thesis:

- Mercedes Rodriguez-Garcia, Montserrat Batet, David Sánchez, "Semantic Noise: Privacy-Protection of Nominal Microdata through Uncorrelated Noise Addition", in: *Proceedings of the 27th IEEE International Conference on Tools with Artificial Intelligence*, ICTAI 2015, Vietri sul Mare, Italy, IEEE, 2015, pp. 1106-1113, ISSN: 1082-3409, https://doi.org/10.1109/ICTAI.2015.157. **CORE B**.
- Mercedes Rodriguez-Garcia, David Sánchez, Montserrat Batet, "Perturbative Data Protection of Multivariate Nominal Datasets", in: *Proceedings of International Conference on Privacy in Statistical Databases*, PSD 2016, Dubrovnik, Croatia, Springer International Publishing, 2016, pp. 94-106, ISBN 978-3-319-45381-1, http://dx.doi.org/10.1007/978-3-319-45381-1_8. **CORE C**.
- Mercedes Rodriguez-Garcia, Montserrat Batet, David Sánchez, "A semantic framework for noise addition with nominal data", *Knowledge-Based Systems*, Available online 24 January 2017, ISSN 0950-7051, http://dx.doi.org/10.1016/j.knosys.2017.01.032. **ISI-JCR Impact factor: 3.325 (1st quartile)**.
- Mercedes Rodriguez-Garcia, Montserrat Batet, David Sánchez, "Utility-preserving privacy protection of medical nominal datasets via semantic rank swapping", *Journal of Biomedical Informatics* (Submitted - under review). **ISI-JCR Impact factor: 2.447 (1st quartile)**.

8.3 Future Work

The work presented in this thesis opens several avenues for future research:

- Thanks to the mathematical consistence of our methods, we plan to treat heterogeneous data involving numerical and nominal attributes by integrating our semantic mechanisms with the standard numerical ones.

- Other more sophisticated semantic distance measures exploiting several knowledge sources [110, 111, 113, 114] may be also considered to better capture the semantics underlying nominal attributes whose values are spread through several domains of knowledge.
- In semantic noise addition, we plan to further refine the strategies used to guide the replacement of nominal values whereby we can better preserve a particular feature of the data, e.g., the average error or the mean, in case the posterior data analysis strongly depends on that feature, or to achieve the best balance between all of them.
- As detailed in Chapter 2, data shuffling [48] is a permutation-based method aimed at protecting numeric confidential attributes that largely preserves the correlation structure of the dataset. Since the masking process combines additive noise perturbation with data swapping, it could also be adapted to the nominal data domain by using our semantic approaches. However, as this method is patented, its adaptation would require the collaboration of the authors.

Specific applications of noise addition also open lines of future work:

- Because our semantic noise addition method is not linked to a specific noise distribution, it can be used to implement other noise-based mechanisms on nominal data, such as Laplace noise, which is widely used to enforce ε-differential privacy. In this sense, it would be interesting to evaluate the performance of different semantic similarity calculation paradigms with respect to the desired noise distribution, so that we can end with a set of the best suited measures for each type of noise.
- In the context of Artificial Neural Networks (ANNs), numerical noise addition is commonly adopted to reduce overfitting [115]: by adding different levels of noise to the training data of ANNs, the system will learn to ignore irrelevant information (noise) during the tune-up process, thereby improving its response capacity in face of new data. By adding different levels of noise to the training data used by ANNs, the system will learn to ignore irrelevant information during the tune-up process, thereby improving its response capacity in the face of new data. Other machine learning paradigms, such as incremental learning [116] and online machine learning [117], also apply this technique to build high-

performance predictive models. Thus, our semantic noise addition methods can be used to add a controlled amount of noise to nominal data, so that these can be used as input to train machine learning algorithms while avoiding overfitting.

- Unlike permutation-based and aggregation-based data protection, which require a set of records as input [5], noise addition is able to deal with records one by one. This feature is particularly useful for online anonymization of transactional data [27, 103], where data streams are dynamic and must be protected on the fly [104]. Mobile aggregation applications, such as large-scale mobile surveys or sensor network aggregation applications, are emerging cases of data streams in which noise addition is used to protect data privacy [118]. This suggests exploring the behavior of semantic noise addition in a transactional nominal data scenario, such as the online protection of query logs [108, 119]. To correlate the noise with the stream behavior, one could use the correlations in different time series while deciding the noise to be added to any particular value, as proposed in [104].

Bibliography

[1] E. Ramírez, J. Brill, M. Ohlhausen, J. Wright, T. Mc-Sweeny, Data brokers: A call for transparency and accountability, U.S. Federal Trade Commission FTC (May 2014).

[2] M. Elliot, K. Purdam, D. Smith, Statistical disclosure control architectures for patient records in biomedical information systems, Journal of Biomedical Informatics 41 (1) (2008) 58–64.

[3] V. Ciriani, S. Vimercati, S. Foresti, P. Samarati, Microdata protection, in: Secure Data Management in Decentralized Systems, Springer US, 2007, pp. 291–321.

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

[5] A. Hundepool, J. Domingo-Ferrer, L. Franconi, S. Giessing, E.S. Nordholt, K. Spicer, P.-P. Wolf, Statistical Disclosure Control, Wiley, 2012.

[6] B.C.M. Fung, K. Wang, R. Chen, P.S. Yu, Privacy-preserving Data Publishing: A Survey of Recent Developments, ACM Computing Surveys 42 (4) (2010) 14:11-14:53.

[7] L. Xu, C. Jiang, J. Wang, J. Yuan, Y. Ren, Information Security in Big Data: Privacy and Data Mining, IEEE Access 2 (2014) 1149-1176.

[8] A. Viejo, D. Sánchez, Enforcing transparent access to private content in social networks by means of automatic sanitization, Expert Systems with Applications 62 (15) (2016) 148-160.

[9] D. Sánchez, M. Batet, A. Viejo, Utility-preserving privacy protection of textual healthcare documents, Journal of Biomedical Informatics 52 (C) (2014) 189-198.

[10] M. Batet, A. Erola, D. Sánchez, J. Castellà-Roca, Utility preserving query log anonymization via semantic microaggregation, Information Sciences 242 (2013) 49-63.

[11] M. Batet, A. Erola, D. Sánchez, J. Castellà-Roca, Semantic anonymisation of set-valued data, in: Proceedings of the International Conference on Agents and Artificial Intelligence (ICAART'14), vol. 1, 2014, pp. 102-112.

[12] V. Torra, Towards knowledge intensive data privacy, Data Privacy Management and Autonomous Spontaneous Security 6514 (2011) 1-7.

[13] S. Martínez, D. Sánchez, A. Valls, A semantic framework to protect the privacy of electronic health records with non-numerical attributes, Journal of biomedical informatics 46 (2) (2013) 294-303.

[14] N. Guarino, Formal Ontology and Information Systems, in: Proceedings of the 1st International Conference on Formal Ontology in Information Systems, FOIS 1998, IOS Press, 1998, pp. 3-15.

[15] M. Batet, D. Sánchez, A. Valls, An ontology-based measure to compute semantic similarity in biomedicine, Journal of Biomedical Informatics 44 (1) (2011) 118-125.

[16] M. Batet, D. Sánchez, A review on semantic similarity, in: Encyclopedia of Information Science and Technology, Third Edition, IGI Global, 2015, pp. 7575-7583.

[17] J. Domingo-Ferrer, K. Muralidhar, New directions in anonymization: Permutation paradigm, verifiability by subjects and intruders, transparency to users, Information Sciences 337–338 (2016) 11-24.

[18] E. McCallister, T. Grance, K. Scarfone, Guide to Protecting the Confidentiality of Personally Identifiable Information (PII), in: Special Publication 800-122, National Institute of Standards and Technology, U.S. Department of Commerce, 2010.

[19] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). http://data.europa.eu/eli/reg/2016/679/oj.

[20] Standard for privacy of individually identifiable health information. *Federal Register, Special Edition*, pages 768–769, October 2007.

[21] T.D. Gunter, N.P. Terry, The emergence of national electronic health record architectures in the United States and Australia: models, costs, and questions, Journal of Medical Internet Research 7 (1) (2005).

[22] HIPAA. Health insurance portability and accountability act, 2004. http://www.hhs.gov/ocr/hipaa/.

[23] European Comission: European Privacy Regulations. http://ec.europa.eu/justice_home/fsj/privacy/index_en.htm (2016)

[24] J. Domingo-Ferrer, D. Sánchez, J. Soria-Comas, Database Anonymization: Privacy Models, Data Utility, and Microaggregation-based Inter-model Connections, in: Synthesis Lectures on Information Security Privacy and Trust, Morgan & Claypool, 2016, pp. 1-136.

[25] C. Dwork, Differential privacy, in: Proceedings of the 33rd International Colloquium on Automata, Languages and Programming (ICALP 2006), vol. 4052 of the series Lecture Notes in Computer Science, Springer Berlin Heidelberg, 2006, pp. 1-12.

[26] J. Soria-Comas, J. Domingo-Ferrer, Probabilistic k-anonymity through microaggregation and data swapping, in: Proceedings of 2012 IEEE International Conference on Fuzzy Systems, 2012, pp. 1-8.

[27] C.C. Aggarwal, P.S. Yu, A General Survey of Privacy-Preserving Data Mining Models and Algorithms, in: Privacy-Preserving Data Mining, Springer US, 2008, pp. 11-52.

[28] L. Willenborg, T. Waal, Elements of Statistical Disclosure Control. Lecture Notes in Statistics, vol. 155, Springer New York, 2001.

[29] W.E. Winkler, Re-identification methods for masked microdata, in: Privacy in Statistical Databases, volume 3050 of Lecture Notes in Computer Science, Berlin Heidelberg Springer, 2004, pp. 216–230.

[30] J.P. Reiter, Releasing multiply-imputed, synthetic public use microdata: An illustration and empirical study, Journal of the Royal Statistical Society, Series A 168 (2005) 185–205.

[31] P. Samarati, L. Sweeney, Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression, in: Technical Report SRI-CSL-98-04, Computer Science Laboratory, SRI International, 1998.

[32] R. Conway, D. Strip, Selective partial access to a database, Cornell University, Tech. Rep. (1976).

[33] P. Tendick, Optimal noise addition for preserving confidentiality in multivariate data, Journal of Statistical Planning and Inference 27 (3) (1991) 341-353.

[34] K. Muralidhar, R. Sarathy, Security of random data perturbation methods, ACM Transactions on Database Systems 24 (2000) 487-493.

[35] J. Kim, A method for limiting disclosure in microdata based on random noise and transformation, in: Proceedings of the ASA Section on Survey Research Methods, 1986, pp. 370-374.

[36] T. Dalenius, S.P. Reiss, Data-swapping: a technique for disclosure control (extended abstract), in: Proceedings of the ASA Section on Survey Research Methods, 1978, pp. 191–194.

[37] S.P. Reiss, Non-reversible privacy transform, in: Proceedings of the ACM Symposium on Principles of Database Systems, Los Angeles, CA, USA, 1982.

[38] B. Greenberg, Rank swapping for masking ordinal microdata, U.S. Bureau of the Census (unpublished manuscript) (1987).

[39] R.A. Moore, Controlled data swapping techniques for masking public use microdata sets, in, Statistical Research Division Report Series RR 96-04, U. S. Bureau of the Census, Washington, DC, 1996.

[40] J. Domingo-Ferrer, V. Torra, A quantitative comparison of disclosure control methods for microdata, in: L.J.I. Doyle P., Theeuwes J.J.M., Zayatz L.V. (Ed.) Confidentiality, disclosure, and data access: Theory and practical applications for statistical agencies, Elsevier, 2001, pp. 111-134.

[41] J. Nin, J. Herranz, V. Torra, Rethinking Rank Swapping to Decrease Disclosure Risk, Data & Knowledge Engineering 64 (1) (2008) 346-364.

[42] D. Defays, M.N. Anwar, Masking microdata using micro-aggregation, Journal of Official Statistics 14 (4) (1998) 449–461.

[43] D. Defays, P. Nanopoulos, Panels of enterprises and confidentiality: the small aggregates method, in: Proceedings of 92 Symposium on Design and Analysis of Longitudinal Surveys, Ottawa, Canada, 1993, pp. 195–204.

[44] J. Domingo-Ferrer, J.M. Mateo-Sanz, A. Oganian, V. Torra, A. Torres, On the security of microaggregation with individual ranking: analytical attacks, International Journal of Uncertainty Fuzziness and Knowledge-Based Systems 18 (5) (2002) 477–492.

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

[46] J. Domingo-Ferrer, V. Torra, Ordinal, continuous and heterogeneous k-anonymity through microaggregation, Data Minining and Knowledge Discovery 11 (2) (2005) 195–212.

[47] V. Torra, Microaggregation for Categorical Variables: A Median Based Approach, in: Lecture Notes in Computer Science (Privacy in Statistical Databases - PSD2004), vol. 3050, Springer Berlin Heidelberg, 2004, pp. 162-174.

[48] K. Muralidhar, R. Sarathy, Data Shuffling - A New Masking Approach for Numerical Data, Management Science 52 (5) (2006) 658 - 670.

[49] J.M. Mateo-Sanz, F. Sebé, J. Domingo-Ferrer, Outlier Protection in Continuous Microdata Masking, in: Proceedings Privacy in Statistical Databases: CASC Project

Final Conference (PSD 2004) Springer Berlin Heidelberg, Barcelona, Spain, 2004, pp. 201-215.

[50] A.F. Karr, C.N. Kohnen, A. Oganian, J.P. Reiter, A.P. Sanil, A framework for evaluating the utility of data altered to protect confidentiality, The American Statistician 60 (3) (2006) 224-232.

[51] K. Muralidhar, R. Sarathy, R. Dandekar, Why Swap When You Can Shuffle? A Comparison of the Proximity Swap and Data Shuffle for Numeric Data, in: Proceedings Privacy in Statistical Databases: CENEX-SDC Project International Conference (PSD 2006), Springer Berlin Heidelberg, 2006, pp. 164-176.

[52] A. Meyerson, R. Williams, On the complexity of optimal k-anonymity, in: Proceedings of the twenty-third ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems (PODS '04), ACM Press, 2004, pp. 223–228.

[53] A. Machanavajjhala, D. Kifer, J. Gehrke, M. Venkitasubramaniam, *l*-Diversity: Privacy beyond *k*-anonymity, ACM Transactions on Knowledge Discovery from Data 1 (1) (2007).

[54] N. Li, T. Li, S. Venkatasubramanian, *t*-Closeness: Privacy beyond *k*-anonymity and *l*-diversity, in: IEEE 23rd International Conference on Data Engineering (ICDE 2007), IEEE, 2007, pp. 106–115.

[55] N. Li, T. Li, S. Venkatasubramanian, Closeness: a new privacy measure for data publishing, IEEE Transactions on Knowledge and Data Engineering 22 (7) (2010) 943–956.

[56] J. Soria-Comas, J. Domingo-Ferrer, D. Sánchez, S. Martínez, t-Closeness through Microaggregation: Strict Privacy with Enhanced Utility Preservation, IEEE Trans. Knowl. Data Eng. 27 (11) (2015) 3098-3110.

[57] C. Dwork, F. Mcsherry, K. Nissim, A. Smith, Calibrating noise to sensitivity in private data analysis, in: Proceedings of the 3rd Theory of Cryptography Conference, Springer, 2006, pp. 265–284.

[58] S. Inusah, T.J. Kozubowski, A discrete analogue of the laplace distribution, Journal of Statistical Planning and Inference 136 (3) (2006) 1090–1102.

[59] A. Ghosh, T. Roughgarden, M. Sundararajan, Universally utility-maximizing privacy mechanisms, in: Proceedings of the ACM Symposium on Theory of Computing (STOC'09), 2009, pp. 351-360.

[60] A. Blum, K. Ligett, A. Roth, A learning theory approach to noninteractive database privacy, in: Proceedings of the 40th Annual ACM Symposium on Theory of Computing (STOC 2008), 2008, pp. 609–618.

[61] C. Dwork, M. Naor, O. Reingold, G.N. Rothblum, S. Vadhan, On the complexity of differentially private data release: efficient algorithms and hardness results, in: Proceedings of the 41st Annual ACM Symposium on Theory of Computing (STOC 2009), 2009, pp. 381–390.

[62] D. Sánchez, J. Domingo-Ferrer, S. Martínez, J. Soria-Comas, Utility-preserving differentially private data releases via individual ranking microaggregation, Information Fusion 30 (2016) 1-14.

[63] J. Soria-Comas, J. Domingo-Ferrer, D. Sánchez, S. Martínez, Enhancing Data Utility in Differential Privacy via Microaggregation-based K-anonymity, The VLDB Journal 23 (5) (2014) 771-794.

[64] P. Kooiman, L. Willenborg, J. Gouweleeuw, PRAM: A method for disclosure limitation of microdata, Research Paper 9705, Statistics Netharlands, P.O. Box 4000, 2270 JM Voorburg, The Netherlands (1997).

[65] H. Giggins, L. Brankovic, Protecting privacy in genetic databases, in: Proceedings of the 6th Engineering Mathematics and Applications Conference (EMAC 2003), vol. 2, 2003, pp. 73-78.

[66] M.Z. Islam, L. Brankovic, Privacy preserving data mining: A noise addition framework using a novel clustering technique, Knowledge-Based Systems 24 (8) (2011) 1214-1223.

[67] F. McSherry, K. Talwar, Mechanism design via differential privacy, in: Proceedings of Annual IEEE Symposium on Foundations of Computer Science (FOCS'07), 2007, pp. 94-103.

[68] H. Giggins, L. Brankovic, VICUS: A Noise Addition Technique for Categorical Data, in: Proceedings of the 10th Australasian Data Mining Conference (AusDM '12), vol. 134, Australian Computer Society, Inc., 2012, pp. 139-148.

[69] D. Abril, G. Navarro-Arribas, V. Torra, On the declassification of confidential documents, Modeling Decision for Artificial Intelligence 6820 (2011) 235-246.

[70] T. Dalenius, S.P. Reiss, Data-swapping: A technique for disclosure control, Journal of Statistical Planning and Inference 6 (1982) 73-85.

[71] J. Schlörer, Security of Statistical Databases: Multidimensional Transformation, ACM Transactions on Database Systems 6 (1) (1981) 95-112.

[72] S.P. Reiss, Practical data-swapping: The first steps, ACM Transactions on Database Systems 9 (1984) 20–37.

[73] S.E. Fienberg, J. McIntyre, Data Swapping: Variations on a Theme by Dalenius and Reiss, in: Privacy in Statistical Databases, series Lecture Notes in Computer Science 3050, Springer Berlin Heidelberg, 2004, pp. 14-29.

[74] J. Domingo-Ferrer, D. Sánchez, J. Soria-Comas, Anonymization Methods for Microdata, in: Database Anonymization: Privacy Models, Data Utility, and Microaggregation-based Inter-model Connections, Morgan & Claypool Publishers, 2016, pp. 15-23.

[75] V. Torra, Rank swapping for partial orders and continuous variables, in: Proceedings of the International Conference on Availability, Reliability and Security (ARES 2009), IEEE, 2009, pp. 888-893.

[76] D. Abril, G. Navarro-Arribas, V. Torra, Towards Semantic Microaggregation of Categorical Data for Confidential Documents, in: Proceedings of the 7th International Conference on Modeling Decisions for Artificial Intelligence, MDAI 2010, Springer Berlin Heidelberg, 2010, pp. 266-276.

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

[78] P. Cimiano, Ontology Learning and Population from Text: Algorithms, Evaluation and Applications, Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006.

[79] C. Fellbaum, WordNet: An Electronic Lexical Database, MIT Press, 1998.

[80] D.B. Lenat, R.V. Guha, Building Large Knowledge-Based Systems: Representation and Inference in the Cyc Project, Addison-Wesley, Reading, Massachusetts, 1990.

[81] K.A. Spackman, SNOMED CT milestones: endorsements are added to already-impressive standards credentials, Healthcare informatics: the business magazine for information and communication systems 21 (9) (2004) 54-56.

[82] M. Batet, D. Isern, L. Marin, S. Martínez, A. Moreno, D. Sánchez, A. Valls, K. Gibert, Knowledge-driven delivery of home care services, Journal of Intelligent Information Systems 38 (1) (2012) 95-130.

[83] A. Valls, K. Gibert, D. Sánchez, M. Batet, Using ontologies for structuring organizational knowledge in Home Care assistance, International Journal of Medical Informatics 79 (5) (2010) 370-387.

[84] L. Ding, T. Finin, A. Joshi, R. Pan, R.S. Cost, Y. Peng, P. Reddivari, V.C. Doshi, J. Sachs, Swoogle: A Search and Metadata Engine for the Semantic Web, in:

Proceedings of the 13th ACM Conference on Information and Knowledge Management, CIKM 2004, ACM Press, 2004, pp. 652-659.

[85] R. Rada, H. Mili, E. Bicknell, M. Blettner, Development and application of a metric on semantic nets, IEEE Transactions on Systems, Man and Cybernetics 19 (1) (1989) 17-30.

[86] C. Leacock, M. Chodorow, Combining local context and WordNet similarity for word sense identification, in: WordNet: An Electronic Lexical Database, MIT Press, 1998, pp. 265-283.

[87] Z. Wu, M. Palmer, Verbs semantics and lexical selection, in: Proceedings of the Annual Meeting of the Association for Computational Linguistics, 1994, pp. 133-139.

[88] A. Tversky, Features of Similarity, Psychological Review 84 (4) (1977) 327-352.

[89] D. Sánchez, M. Batet, D. Isern, A. Valls, Ontology-based semantic similarity: A new feature-based approach, Expert Systems with Applications 39 (9) (2012) 7718-7728.

[90] M.A. Rodríguez, M.J. Egenhofer, Determining semantic similarity among entity classes from different ontologies, IEEE Transactions on Knowledge and Data Engineering 15 (2) (2003) 442-456.

[91] P. Resnik, Using Information Content to Evalutate Semantic Similarity in a Taxonomy, in: Proceedings of the 14th International Joint Conference on Artificial Intelligence, IJCAI 1995, vol. 1, 1995, pp. 448-453.

[92] D. Lin, An Information-Theoretic Definition of Similarity, in: Proceedings of the 15th International Conference on Machine Learning, ICML 1998, 1998, pp. 296-304.

[93] J.J. Jiang, D.W. Conrath, Semantic similarity based on corpus statistics and lexical taxonomy, in: Proceedings of 10th International Conference on Research in Computational Linguistics, ROCLING'97, 1997.

[94] N. Seco, T. Veale, J. Hayes, An Intrinsic Information Content Metric for Semantic Similarity in WordNet, in: Proceedings of the 16th Eureopean Conference on Artificial Intelligence (ECAI 2004) including Prestigious Applicants of Intelligent Systems (PAIS 2004), 2004, pp. 1089-1090.

[95] Z. Zhou, Y. Wang, J. Gu, A New Model of Information Content for Semantic Similarity in WordNet, in: Proceedings of the Second International Conference on Future Generation Communication and Networking Symposia, FGCNS 2008, 2008, pp. 85-89.

[96] D. Sánchez, M. Batet, D. Isern, Ontology-based Information Content computation, Knowledge-based Systems 24 (2) (2011) 297-303.

[97] S. Martínez, A. Valls, D. Sánchez, Semantically-grounded construction of centroids for datasets with textual attributes, Knowledge-Based Systems 35 (2012) 160-172.

[98] D. Sánchez, M. Batet, S. Martínez, J. Domingo-Ferrer, Semantic variance: An intuitive measure for ontology accuracy evaluation, Engineering Applications of Artificial Intelligence 39 (2015) 89-99.

[99] G.J. Székely, M.L. Rizzo, N.K. Bakirov, Measuring and testing dependence by correlation of distances, Annals of Statistics 35 (6) (2007) 2769-2794.

[100] J. Kong, B.E.K. Klein, R. Klein, K. Lee, G. Wahba, Using distance correlation and SS-ANOVA to assess associations of familial relationships, lifestyle factors, diseases, and mortality, in: Proceedings of the National Academy of Sciences, vol. 109, 2012, pp. 20352-20357.

[101] M. Omelka, Š. Hudecová, A comparison of the Mantel test with a generalised distance covariance test, Environmetrics 24 (7) (2013) 449-460.

[102] J. Domingo-Ferrer, D. Sánchez, G. Rufian-Torrell, Anonymization of nominal data based on semantic marginality, Information Sciences 242 (2013) 35-48.

[103] G. Krempl, I. Zliobaite, D. Brzezinski, E. Hüllermeier, M. Last, V. Lemaire, T. Noack, A. Shaker, S. Sievi, M. Spiliopoulou, J. Stefanowski, Open Challenges for Data Stream Mining Research, ACM SIGKDD Explorations Newsletter 16 (1) (2014) 1-10.

[104] F. Li, J. Sun, S. Papadimitriou, G. Mihaila, I. Stanoi, Hiding in the Crowd: Privacy Preservation on Evolving Streams through Correlation Tracking, in: Proceedings of the 23rd IEEE International Conference on Data Engineering (ICDE 2007), 2007, pp. 686-695.

[105] C. Makris, Y. Plegas, S. Stamou, Web query disambiguation using PageRank, Journal of the American Society for Information Science and Technology 63 (8) (2012) 1581–1592.

[106] X. Shi, C.C. Yang, Mining related queries from web search engine query logs using an improved association rule mining model, Journal of the American Society for Information Science and Technology 58 (12) (2007) 1871–1883.

[107] A. Viejo, D. Sánchez, Profiling social networks to provide useful and privacy-preserving web search, Journal of the Association for Information Science and Technology (JASIST) 65 (12) (2014) 2444-2458.

[108] D. Sánchez, J. Castellà-Roca, A. Viejo, Knowledge-based scheme to create privacy-preserving but semantically-related queries for web search engines, Information Sciences 218 (2013) 17-30.

[109] D. Sánchez, A. Solé-Ribalta, M. Batet, F. Serratosa, Enabling semantic similarity estimation across multiple ontologies: An evaluation in the biomedical domain, Journal of Biomedical Informatics 45 (1) (2012) 141-155

[110] M. Batet, S. Harispe, S. Ranwez, D. Sánchez, V. Ranwez, An information theoretic approach to improve semantic similarity assessments across multiple ontologies, Information Sciences 283 (2014) 197-210.

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

[112] D. Sánchez, M. Batet, Semantic similarity estimation in the biomedical domain: An ontology-based information-theoretic perspective, Journal of Biomedical Informatics 44 (5) (2011) 749-759.

[113] A. Solé-Ribalta, D. Sánchez, M. Batet, F. Serratosa, Towards the Estimation of Feature-based Semantic Similarity Using Multiple Ontologies, Knowledge-Based Systems 55 (2014) 101-113.

[114] D. Sánchez, M. Batet, A semantic similarity method based on information content exploiting multiple ontologies, Expert Systems with Applications 40 (4) (2013) 1393-1399.

[115] R.M. Zur, Y. Jiang, L.L. Pesce, K. Drukker, Noise injection for training artificial neural networks: A comparison with weight decay and early stopping, Medical Physics 36 (10) (2009) 4810-4818.

[116] X. Geng, K. Smith-Miles, Incremental Learning, in: Encyclopedia of Biometrics, Springer US, 2009, pp. 731-735.

[117] N. Cesa-Bianchi, S. Shalev-Shwartz, O. Shamir, Online Learning of Noisy Data, IEEE Transactions on Information Theory 57 (12) (2011) 7907-7931.

[118] H. Zhang, N. Yu, H. Hu, The Optimal Noise Distribution for Privacy Preserving in Mobile Aggregation Applications, International Journal of Distributed Sensor Networks 10 (2) (2014).

[119] D. Sánchez, M. Batet, C-sanitized: A privacy model for document redaction and sanitization, Journal of the Association for Information Science and Technology (JASIST) 67 (1) (2016) 148-163.

UNIVERSITAT
ROVIRA i VIRGILI