A semantic framework for noise addition with nominal data

Mercedes Rodriguez-Garcia[a1], Montserrat Batet[b], David Sánchez[a]

[a]*UNESCO Chair in Data Privacy, Department of Computer Science and Mathematics, Universitat Rovira i Virgili, Av. Països Catalans, 26, 43007 Tarragona, Catalonia, Spain*

[b]*Internet Interdisciplinary Institute (IN3), Universitat Oberta de Catalunya, Av. Carl Friedrich Gauss, 5, Parc Mediterrani de la Tecnologia, 08860 Castelldefels (Barcelona), Catalonia, Spain*

Abstract

Noise addition is a data distortion technique widely used in data intensive applications. For example, in machine learning tasks it helps to reduce overfitting, whereas in data privacy protection it adds uncertainty to personally identifiable information. Yet, due to its mathematical operating principle, noise addition is a method mainly intended for continuous numerical data. In fact, despite the large amount of nominal data that are being currently compiled and used in data analysis, only a few alternative techniques have been proposed to distort nominal data in a similar way as standard noise addition does for numerical data. Furthermore, all these alternative methods rely on the distribution of the data rather than on the semantics of nominal values, which negatively affects the utility of the distorted outcomes. To tackle this issue, in this paper we present a semantically-grounded alternative to numerical noise suitable for nominal data, which we name *semantic noise*. By means of semantic noise, and by exploiting structured knowledge sources such as ontologies, we are able to distort nominal data while preserving better their semantics and thus, their analytical utility. To that end, we provide semantically and mathematically coherent versions of the statistical operators required in the noise addition process, which include the difference, the mean, the variance and the covariance. Then, we propose semantic noise addition algorithms that cope with the finite, discrete and non-ordinal nature of nominal data. The proposed algorithms cover both uncorrelated noise addition, which is suited to independent attributes, and correlated noise addition, which can cope with multivariate datasets with dependent attributes. Empirical results show that our proposals offer general and configurable mechanisms to distort nominal data while preserving data semantics better than baseline methods based only on the distribution of the data.

Keywords: Noise addition, nominal data, semantics, ontologies, medical ontologies.

[1] Corresponding author. Address: Department of Computer Engineering and Mathematics. Universitat Rovira i Virgili. Avda. Països Catalans, 26. 43007. Tarragona, Catalonia (Spain)
Tel.: +34 977 559657; Fax: +34 977 559710;
E-mail: mercedes.rodriguez@uca.es

1. Introduction

Noise addition is a technique widely used in computer science to distort data. It consists of adding to the input data a noise sequence, typically drawn from a probability distribution such as Gauss or Laplace distribution. The randomized distortion resulting from adding noise is used for a variety of purposes, such as reducing the effect of quantization error [1] or detecting image tampering in digital forensics [2]. In the context of Artificial Neural Networks (ANNs), noise addition is adopted to reduce overfitting [3]. 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 [4] and online machine learning [5], also apply this technique to build high-performance predictive models. In the field of data privacy, within the disciplines of Statistical Disclosure Control (SDC) [6] and Privacy Preserving Data Mining (PPDM) [7, 8], the distortion caused by noise addition is commonly used to protect sensitive data while preserving their analytical utility. In other words, noise addition brings uncertainty to disclosure inferences, so that the identities or the confidential information inferred from the protected data are no longer unequivocal. Moreover, unlike other methods commonly used in SDC and PPDM, such as microaggregation or rank swapping, which require a set of records as input [6], noise addition is able to deal with records one by one. This feature is particularly useful for online anonymization of transactional data [7, 9], where data streams are dynamic and must be protected on the fly [10]. 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 [11].

Because of its mathematical operating principle, noise addition is commonly seen as a method exclusively intended for numerical data [6]. However, most of the data that are currently being gathered from social networks, electronic healthcare records or web browsing logs are nominal [12]. These data are crucial to improve decision-making in business and healthcare, or to offer personalized services that enhance the online experience [8]. Hence, it is particularly important to provide mechanisms whereby noise addition can be employed on nominal data in fields and applications such as those mentioned above.

Nevertheless, unlike numerical data, nominal data are finite, discrete, textual and non-ordinal. For such data, it is not possible to directly apply the arithmetical operations required by noise addition, which are the difference, the mean, the variance and the covariance. Therefore, a first challenge will be to adapt such operations to the domain of nominal data. On the other hand, noise addition should consider and preserve the features of the data in order to maximize the analytical utility of the noise-added outcomes. Whereas such utility for numerical data is a

function of the statistical features of the data, nominal data utility is closely related to the preservation of data *semantics* [13]. Hence, data semantics should be carefully considered during the operations applied to the data [14]. Because of these difficulties, noise addition has been rarely applied to data types other than numerical and, in the few existing applications, the methods proposed in the literature have systematically neglected data semantics [15-19], which has a negative impact on data utility.

To tackle these challenges, in this paper 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 and multivariate nominal datasets. Being able to deal with the multivariate case is especially relevant: indeed, unlike related works focusing on nominal data [15-19], our proposal is able to distort multivariate nominal datasets while reasonably preserving the semantic correlation among attributes.

The rest of the paper is organized as follows. Section 2 discusses related works proposing distortion mechanisms for nominal data. Section 3 gives theoretical background on standard noise addition. Section 4 describes the adaptation to the semantic domain of the statistical operators needed in noise addition and presents our semantically-grounded noise addition algorithms. Section 5 details the empirical work we have conducted and compares our results against baseline methods regarding the preservation of data semantics. Section 6 contains the conclusions and identifies some lines of future research.

2. Related work

Nominal data have been scarcely considered in noise addition methods and, in all cases, 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. [15]. 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 [6].

From another perspective, given that nominal values lack a natural order, some authors [16] 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 [20], 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 [21], 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 [17] 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 [18] 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 particular 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.

However, 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 [19] 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. [22] 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 methods discussed above manage individual attributes independently and, therefore, neglect the potential correlations between attribute pairs. This crucial issue may negatively affect data analysis, which usually exploit attribute correlations to perform inferences. The methods we present in the next section aim at tackling both issues.

3. Background on numerical noise addition

The main approaches to noise addition for numerical data are *uncorrelated noise*, for individual attributes, and *correlated noise*, for multivariate datasets [6].

Uncorrelated noise addition [23] is based on adding sequences of normally distributed random noise to individual attributes from an input dataset. Following the notation in [24], the input dataset X is managed as a set of p attributes or columns, each one corresponding to a different feature of a set of n records.

$$X = \{X_1, \ldots, X_j, \ldots, X_p\}, \tag{1}$$

where $X_j = (x_{1j}, \ldots, x_{ij}, \ldots, x_{nj})$ is the j-th attribute of the dataset and x_{ij} is the value of the attribute j for the record i.

In order to distort the attribute X_j, each value x_{ij} is replaced by a noisy version z_{ij}:

$$Z_j = X_j + \varepsilon_j, \tag{2}$$

where $Z_j = (z_{1j}, \ldots, z_{ij}, \ldots, z_{nj})$ is the distorted attribute and $\varepsilon_j = (\epsilon_{1j}, \ldots, \epsilon_{ij}, \ldots, \epsilon_{nj}) \sim N(0, \sigma_{\varepsilon_j}^2)$ is a noise sequence randomly drawn from a normal distribution with mean zero and variance $\sigma_{\varepsilon_j}^2$. The error variance $\sigma_{\varepsilon_j}^2$ is proportional to the original attribute variance $\sigma_{X_j}^2$ as follows:

$$\sigma_{\varepsilon_j}^2 = \alpha \sigma_{X_j}^2, \quad \alpha > 0 \tag{3}$$

The parameter α determines the amount of noise to be added, whose value usually ranges between 0.1 and 0.5 [25]. The greater the α, the greater the distortion level.

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

$$\begin{aligned} \mu_{Z_j} &= \mu_{X_j} + \mu_{\varepsilon_j} = \mu_{X_j} \\ \sigma_{Z_j}^2 &= \sigma_{X_j}^2 + \sigma_{\varepsilon_j}^2 = (1+\alpha)\sigma_{X_j}^2 \end{aligned} \tag{4}$$

In order to distort multiple attributes, given the uncorrelated character of the method, the noise must be applied to each attribute independently [25, 26], without considering the noise applied to previous attributes. Consequently,

$$Cov(\varepsilon_t, \varepsilon_l) = 0, \quad \forall t \neq l \tag{5}$$

Because the covariance between any two noise vectors ε_t (added to an attribute X_t) and ε_l (added to an attribute X_l) is 0, any correlation between attributes is not preserved after adding noise. Hence, the method is suitable for statistical analyses over individual attributes but not over records with non-independent attributes.

In order to solve this limitation, Kim [27] proposes a method to add correlated random noise to several attributes in a dataset, such that:

$$Z = X + \varepsilon, \tag{6}$$

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

$$\Sigma_\varepsilon = \alpha \Sigma_X, \quad \alpha > 0, \tag{7}$$

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,

$$\begin{aligned} \mu_Z &= \mu_X + \mu_\varepsilon = \mu_X \\ \Sigma_Z &= \Sigma_X + \Sigma_\varepsilon = (1+\alpha)\Sigma_X \\ \rho_{Z_t, Z_l} &= \frac{Cov(Z_t, Z_l)}{\sqrt{\sigma^2_{Z_t}\sigma^2_{Z_l}}} = \frac{(1+\alpha)Cov(X_t, X_l)}{(1+\alpha)\sqrt{\sigma^2_{X_t}\sigma^2_{X_l}}} = \rho_{X_t, X_l} \end{aligned} \tag{8}$$

4. Semantic noise addition methods for nominal data

We propose two semantically-grounded noise addition methods for nominal attributes, a first one with uncorrelated noise and a second one with correlated noise. Both methods exploit the formal semantics provided by an ontology to properly manage and better preserve the semantics underlying the nominal values. The common ground of both methods is that nominal values of each attribute in the input dataset are replaced by other concepts from the same semantic domain extracted from an ontology, which are as semantically distant as the random magnitude of noise defined by the user.

To better preserve the semantics, and therefore, the utility, of noise-added data we adapt to the *semantic domain* the difference, mean, variance and covariance measures that are used to guide the noise addition process. For such purpose, we first introduce the notion of *semantic domain* for nominal data, which is based on an underlying ontology, and discuss which semantic measures are best suited to compare nominal values in a noise addition scenario. Then, we define semantic versions of the statistical operators used in noise addition, which are able to capture the meaning encompassed by nominal values. Finally, by relying on such operators, we propose several heuristic algorithms for *uncorrelated* and *correlated semantic noise addition* for nominal data.

4.1. Defining the semantic domain of a nominal attribute

Contrary to numerical data, nominal data are finite, discrete, textual and non-ordinal. Nominal domains can be expressed either as unstructured term lists or as semantically organized sets of concepts in a knowledge base. Because unstructured term lists neglect data semantics, noise addition would be nearly random, and may distort the meaning of the data much more than is expected from the magnitude of the noise. In this respect, the exploitation of the formal semantics modelled in knowledge bases such as ontologies is more desirable, because with this we can guide the noise addition process toward preserving the *meaning* of nominal data.

An ontology is a structured knowledge source that explicitly and consensually represents the concepts and the semantic interrelations of a domain of knowledge [28]. According to the formal definition proposed in [29], 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 (is-a links), and non-taxonomic relationships, e.g., meronymy (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.

In semantic noise addition, we use an ontology O to replace the original values of a nominal attribute in the dataset by concepts in O that are as semantically distant from the original value as defined by the magnitude of the noise to be added. To ensure the semantic coherence of the results, the concepts used as noisy replacements must belong to the concept hierarchy $\leq_C$ to which the attribute domain refers; that is, if the domain of the attribute values are diseases, the noise-added outcomes must also be diseases.

Formally, let $A = (a_1, \ldots, a_i, \ldots, a_n)$ be a nominal attribute from a dataset X and a_i the value of that attribute in the record i. We assume that the nominal values of A have been associated with concepts c modeled in an ontology O (manually or by lexically matching the strings of nominal values and concept labels, as done in [30]). The semantic domain of the attribute A, denoted by $D(A)$, encompasses all the concepts c in O to which the values in the domain of A may be associated.

Definition 1. Let $S(G)$ be the set of subsumers of a set G of concepts from the ontology O. The *least common subsumer* of G, denoted by $LCS(G)$, is the most specific concept in $S(G)$.

$$S(G) = \bigcup_{c_i \in O} \{c_i \mid \forall c_j \in G: \ c_j \leq_C c_i\}$$
$$LCS(G) = \{c \in S(G) \mid \forall c_i \in S(G): \ c \leq_C c_i\} \tag{9}$$

Definition 2. The *taxonomy associated with the semantic domain of an attribute A* w.r.t. the ontology O, denoted by $\tau(D(A))$, is the concept hierarchy extracted from O that includes all concepts that are taxonomic specializations of $LCS(D(A))$.

$$\tau(D(A)) = \bigcup_{c_i \in O} \{c_i \mid c_i \leq_C LCS(D(A))\}. \tag{10}$$

Note that $\tau(D(A))$ delimits the noise application range in O and, therefore, determines the set of concepts in O that are candidates to replace the original values. Figure 1 shows an example of the *taxonomy associated with the domain of an attribute* on a fragment of the medical ontology SNOMED-CT [31]. Let A={*gastritis*, *hematoma*, *gingivitis*} be a nominal attribute that stores the diseases of a set of 3 patients. The semantic domain of A is the set of all diseases in O, therefore, $LCS(D(A))$ is *Disease* and $\tau(D(A))$ is the taxonomy consisting of the concepts that are taxonomic specializations of $LCS(D(A))$, which are shown in gray in Figure 1.

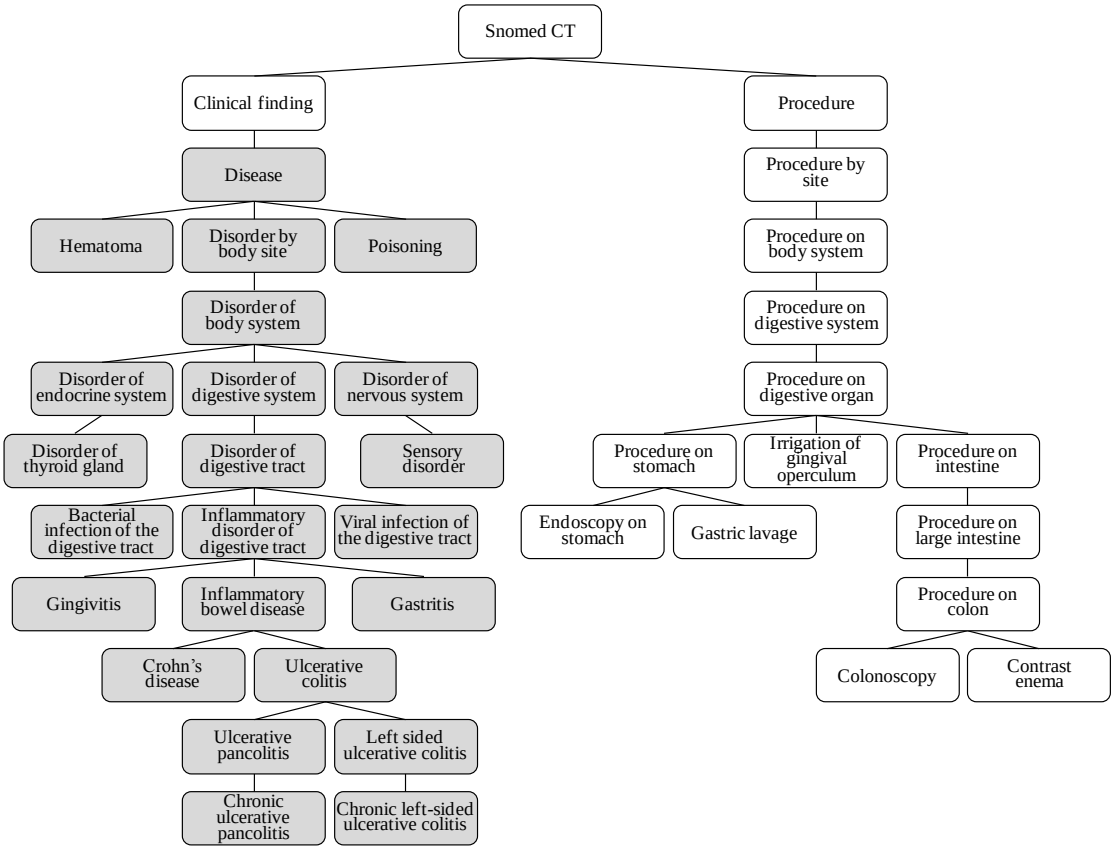

Figure 1. Example of taxonomy associated with the domain of the attribute *Disease* (gray-shaded concepts), extracted from the SNOMED-CT medical ontology.

4.2. Computing the semantic distance between nominal values

In numerical noise addition, many of the operations carried out to manage and transform data 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. To support this in the nominal domain, and because nominal values should be managed according to the semantics of the concepts to which they refer, we replace the arithmetical difference operator by the notion of *semantic distance*.

Semantic distance, $sd: c_1 \times c_2 \rightarrow \mathfrak{R}$, 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 [32].

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 [33].

Different ontology-based distance measures have been proposed in the literature [32]. Firstly, edge-counting measures stand out for their simplicity and low computational cost. The simplest method [34] bases its calculation on the length of the shortest taxonomic path and, therefore, outputs non-normalized values. Other edge-counting methods [35, 36] provide normalized results that better mimic human judgments on similarity by estimating the specificity of concepts from their taxonomic depth, although only [35] performs a non-logarithmic and non-exponential assessment. Secondly, feature-based measures [37, 38] 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 [39], many of these latter

measures use non-taxonomic relationships, a form of knowledge partially modeled in most ontologies [40]. In this regard, [39] 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 [41, 42]. Specifically, the occurrence frequency is used to estimate concept specificity, i.e., infrequent concepts are considered more specific because they encompass more information and, therefore, semantics. 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 [43, 44] 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 [35] because it fulfills the above requirements and reasonably mimic human judgements on semantic similarity [39, 45]. According to Wu and Palmer, the semantic similarity between two concepts modeled in a taxonomy is defined as follows,

$$sim_{wp}(c_1,c_2)=\frac{2\times depth(LCS(c_1,c_2))}{2\times depth(LCS(c_1,c_2))+path(c_1,LCS(c_1,c_2))+path(c_2,LCS(c_1,c_2))}, \quad (11)$$

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*; $path(c_1,LCS(c_1,c_2))$ is the number of taxonomic links in the shortest path between c_1 and $LCS(c_1,c_2)$, similarly for $path(c_2,LCS(c_1,c_2))$. As stated above, 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 [35]. In our proposal, the taxonomy considered to measure semantic similarities is limited to $\tau(D(A))$. Therefore, c_1, c_2, $LCS(c_1, c_2) \in \tau(D(A))$ and the node *root* is the top node of $\tau(D(A))$.

Because sim_{wp} evaluates the similarity between concepts, we formulate sd_{wp} to compute the desired semantic distance, as the opposite of sim_{wp}:

$$sd_{wp}(c_1,c_2)=1-sim_{wp}(c_1,c_2) \quad (12)$$

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 1.

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

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

4.3. Semantic operators for noise addition

To preserve the utility of numerical data, the standard noise addition mechanism detailed in Section 3 relies on the calculation of several statistical features of the data. Specifically, uncorrelated noise addition requires computing the *mean* and the *variance* of the attributes and the correlated method additionally requires computing the *covariance* and the *correlation coefficient* between attribute pair. In our approach, we adapt these statistical measures to the semantic domain because (i) standard arithmetical operations cannot be directly applied on nominal values, and (ii) measures guiding the noise addition process should capture the semantics of nominal data in order to properly preserve the meaning of noise-added data.

By applying the notion of *centroid* of a sample of discrete values [46], and by using the semantic distance discussed above, we define the *semantic mean* of a nominal attribute as follows.

Definition 3. The *semantic mean* of a nominal attribute *A*, denoted by *sMean*(*A*), is the concept *c* from $\tau(D(A))$ that minimizes the sum of the semantic distances with respect to all a_i in *A*.

$$sMean(A) = \underset{c \in \tau(D(A))}{\arg\min} \left(\sum_{a_i \in A} sd(c, a_i) \right) \tag{13}$$

With this definition, any concept in $\tau(D(A))$ can be the mean of the attribute, regardless whether it was present in *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.

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* [47]. 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 4. The *semantic variance* of a nominal attribute *A*, denoted by *sVar*(*A*), is the average of squared semantic distances between each concept a_i in *A* and the semantic mean *sMean*(*A*).

$$sVar(A) = \frac{\sum_{a_i \in A} \left(sd\left(a_i, sMean(A)\right)\right)^2}{n}, \tag{14}$$

where *n* is the number of values in *A*.

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 [48] 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 (see equation (8)), 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 [49, 50]; 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 A and B be two nominal attributes of a dataset X. If their samples have n records, we obtain the following set of value pairs $(A,B) = \{(a_i,b_i)\text{: i} = 1, \ldots, \text{n}\}$, where the pair (a_i,b_i) represents the value of the attributes A and B for the record i. For example, in a dataset X of n patients from a hospital, where the attribute A stores diagnoses and the attribute B stores medical procedures, the pair (a_i,b_i) 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_A as the $(n\times n)$ *semantic distance matrix* of attribute $A = (a_1, \ldots, a_n)$ and SD_B as the matrix of attribute $B = (b_1, \ldots, b_n)$

$$SD_A = \left(sd_{ij}^A\right)_{i,j=1}^n , \quad SD_B = \left(sd_{ij}^B\right)_{i,j=1}^n , \tag{15}$$

where elements sd_{ij}^A and sd_{ij}^B are semantic distances. Thereby, $sd_{ij}^A = sd(a_i, a_j)$ is the semantic distance between the values of the attribute A in positions i and j. In line with the previous example, sd_{ij}^A would express the semantic distance between the main diagnoses of patients i and j. Analogously, sd_{ij}^B represents the semantic distance between the values of the attribute 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_A and SD_B have a zero diagonal because, as stated in Section 4.2 for the Wu and Palmer measure, 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_A and Δ_B be two $(n\times n)$ double centered semantic distance matrices whose elements δ_{ij}^A and δ_{ij}^B are computed from their respective matrices SD_A and SD_B as follows

$$\begin{aligned} \Delta_A &= \left(\delta_{ij}^A\right)_{i,j=1}^n = \left(sd_{ij}^A - \overline{sd}_{i.}^A - \overline{sd}_{.\text{j}}^A + \overline{sd}_{..}^A\right)_{i,j=1}^n \\ \Delta_B &= \left(\delta_{ij}^B\right)_{i,j=1}^n = \left(sd_{ij}^B - \overline{sd}_{i.}^B - \overline{sd}_{.\text{j}}^B + \overline{sd}_{..}^B\right)_{i,j=1}^n \end{aligned} \tag{16}$$

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

$$\overline{sd}_{i.}^A = \frac{1}{n}\sum_{j=1}^n sd_{ij}^A, \qquad \overline{sd}_{.\text{j}}^A = \frac{1}{n}\sum_{i=1}^n sd_{ij}^A, \qquad \overline{sd}_{..}^A = \frac{1}{n^2}\sum_{ij=1}^n sd_{ij}^A \tag{17}$$

Note that, when i is equal to j, $\overline{sd}_{i.}^{A}$ is equal to $\overline{sd}_{.j}^{A}$ by the commutative property of the semantic distance measure. Analogously, $\overline{sd}_{i.}^{B}$ is the mean of i-th row from matrix SD_B, $\overline{sd}_{.j}^{B}$ is the mean of j-th column from matrix SD_B and $\overline{sd}_{..}^{B}$ is the mean of all values from matrix SD_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 5. The *semantic distance covariance* between two nominal attributes A and B, denoted by $sdCov(A,B)$, is the square root of the arithmetic mean of the product $\delta_{ij}^{A}\delta_{ij}^{B}$.

$$sdCov(A,B) = \frac{1}{n}\sqrt{\sum_{i,j=1}^{n}\delta_{ij}^{A}\delta_{ij}^{B}} \tag{18}$$

According to [48], the distance covariance satisfies $sdCov(A,B) \geq 0$. Further, $sdCov(A,B) = 0$ if and only if A and B are independent. This property is a consequence of dealing with centered distances and allows measuring nonlinear associations.

Definition 6. The *semantic distance correlation* between two nominal attributes A and B, denoted by $sdCor(A,B)$, is the nonnegative number obtained by dividing the distance covariance by the product of the distance standard deviations of the attributes.

$$sdCor(A,B) = \begin{cases} \dfrac{sdCov(A,B)}{\sqrt{sdVar(A)\ sdVar(B)}}, & sdVar(A)\,sdVar(B) > 0 \\ 0, & sdVar(A)\,sdVar(B) = 0 \end{cases} \tag{19}$$

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

$$\begin{aligned} sdVar(A) &= sdCov(A,A) = \frac{1}{n}\sqrt{\sum_{i,j=1}^{n}\delta_{ij}^{A}\delta_{ij}^{A}} \\ sdVar(B) &= sdCov(B,B) = \frac{1}{n}\sqrt{\sum_{i,j=1}^{n}\delta_{ij}^{B}\delta_{ij}^{B}} \end{aligned} \tag{20}$$

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

4.4. Semantic uncorrelated noise addition method

By means of the distance measure discussed in Section 4.2, which enables us to semantically compare nominal values while being consistent with the noise distribution, and the semantically-grounded versions of the mean and variance measures proposed in Section 4.3, we can adapt the numerical uncorrelated noise addition method to the semantic domain of nominal data. 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 A through the uncorrelated noise addition method are shown in Figure 2. Firstly, the taxonomy $\tau(D(A))$ associated with the domain of attribute A is obtained from the ontology O, as described in Section 4.1. Thereby, we delimit the set of concepts from O that are candidates to replace the original values during the noise-addition process.

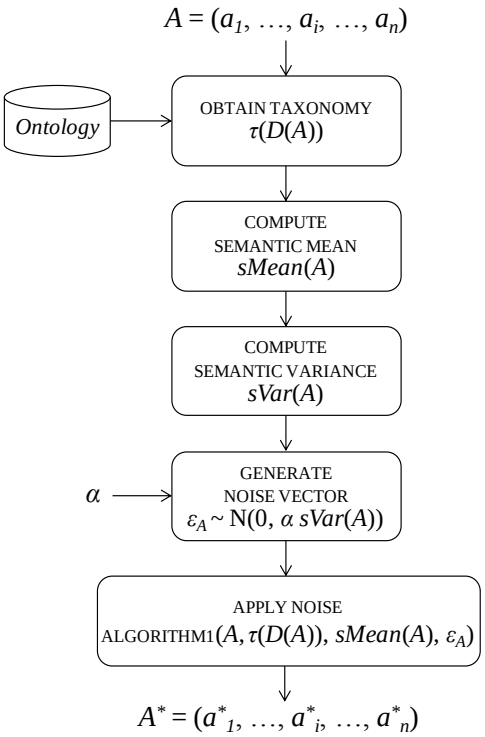

Figure 2. *Semantic uncorrelated noise addition* method for a nominal attribute *A*.

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 $Var(\varepsilon_A) = \alpha\ sVar(A)$. To compute the semantic variance $sVar(A)$ of the attribute A, we use equation (14), which requires the calculation of the semantic mean $sMean(A)$ by equation (13). After that, the noise sequence consisting of a vector of $n = |A|$ random numbers $\varepsilon_A = (\epsilon_{a_1}, \ldots, \epsilon_{a_i}, \ldots, \epsilon_{a_n})$, which follows a normal distribution $\varepsilon_A \sim N(0, \alpha\ sVar(A))$ with mean 0 and variance $\alpha\ sVar(A)$, is generated. Finally, after obtaining ε_A, the error values ϵ_{a_i} are applied to the original values a_i of attribute 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, a_i, and their noisy versions, a_i^*. 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., $sd(a_i, a_i^*) = |\epsilon_{a_i}|$. 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.

$$a_i^* = \underset{c \in \tau(D(A))}{\arg\min} \left\{ sd(c, a_i) \mid sd(c, a_i) \geq \left| \epsilon_{a_i} \right| \right\} \tag{21}$$

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 4.3, 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 is positive, the concept c in $\tau(D(A))$ that will replace the original value a_i must be farther from $sMean(A)$ than a_i, i.e., $sd(c, sMean(A)) > sd(a_i, sMean(A))$. For example, by applying this condition to the original nominal value a_i=*Ulcerative colitis* in Figure 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 a_i.
- If the error ϵ_i is negative, the concept c in $\tau(D(A))$ that will replace the original value a_i must be closer to $sMean(A)$ than a_i, i.e., $sd(c, sMean(A)) < sd(a_i, sMean(A))$. 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 4.

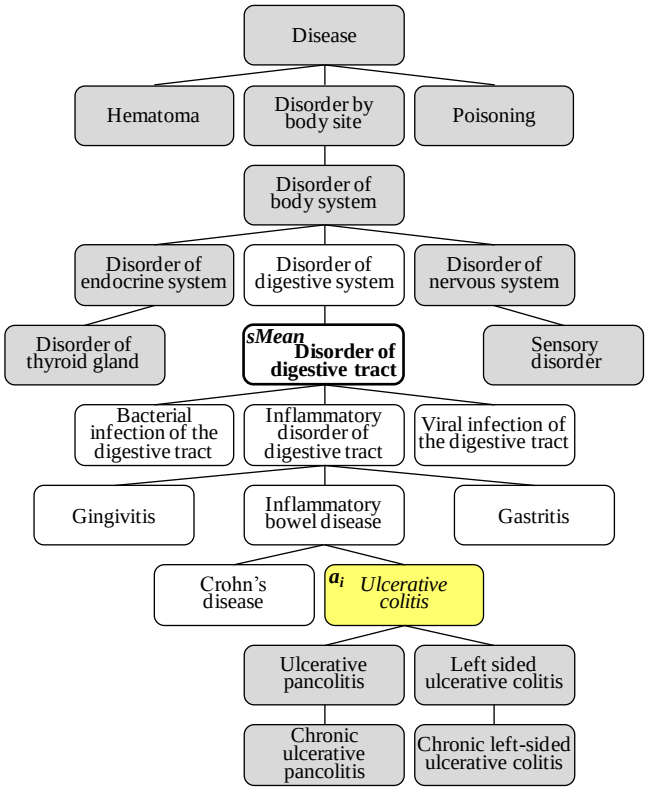

Figure 3. Example of replacement candidates (gray-shaded concepts) for an original value a_i (*Ulcerative colitis*) when the error sign is positive.

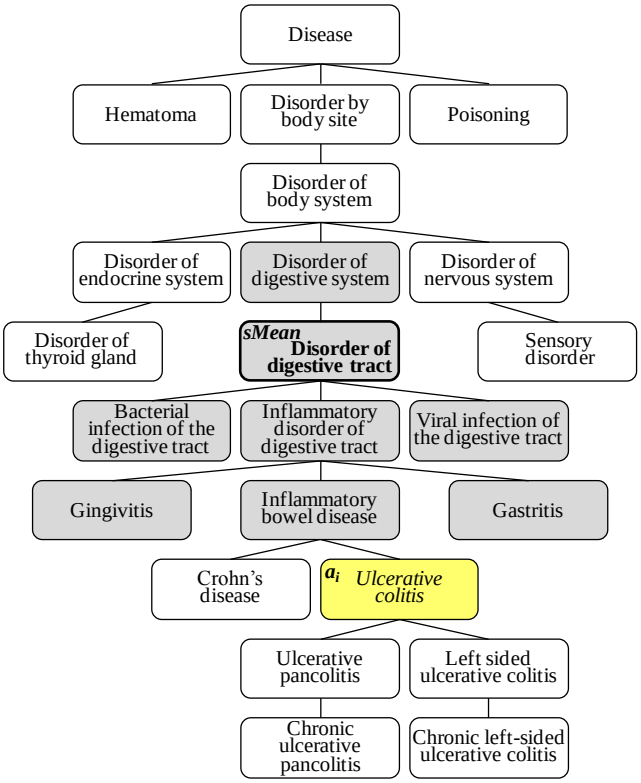

Figure 4. Example of replacement candidates (gray-shaded concepts) for an original value a_i (*Ulcerative colitis*) when the error sign is negative.

Finally, our method will select as noise-added value a_i^* the candidate concept c that best approximates the error magnitude $|\epsilon_{a_i}|$ according to equation (21). 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 A is shown in Algorithm1. Together with A, $\tau(D(A))$ and the noise vector ε_A, $sMean(A)$ is passed as input parameter to the algorithm because it is necessary to balance value replacements with respect to the semantic mean of A. In order to select the noise-added values a_i^*, we apply the noise magnitude $|\epsilon_{a_i}|$ to each original value a_i by replacing it by a concept c in $\tau(D(A))$ that ideally matches the error magnitude or that, while exceeding the error magnitude, minimizes its distance from a_i (lines 7 and 9). At this step, the interpretation of the error sign proposed above is used: line 7 when ϵ_{a_i} is positive and line 9 when ϵ_{a_i} is negative; if ϵ_{a_i} is zero, the noise-added value a_i^* is exactly a_i (line 3). Finally, when a_i matches $sMean(A)$, the noise-added value a_i^* will simply be the concept c in $\tau(D(A))$ that ideally matches the error magnitude or that, while exceeding the error magnitude, minimizes its distance from a_i (line 5). In any case, if no concept c in $\tau(D(A))$ with $sd(c, a_i) \geq |\epsilon_{a_i}|$ exists, i.e., we cannot get further enough within $\tau(D(A))$, 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.

Algorithm1. Method to apply the noise vector to an attribute *A* by using the *mean* concept as reference point in the replacements.

Input :

A: nominal attribute with n records

$\tau(D(A))$: taxonomy associated with the domain of A //according to definition (2)

$sMean(A)$: semantic mean of A //according to equation (12)

ε_A: noise vector //according to Section 4.4

Output :

A^* : noise-added nominal attribute

1: **for all** a_i in A **do**

2: **if** $\epsilon_{a_i} = 0$ **then**

3: $a_i^* \leftarrow a_i$

4: **else if** $a_i = sMean(A)$ **then**

5: $a_i^* \leftarrow \underset{c \in \tau(D(A))}{\arg\min} \left\{ sd(c, a_i) \mid sd(c, a_i) \geq \left| \epsilon_{a_i} \right| \right\}$

6: **else if** ϵ_{a_i} is positive **then**

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

8: **else if** ϵ_{a_i} is negative **then**

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

10: **end if**

11: **end for**

12: **return** 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 3, to add noise to a multivariate dataset with *p* nominal attributes through uncorrelated noise, Algorithm1 must be applied to each attribute independently. Therefore, correlation among attributes will not be preserved.

4.5. Semantic correlated noise addition method

By relying on the semantically-grounded versions of the distance covariance and distance correlation measures proposed in Section 4.3, we can also adapt the numerical correlated noise addition method detailed in Section 3 to the semantic domain. 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 *A* and *B* with *n* records, whose values refer to concepts modeled in an ontology *O*.

As shown in Figure 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 3, the generated noise consists of a (*n*×*2*) matrix of random numbers $\varepsilon_{A,B} = \{(\epsilon_{a_1}, \epsilon_{b_1}), ..., (\epsilon_{a_i}, \epsilon_{b_i}), ..., (\epsilon_{a_n}, \epsilon_{b_n})\}$ that follows a multivariate normal distribution $\varepsilon_{A,B} \sim N(0, \alpha\Sigma_{A,B})$ with mean the vector 0 and covariance matrix $\alpha\Sigma_{A,B}$, where the parameter α determines the desired degree of semantic noise and $\Sigma_{A,B}$ represents the semantic covariance matrix of the attributes. In the semantic domain, $\Sigma_{A,B}$ 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 (20) and (18) respectively.

$$\Sigma_{A,B} = \begin{pmatrix} sdVar(A) & sdCov(A,B) \\ sdCov(B,A) & sdVar(B) \end{pmatrix} \quad (22)$$

Finally, as shown in the last step depicted in Figure 5, the noise vectors ε_A and ε_B from $\varepsilon_{A,B}$ are applied to the attributes *A* and *B*, respectively. To do this, we propose three noise addition strategies, which are detailed below.

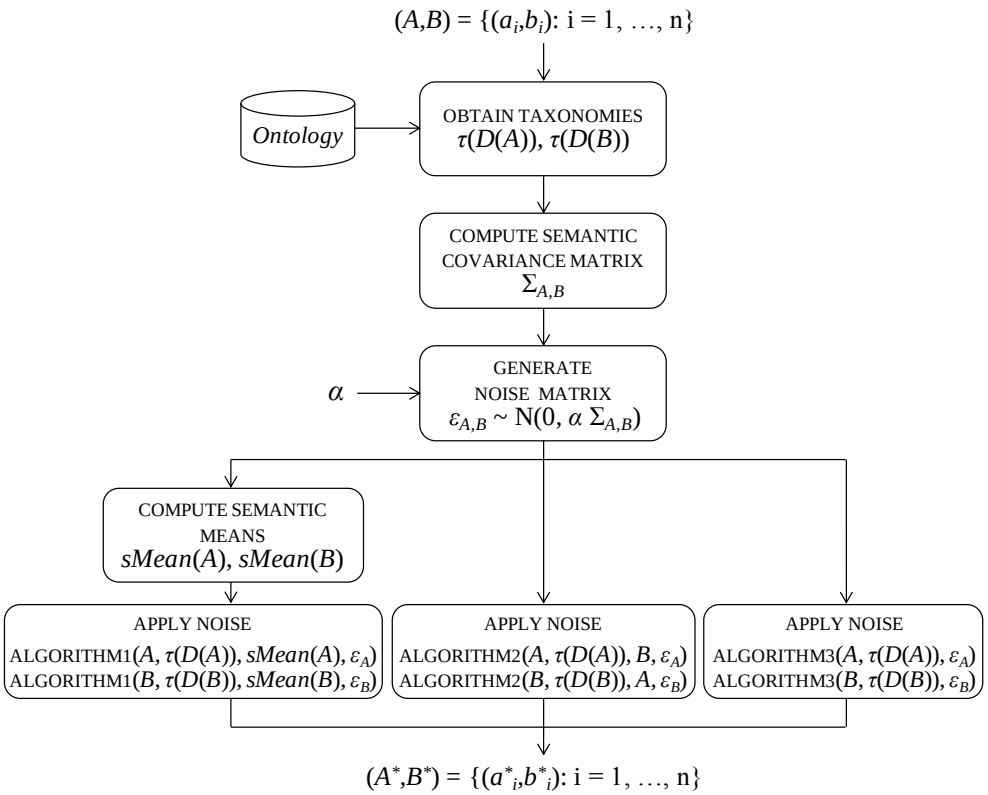

Figure 5. *Semantic correlated noise addition* method for two nominal attributes *A* and *B*.

The first approach follows the strategy detailed in 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 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 *A* influence attribute *B*, and vice versa. Specifically, to preserve the correlation of the data (a_i,b_i), if a positive error is added to an original numerical value $a_i > b_i$, the noisy value will be further away from b_i at the same magnitude; on the other hand, if the error is negative, the new value will get closer to b_i. 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 b_i corresponding to the value a_i that is being replaced, and vice versa.

Formally, as shown in Algorithm2, if the error ϵ_{a_i} is positive, the concept c in $\tau(D(A))$ that will replace the original value a_i must be farther from b_i than a_i, i.e., $sd(c, b_i) > sd(a_i, b_i)$, and vice

versa. Otherwise, if the error ϵ_{a_i} is negative, the concept *c* must be closer to b_i than a_i, i.e., $sd(c, b_i) < sd(a_i, b_i)$, and vice versa. Understandably, both attributes must belong to the same semantic domain, i.e., $\tau(D(A)) = \tau(D(B))$. For each attribute Algorithm2 will be called instead of Algorithm1 in the last step depicted in Figure 5.

Algorithm2. Method to apply the noise vector to an attribute *A* by using the values of the attribute *B* as reference points in the replacements.

Input :
A, B : nominal attributes with *n* records
$\tau(D(A))$: taxonomy associated with the domain of *A*
ε_A: noise vector
Output :
A^* : noise-added nominal attribute

1: **for all** a_i in *A* **do**
2: **if** $\epsilon_{a_i} = 0$ **then**
3: $a_i^* \leftarrow a_i$
4: **else if** ϵ_{a_i} is positive **then**
5: $a_i^* \leftarrow \underset{c \in \tau(D(A))}{\arg\min} \left\{ sd(c, a_i) \mid sd(c, a_i) \geq \left| \epsilon_{a_i} \right| \wedge sd(c, b_i) > sd(a_i, b_i) \right\}$
6: **else if** ϵ_{a_i} is negative **then**
7: $a_i^* \leftarrow \underset{c \in \tau(D(A))}{\arg\min} \left\{ sd(c, a_i) \mid sd(c, a_i) \geq \left| \epsilon_{a_i} \right| \wedge sd(c, b_i) < sd(a_i, b_i) \right\}$
8: **end if**
9: **end for**
10: **return** A^*

A solution for attributes belonging to different semantic domains, i.e., $\tau(D(A)) \neq \tau(D(B))$, for example, $\tau(D(A))$={*diseases*} and $\tau(D(B))$={*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 Algorithm3.

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

Input :
A: nominal attribute with n records
$\tau(D(A))$: taxonomy associated with the domain of A
ε_A: noise vector
Output :
A^*: noise-added nominal attribute
1: **for all** a_i in A **do**
2: **if** $\epsilon_{a_i} = 0$ **then**
3: $a_i^* \leftarrow a_i$
4: **else if** ϵ_{a_i} is positive **then**
5: $a_i^* \leftarrow \underset{c \in \tau(D(A))}{\arg\min} \left\{ sd(c,a_i) \mid sd(c,a_i) \geq \left|\epsilon_{a_i}\right| \wedge sd(c,root) > sd(a_i,root) \right\}$
6: **else if** ϵ_{a_i} is negative **then**
7: $a_i^* \leftarrow \underset{c \in \tau(D(A))}{\arg\min} \left\{ sd(c,a_i) \mid sd(c,a_i) \geq \left|\epsilon_{a_i}\right| \wedge sd(c,root) < sd(a_i,root) \right\}$
8: **end if**
9: **end for**
10: **return** 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 [51-53] 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 (Algorithms 1 and 3) 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, 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 $X = (A,B,C,D) = \{(a_i,b_i,c_i,d_i)$: i = 1, …, n$\}$ be a dataset with four nominal attributes; there are three options to apply Algorithm2: pairs (*A*,*B*) and (*C*,*D*); pairs (*A*,*C*) and (*B*,*D*), and pairs (*A*,*D*) and (*B*,*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 potential 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.

5. Empirical study

In this section, we evaluate the semantic noise addition methods we propose in Section 4 with several nominal datasets and under different evaluation metrics; we also compare their results with two non-semantic data perturbation methods based on the distribution of the 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[2]. Each record of the database details the healthcare discharge of a patient and, among others, contains several nominal attributes stating the *principal diagnosis*, *secondary diagnosis* and the *medical procedure* applied to the patient, which we selected as evaluation data. The database is especially suitable to illustrate the need for semantic noise addition methods, because nominal discharge patient data can be used as input for machine learning algorithms in healthcare research, for which noise addition may contribute to avoid overfitting. Moreover, because comorbidity of diagnoses may produce rare or even unique combinations of diseases that may disclose the identity of certain patients, noise addition can also be used to distort values and protect patients' privacy [54]. Finally, because correlations between medical attributes are crucial for research, noise addition methods should preserve them as much as possible in order to retain the analytical utility of noise-added data.

Diagnosis and procedure codes have been mapped to healthcare concepts in the SNOMED-CT medical ontology[3], which is especially well-suited to assist semantic similarity assessments of medical-related data because of its large size and taxonomic detail: it contains more than 311,000 unique concepts organized in 18 overlapping hierarchies with more than 1.36 million relationships.

To quantify the accuracy of the noise-added results from the perspective of data utility preservation, we have considered the following semantic features and evaluation metrics:

1. The *semantic mean* of the input attribute A, $sMean(A)$, and of the noise-added attribute A^*, $sMean(A^*)$, by using equation (13), and the *semantic distance* between both, $sd(sMean(A^*), sMean(A))$, by using equation (12). Distance=0 indicates that the mean has been perfectly preserved after the noise-addition process.
2. The *semantic variance* of input and noise-added attributes, $sVar(A)$ and $sVar(A^*)$, by using equation (14), and the absolute difference between the actual *semantic variance*

[2] http://www.oshpd.ca.gov/HID/Products/PatDischargeData/PublicDataSet/index.html
[3] https://www.nlm.nih.gov/healthit/snomedct/index.html

of the noise-added attribute values and the expected *semantic variance* after adding noise with a noise parameter α, i.e., $|sVar(A^*) - (1+\alpha)\, sVar(A)|$. Differences near 0 indicate that the variance of the noise-added results has been well-controlled.

3. The *root mean square error* (*RMSE*), measured as the root average square *semantic distance* between original and noise-added value pairs, $RMSE_{Actual}(A)$. This measures the overall loss of semantics in the noise-added values with respect to the original ones, which should be similar to the *target error* defined by the desired magnitude of the noise to be added, $RMSE_{Target}(A)$.
4. The *semantic distance covariance* of input and noise-added attribute pairs, $sdCov(A,B)$ and $sdCov(A^*,B^*)$, by using equation (18), and the absolute difference between the actual *semantic distance covariance* of pairs of noise-added attributes and the expected *semantic distance covariance* after adding noise with a noise parameter α, i.e., $|sdCov(A^*,B^*) - (1+\alpha)\, sdCov(A,B)|$. Differences around 0 indicate that covariances of noise-added attributes have been well preserved.
5. The *semantic distance correlation* of input and noise-added attribute pairs, $sdCor(A,B)$ and $sdCor(A^*,B^*)$, by using equation (19), and the absolute difference between the actual *semantic distance correlation* of pairs of noise-added attributes and original attributes, i.e., $|sdCor(A^*,B^*) - sdCor(A,B)|$. Difference=0 indicates that the correlation has been perfectly preserved after the noise-addition process.

The first experiment was carried out with a pair of strongly correlated attributes *A=principal diagnosis* and *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 1.

Table 1. Semantic features of *Dataset1*: 1,350 patients with two strongly correlated attributes *A= principal diagnosis* and *B=secondary diagnosis*, both with the same associated taxonomy.

Semantic feature	Value
$sMean(A)$	Furuncle of chest wall
$sMean(B)$	Viral hepatitis with hepatic coma
$sVar(A)$	0.22
$sVar(B)$	0.24
$sdCov(A,B)$	0.26
$sdCor(A,B)$	0.94

The fact that the attributes are strongly correlated, $sdCor(A,B) = 0.94$, 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 4.4 with the noise-addition strategy defined in Algorithm1 (*Uncorrelated-Algorithm1*), the correlated method discussed in Section 4.5 with the noise-addition strategy defined in Algorithm1 (*Correlated-Algorithm1*) and the correlated method with the noise-addition strategy designed to

optimize the preservation of the correlation between attributes defined in Algorithm2 (*Correlated-Algorithm2*), since both attributes are drawn from the same taxonomy. Tables 2, 3 and 4 collect the evaluation metrics of the results provided by these methods for several usual values of the noise parameter α={0.1, 0.3, 0.5, 1}.

Table 2. Evaluation metrics of the noise-added dataset obtained with *Uncorrelated-Algorithm1* for *Dataset1* (*A= principal diagnosis* and *B=secondary diagnosis*).

Metric	**α=0.1**	**α=0.3**	**α=0.5**	**α=1**
$sMean(A^*)$	Blister of axilla with infection	Axillary hydra-denitis suppurativa	Blister of axilla with infection	Blister of axilla with infection
$sd(sMean(A^*), sMean(A))$	0.20	0.20	0.20	0.20
$sVar(A^*) \mid \lvert sVar(A^*) - (1+\alpha)\, sVar(A)\rvert$	0.24 | 0	0.24 | 0.05	0.25 | 0.08	0.27 | 0.17
$RMSE_{Actual}(A) \mid RMSE_{Target}(A)$	0.23 | 0.15	0.31 | 0.25	0.38 | 0.34	0.49 | 0.48
$sMean(B^*)$	Viral hepatitis with hepatic coma	Viral hepatitis with hepatic coma	Viral hepatitis with hepatic coma	Inflammatory disease of liver
$sd(sMean(B^*), sMean(B))$	0	0	0	0.18
$sVar(B^*) \mid \lvert sVar(B^*) - (1+\alpha)\, sVar(B)\rvert$	0.23 | 0.03	0.24 | 0.07	0.26 | 0.1	0.30 | 0.18
$RMSE_{Actual}(B) \mid RMSE_{Target}(B)$	0.19 | 0.15	0.30 | 0.27	0.37 | 0.35	0.50 | 0.51
$sdCov(A^*,B^*) \mid \lvert sdCov(A^*,B^*) - (1+\alpha)\, sdCov(A,B)\rvert$	0.15 | 0.14	0.11 | 0.23	0.09 | 0.30	0.08 | 0.44
$sdCor(A^*,B^*) \mid \mid sdCor(A^*,B^*) - sdCor(A,B)\rvert$	0.78 | 0.16	0.65 | 0.29	0.57 | 0.37	0.38 | 0.56

Table 3. Evaluation metrics of the noise-added dataset obtained with *Correlated-Algorithm1* for *Dataset1* (*A= principal diagnosis* and *B=secondary diagnosis*).

Metric	**α=0.1**	**α=0.3**	**α=0.5**	**α=1**
$sMean(A^*)$	Blister of axilla with infection	Blister of axilla with infection	Blister of axilla with infection	Blister of axilla with infection
$sd(sMean(A^*), sMean(A))$	0.20	0.20	0.20	0.20
$sVar(A^*) \mid \lvert sVar(A^*) - (1+\alpha)\, sVar(A)\rvert$	0.24 | 0	0.24 | 0.05	0.26 | 0.07	0.29 | 0.15
$RMSE_{Actual}(A) \mid RMSE_{Target}(A)$	0.24 | 0.17	0.33 | 0.28	0.40 | 0.37	0.51 | 0.52
$sMean(B^*)$	Viral hepatitis with hepatic coma	Inflammatory disease of liver	Mouth-gen. ulcers inflam. cartil. synd.	Mouth-gen. ulcers inflam. cartil. synd.
$sd(sMean(B^*), sMean(B))$	0	0.18	0.45	0.45
$sVar(B^*) \mid \lvert sVar(B^*) - (1+\alpha)\, sVar(B)\rvert$	0.22 | 0.04	0.23 | 0.08	0.24 | 0.12	0.28 | 0.2
$RMSE_{Actual}(B) \mid RMSE_{Target}(B)$	0.20 | 0.17	0.30 | 0.28	0.38 | 0.37	0.49 | 0.52
$sdCov(A^*,B^*) \mid \lvert sdCov(A^*,B^*) - (1+\alpha)\, sdCov(A,B)\rvert$	0.15 | 0.14	0.12 | 0.22	0.10 | 0.29	0.09 | 0.43
$sdCor(A^*,B^*) \mid \mid sdCor(A^*,B^*) - sdCor(A,B)\rvert$	0.82 | 0.12	0.73 | 0.21	0.69 | 0.25	0.59 | 0.35

Table 4. Evaluation metrics of the noise-added dataset obtained with *Correlated-Algorithm2* for *Dataset1* (*A= principal diagnosis* and *B=secondary diagnosis*).

Metric	**α=0.1**	**α=0.3**	**α=0.5**	**α=1**
$sMean(A^*)$	Blister of axilla with infection	Blister of axilla with infection	Blister of axilla with infection	Granuloma inguinale
$sd(sMean(A^*), sMean(A))$	0.20	0.20	0.20	0.20
$sVar(A^*) \mid \lvert sVar(A^*) - (1+\alpha)\, sVar(A)\rvert$	0.24 | 0	0.26 | 0.03	0.28 | 0.05	0.32 | 0.12
$RMSE_{Actual}(A) \mid RMSE_{Target}(A)$	0.25 | 0.17	0.35 | 0.28	0.41 | 0.37	0.52 | 0.52
$sMean(B^*)$	Inflammatory disease of liver	Inflammatory disease of liver	Inflammatory disease of liver	Mouth-gen. ulcers inflam. cartil. synd.
$sd(sMean(B^*), sMean(B))$	0.18	0.18	0.18	0.45
$sVar(B^*) \mid \lvert sVar(B^*) - (1+\alpha)\, sVar(B)\rvert$	0.23 | 0.03	0.25 | 0.06	0.27 | 0.09	0.31 | 0.17
$RMSE_{Actual}(B) \mid RMSE_{Target}(B)$	0.22 | 0.17	0.32 | 0.28	0.40 | 0.37	0.51 | 0.52
$sdCov(A^*,B^*) \mid \lvert sdCov(A^*,B^*) - (1+\alpha)\, sdCov(A,B)\rvert$	0.16 | 0.13	0.13 | 0.21	0.11 | 0.28	0.09 | 0.43
$sdCor(A^*,B^*) \mid \mid sdCor(A^*,B^*) - sdCor(A,B)\rvert$	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. The actual covariance between the two attributes follows a pattern similar to the actual variance.

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-Algorithm1* provides a slightly better mean than *Correlated-Algorithm2*, because *Correlated-Algorithm1* has been designed to optimize this feature. On the contrary, the correlation is better preserved by *Correlated-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 2, in Figures 6-8, we compare the accuracy of our algorithms with respect to 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.

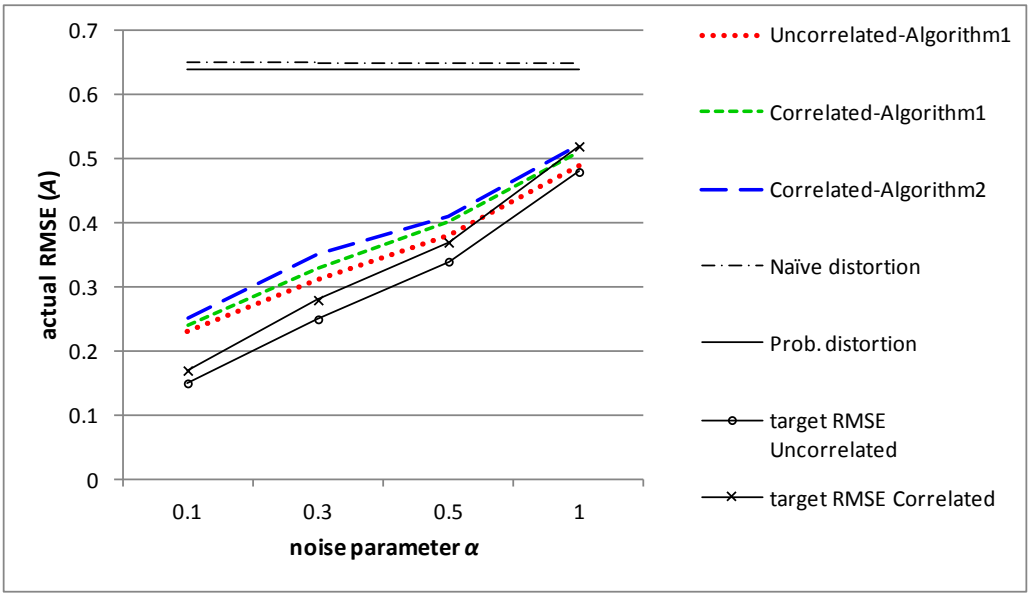

Figure 6. Evaluation of the actual RMSE of attribute *A* for the *naïve distortion*, *probabilistic distortion* and our semantic methods (*Uncorrelated-Algorithm1*, *Correlated-Algorithm1*, *Correlated-Algorithm2*) in *Dataset1*.

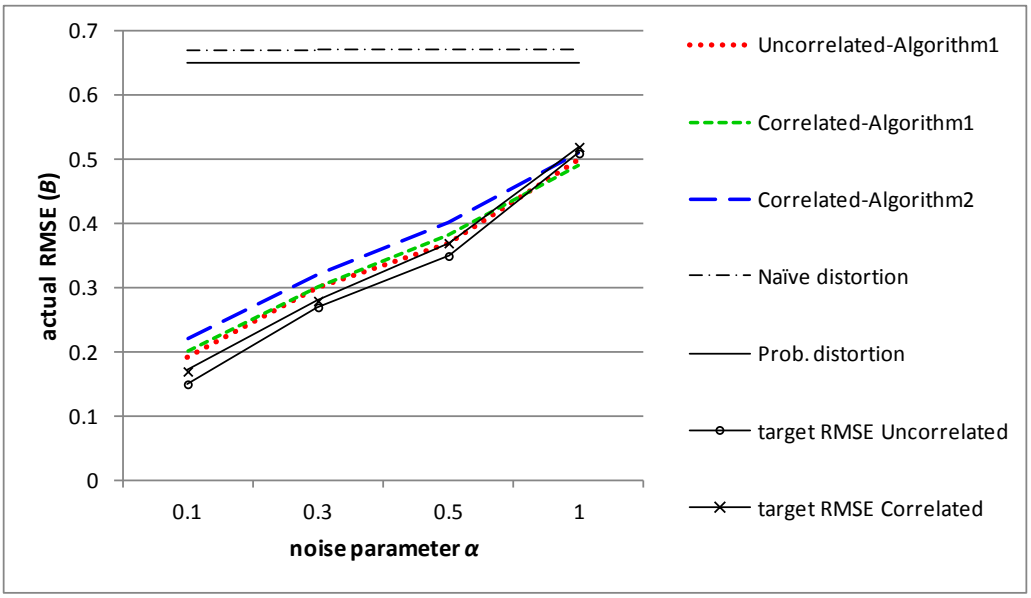

Figure 7. Evaluation of the actual RMSE of attribute *B* for the *naïve distortion*, *probabilistic distortion* and our semantic methods (*Uncorrelated-Algorithm1*, *Correlated-Algorithm1*, *Correlated-Algorithm2*) in *Dataset1*.

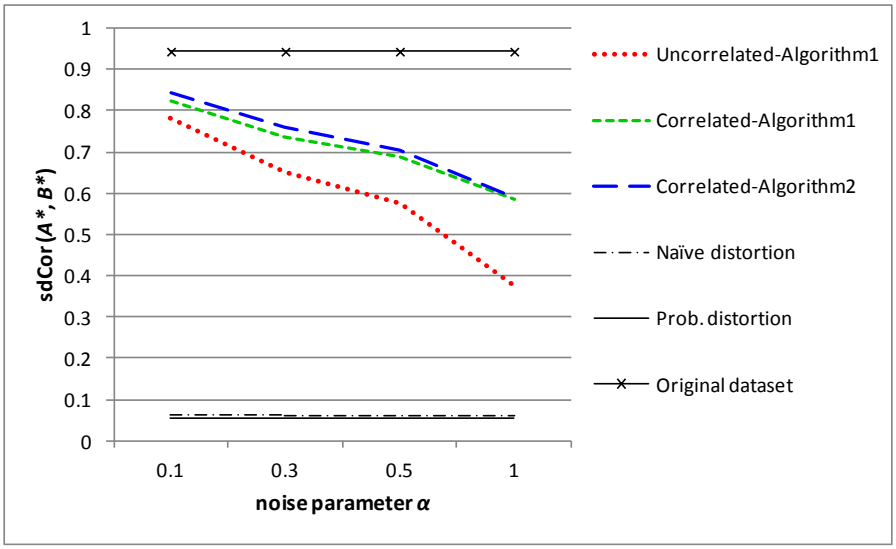

Figure 8. Evaluation of the *semantic distance correlation* for the *naïve distortion*, *probabilistic distortion* and our semantic methods (*Uncorrelated-Algorithm1*, *Correlated-Algorithm1*, *Correlated-Algorithm2*) in *Dataset1*.

However, distribution-based data distortions are able to preserve the mean better in some cases, as shown Table 5. This is due to the small spectrum of different categories in the dataset, that is, only 17 different diseases for attribute *A* and 19 for *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 5. Evaluation of the *semantic mean* for the *naïve distortion*, *probabilistic distortion* and our semantic methods (*Uncorrelated-Algorithm1*, *Correlated-Algorithm1*, *Correlated-Algorithm2*) in *Dataset1*.

Metric	Naïve distortion	Probabilistic distortion	Uncorrelated Algorithm1 (α=1)	Correlated Algorithm1 (α=1)	Correlated Algorithm2 (α=1)
$sd(sMean(A^*), sMean(A))$	0.20	0.04	0.20	0.20	0.20
$sd(sMean(B^*), sMean(B))$	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: *A=principal diagnosis*, which is associated with the taxonomy of *diseases* and *B=medical procedure*, which is associated with the taxonomy of *procedures*, both from SNOMED-CT. This allows us to compare *Correlated-Algorithm1* with the version designed to optimize the preservation of the correlation between attributes with domains in different taxonomies, *Correlated-Algorithm3*. Table 6 depicts the semantic features of *Dataset2*, which also shows a strong correlation of 0.87.

Table 6. Semantic features of *Dataset2*: 1,316 patients with two strongly correlated attributes, *A= principal diagnosis, B= medical procedure* with different associated taxonomies.

Semantic feature	Value
$sMean(A)$	Malignant neoplasm of costovertebral joint
$sMean(B)$	Arthrectomy of hip
$sVar(A)$	0.15
$sVar(B)$	0.07
$sdCov(A,B)$	0.19
$sdCor(A,B)$	0.88

Evaluation metrics for the results provided by *Correlated-Algorithm1* and *Correlated-Algorithm3* with *Dataset2* are shown in Tables 7 and 8.

Table 7. Evaluation metrics of the noise-added dataset provided by *Correlated-Algorithm1* for *Dataset2* (*A=principal diagnosis* and *B=medical procedure*).

Metric	α=0.1	α=0.3	α=0.5	α=1
$sMean(A^*)$	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
$sd(sMean(A^*), sMean(A))$	0.33	0	0	0
$sVar(A^*) \mid \lvert sVar(A^*) - (1+\alpha)\, sVar(A)\rvert$	0.18 | 0.02	0.19 | 0.01	0.2 | 0.03	0.23 | 0.07
$RMSE_{Actual}(A) \mid RMSE_{Target}(A)$	0.23 | 0.16	0.32 | 0.28	0.40 | 0.37	0.50 | 0.52
$sMean(B^*)$	Arthrectomy of hip	Arthrectomy of hip	Arthrectomy of hip	Arthrectomy of hip
$sd(sMean(B^*), sMean(B))$	0	0	0	0
$sVar(B^*) \mid \lvert sVar(B^*) - (1+\alpha)\, sVar(B)\rvert$	0.08 | 0	0.09 | 0	0.1 | 0.01	0.14 | 0
$RMSE_{Actual}(B) \mid RMSE_{Target}(B)$	0.15 | 0.13	0.24 | 0.23	0.30 | 0.29	0.41 | 0.40
$sdCov(A^*,B^*) \mid \lvert sdCov(A^*,B^*) - (1+\alpha)\, sdCov(A,B)\rvert$	0.11 | 0.10	0.08 | 0.17	0.07 | 0.22	0.06 | 0.32
$sdCor(A^*,B^*) \mid \lvert sdCor(A^*,B^*) - sdCor(A,B)\rvert$	0.68 | 0.20	0.56 | 0.32	0.48 | 0.40	0.42 | 0.46

Table 8. Evaluation metrics of the noise-added dataset provided by *Correlated-Algorithm3* for *Dataset2* (*A=principal diagnosis* and *B=medical procedure*).

Metric	α=0.1	α=0.3	α=0.5	α=1
$sMean(A^*)$	Fracture of prox. end of femur	Fracture of prox. end of femur	Recurrent dislocation of joint	Recurrent dislocation of joint
$sd(sMean(A^*), sMean(A))$	0.23	0.23	0.23	0.23
$sVar(A^*) \mid \lvert sVar(A^*) - (1+\alpha)\, sVar(A)\rvert$	0.2 | 0.04	0.22 | 0.03	0.23 | 0.01	0.26 | 0.04
$RMSE_{Actual}(A) \mid RMSE_{Target}(A)$	0.27 | 0.16	0.34 | 0.28	0.41 | 0.37	0.51 | 0.52
$sMean(B^*)$	Arthrectomy of hip	Arthrectomy of hip	Arthrectomy of hip	Arthrectomy of hip
$sd(sMean(B^*), sMean(B))$	0	0	0	0
$sVar(B^*) \mid \lvert sVar(B^*) - (1+\alpha)\, sVar(B)\rvert$	0.08 | 0	0.11 | 0.02	0.13 | 0.03	0.2 | 0.06
$RMSE_{Actual}(B) \mid RMSE_{Target}(B)$	0.16 | 0.13	0.25 | 0.23	0.30 | 0.29	0.41 | 0.40
$sdCov(A^*,B^*) \mid \lvert sdCov(A^*,B^*) - (1+\alpha)\, sdCov(A,B)\rvert$	0.12 | 0.09	0.09 | 0.16	0.08 | 0.21	0.07 | 0.31
$sdCor(A^*,B^*) \mid \lvert sdCor(A^*,B^*) - sdCor(A,B)\rvert$	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-Algorithm1* better preserves the mean, while *Correlated-Algorithm3* provides a better-preserved correlation, despite using the same noise sequences in both methods.

6. Conclusions and future work

We have presented here 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 and thus, their analytical utility. In particular, we have proposed solutions for the two main families of noise addition methods: uncorrelated noise for individual attributes and correlated noise for multivariate datasets. To capture and manage the semantics underlying the nominal values to be distorted, our algorithms exploit the formal knowledge modeled in ontologies by means of ontology-based semantic similarity measures. Moreover, we have adapted the statistical operators used in the standard noise addition mechanism to the semantic domain, in order to be able to cope with the discrete, finite and non-ordinal nature of nominal data, and whereby data distortion is done consistently with their semantics. 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. 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. Finally, unlike perturbation methods based on the distribution of the data, our algorithms provide a configurable distortion level and thus, of controlling the information loss of the noise-added data. Our methods are general and can support different noise distributions such as Normal or Laplace.

After a comprehensive empirical study over several nominal multivariate datasets, we have found 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 perturbation 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 noise-added data that is useful for statistical analysis, in privacy preserving data mining or for artificial intelligence learning algorithms. As a summary and guide for practitioners and researchers, Table 9 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 9. Best suited methods according to the type of dataset and semantic feature to be optimized.

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

As future work, 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. We also plan to study 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. Finally, we plan to test the benefits of our approach in specific applications, such as machine learning or data protection methods.

Acknowledgements and disclaimer

We thank Josep Domingo-Ferrer for his useful comments. This work was partly supported by the European Commission (projects H2020-644024 “CLARUS” and H2020-700540 “CANVAS”), by the Spanish Government (projects TIN2014-57364-C2-R “SmartGlacis”, 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. M. Batet is supported by a Postdoctoral grant from Ministry of Economy and Competitiveness (MINECO) (FPDI-2013-16589).

References

[1] D. Marco, D.L. Neuhoff, The Validity of the Additive Noise Model for Uniform Scalar Quantizers, IEEE Transactions on Information Theory 51 (5) (2005) 1739-1755.

[2] G. Cao, Y. Zhao, R. Ni, B. Ou, Y. Wang, Forensic detection of noise addition in digital images, Journal of Electronic Imaging 23 (2) (2014).

[3] 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.

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

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

[6] A. Hundepool, J. Domingo-Ferrer, L. Franconi, S. Giessing, E.S. Nordholt, K. Spicer, P.-P. Wolf, Microdata, in: Statistical Disclosure Control, Wiley, 2012, pp. 23-130.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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).

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

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

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

[15] 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).

[16] 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.

[17] 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.

[18] 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.
[19] 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.
[20] 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.
[21] C. Dwork, Differential privacy, Automata, Languages and Programming 4052 (2006) 1-2.
[22] D. Abril, G. Navarro-Arribas, V. Torra, On the declassification of confidential documents, Modeling Decision for Artificial Intelligence 6820 (2011) 235-246.
[23] R. Conway, D. Strip, Selective partial access to a database, in, Cornell University, Tech. Rep., 1976.
[24] R. Brand, Microdata protection through noise addition, in: J. Domingo-Ferrer (Ed.) Inference Control in Statistical Databases, Springer Berlin Heidelberg, 2002, pp. 97-116.
[25] P. Tendick, Optimal noise addition for preserving confidentiality in multivariate data, Journal of Statistical Planning and Inference 27 (3) (1991) 341-353.
[26] K. Muralidhar, R. Sarathy, Security of random data perturbation methods, ACM Transactions on Database Systems 24 (2000) 487-493.
[27] 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.
[28] 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.
[29] P. Cimiano, Ontology Learning and Population from Text: Algorithms, Evaluation and Applications, Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006.
[30] M. Batet, A. Erola, D. Sánchez, J. Castellà-Roca, Utility preserving query log anonymization via semantic microaggregation, Information Sciences 242 (2013) 49-63.
[31] 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.
[32] M. Batet, D. Sánchez, A review on semantic similarity, in: Encyclopedia of Information Science and Technology, Third Edition, IGI Global, 2015, pp. 7575-7583.
[33] 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.
[34] 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.
[35] 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.
[36] 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.
[37] A. Tversky, Features of Similarity, Psychological Review 84 (4) (1977) 327-352.

[38] 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.
[39] 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.
[40] 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.
[41] 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.
[42] D. Lin, An Information-Theoretic Definition of Similarity, in: Proceedings of the 15th International Conference on Machine Learning, ICML 1998, 1998, pp. 296-304.
[43] 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.
[44] 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.
[45] 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.
[46] S. Martínez, A. Valls, D. Sánchez, Semantically-grounded construction of centroids for datasets with textual attributes, Knowledge-Based Systems 35 (2012) 160-172.
[47] 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.
[48] 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.
[49] 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.
[50] M. Omelka, Š. Hudecová, A comparison of the Mantel test with a generalised distance covariance test, Environmetrics 24 (7) (2013) 449-460.
[51] 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
[52] 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.
[53] M. Batet, D. Sanchez, A. Valls, K. Gibert, Semantic similarity estimation from multiple ontologies, Applied Intelligence 38 (1) (2013) 29-44.

[54] 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.