Utility-preserving privacy protection of nominal data sets via semantic rank swapping

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

Personal data are of great interest for research but, at the same time, they pose a serious privacy risk. Therefore, appropriate data protection measures should be undertaken by the data controller before making personal data available for secondary use. Also, such data protection should be done in a way that data are still useful for analysis. In the last years, a plethora of data protection mechanisms have been proposed. Among them, *rank swapping* is considered one of the best with respect to disclosure risk minimization and data utility preservation. Because rank swapping is based on sorting input data to swap values that are close to each other, in principle, it is a method restricted to numerical and ordinal categorical data. However, a significant amount of personal data currently compiled and used in data analysis are nominal, and their utility depends on the semantics they convey. To properly cope with this type of data, in this paper, we present rank swapping methods capable of protecting nominal data from a semantic perspective. Specifically, by exploiting ontologies, our methods are able to protect nominal data while properly preserving their semantics and, thus, their analytical utility. For that, we provide a suitable binary relation to semantically sort nominal data. Our proposal is capable of managing both independent individual attributes and non-independent multivariate data sets, being the latter especially relevant for data analysis. Empirical experiments carried on real clinical records and using a standard medical ontology show that our methods are able to preserve the semantic features of nominal data significantly better than standard permutation mechanisms.

Keywords: Rank swapping, nominal data, semantics, 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 977559657; Fax: +34 977 559710;
E-mail: mercedes.rodriguez@uca.es

1. Introduction

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

In this scenario, governmental agencies and current legislations on data protection, such as the General Data Protection Regulation (GDPR) [1], emphasize the need of adequately protecting Personally Identifiable Information (PII) [2] to preserve individuals' privacy. PII includes not only identifying data, such as social security numbers, but also any non-identifying data that, in combination with other non-identifying data, can be used by attackers to re-identify individuals by linking them with external data sources, as shown several studies [3-5]. These non-identifying attributes that, in aggregate, can be used to unequivocally re-identify individuals are known as *quasi-identifier attributes* and they cause real privacy threats. Quasi-identifiers are currently employed by data brokers to compile and aggregate individuals' data and, from these, build user profiles that are later used or sold to third parties for commercial and business purposes [6].

To minimize the chance of re-identification, quasi-identifying attributes should be subjected to *anonymization*. In turn, data anonymization should be done in a way that the protected data still retain as much analytical utility as possible, so that conclusions or inferences extracted from the analysis of the anonymized data set are similar to those of the original data set. For that, different masking methods have been proposed within the disciplines of Statistical Disclosure Control (SDC) [7] and Privacy-Preserving Data Publishing (PPDP) [8]. Among them, perturbative masking methods are the most widespread, which include *noise addition*, *microaggregation* or *rank swapping*. These mechanisms generate a modified version of the original data by distorting or introducing ambiguity on the quasi-identifying attributes while preserving certain statistical features. As shown in several studies [9, 10], *rank swapping* is considered one of the best perturbative mechanisms w.r.t. disclosure risk minimization and data utility preservation. This method, which is based on the idea of proximity swapping [11], ranks the values of each attribute in ascending order for later swapping each value with another one

randomly chosen within a restricted size range. Thus, the higher the range size, the higher the ambiguity in the re-identification inferences and the lower the disclosure risk; but also, the lower the data utility, because swapped values would tend to be less similar. Concerning data utility, and on the contrary to other data protection mechanisms [7], rank swapping perfectly preserves univariate statistics, such as the mean, the variance and the frequency distribution, because the values in the protected attribute are the same as those in the original attribute but permuted. For this same reason, rank swapping also preserves other very useful features for data analysis, such as data granularity or outlying values.

In addition to the above advantages, a recent study [12] has shown that any anonymization method is functionally equivalent to a permutation plus a small amount of noise; this turns the spotlight on the permutation-based data transformation implemented by the rank swapping mechanism as the essential principle underlying any data anonymization.

Because rank swapping relies on the ability to sort attribute values, it has been designed to deal with numerical attributes (e.g., income) and ordinal categorical attributes, i.e., data that admit order relationships (e.g., color, where the different colors may be ranked on basis of their wave lengths) [7]. However, a significant amount of personal data that are currently gathered by data brokers for categorizing individuals (e.g. from social networks, electronic healthcare records or web browsing logs) and that should be subject of anonymization, are of nominal nature [6]. Unlike other data types, *nominal categorical* attributes (e.g., occupation, race, religion, etc.) are finite, discrete, textual and non-ordinal; thus, they do not admit order relationships. In this scenario, in principle, it is not possible to carry out the sorting operation needed to rank the values of the data set during the permutation process. Moreover, because nominal data utility is closely related to the preservation of data semantics [13-15], any data transformation performed to anonymize nominal data, such as ranking, should consider the *meaning* of the attribute values. So far, only microaggregation and noise addition methods have been adapted to work with nominal data from a semantic perspective [16-19]. To do so, they exploit the formal semantics modeled in ontologies, which are knowledge structures that formally describe the concepts of a domain and the semantic relationships between them.

In this paper we present rank swapping methods capable of protecting nominal data from a semantic perspective. Our objective is twofold: (i) to provide a binary relation capable of semantically sorting nominal data by exploiting the formal semantics modeled in ontologies, and (ii) to provide mechanisms to control the degree of permutation in order to enforce a certain level of protection while preserving, as much as possible, the semantic features, and thus, the analytical utility of the data. In particular, we propose semantically-grounded rank swapping

solutions to perturb individual nominal attributes and multivariate nominal data sets. The latter is especially relevant because it is capable of protecting multivariate nominal data sets while reasonably preserving the correlation among attributes, which is of outmost importance for data analysis.

The rest of the paper is organized as follows. Section 2 discusses related works on permutation-based methods for data protection. Section 3 defines a suitable binary relation to semantically sort nominal data and presents our semantically-grounded rank swapping algorithms. Section 4 details the empirical experiments we carried out on real clinical records and by exploiting a standard medical ontology, and measures and compares the data utility preserved by our methods against several baselines. Section 5 contains the conclusions and depicts some lines of future research.

2. Related work

Various permutation-based methods have been proposed to protect data sets while preserving certain statistical features. The first one, named data swapping [20], is based on swapping the values of each attribute from a data set of t categorical attributes to yield a permuted data set whose t-order frequency counts, or t-order statistics, are the same as those of the original data set, i.e., a t-order equivalent data set. Since the t-order statistics are preserved, the inferences that derive from them are not altered. To distort a given attribute, the method builds all the equivalence classes for that attribute and randomly swaps the values within each class. Because of the way in which data swapping operates, this method is not suitable when most equivalence classes in the data set are composed of one or few records, which is the case of data sets with fine grained nominal attributes, since the swaps can hardly be carried out. On the other hand, if the data are released as microdata, it is necessary to add enough uncertainty on the true values of the individuals' data to reasonably protect their privacy. However, identifying a large number of swaps that preserve the t-order statistics is computationally impractical [21, 22]. As a feasible approach for the release of microdata, Reiss proposes in [21] a variation of data swapping where the t-order frequency counts are approximately preserved. Firstly, the method computes the relevant frequency tables from the original data set, and then constructs a new data set consistent with these tables. To do this, the values of an attribute are randomly selected according to the probability distribution derived from the original frequency tables; because this may produce values not appearing in the original data set, this makes it a synthetic method, rather than a strict data swapping one.

Because the above methods do not limit the swapping range, very different values may be swapped, thereby increasing the loss of utility. In order to limit the scope of the swaps and, therefore, maintain each permuted value within a certain rank-distance from the original one, Greenberg, in an unpublished manuscript [23] described by Moore in [11], presents *rank swapping*, a method initially defined for ordinal categorical data and subsequently applied to numerical data [9]. This method restricts the swapping range by ranking the values of the attribute in ascending order and, then, by swapping each value with another unswapped one randomly chosen within a user-defined range p, p being a percent of the total number of records. In this way, the rank of two swapped values cannot differ by more than p% of the total number of records. Large values of p lead to greater permutations whereas smaller values of p incur in higher disclosure risk. In [9], rank swapping is pointed out as the best performer data protection mechanism in terms of disclosure risk minimization and data utility preservation.

Rank swapping has been designed to deal with numerical or ordinal categorical data. In both cases, total orders are available to build the value ranks in which the algorithms rely on. However, nominal data are not ordinal and, thus, lack natural total orders. For such data, rank swapping has been considered either non-applicable [7, 24] or it has been suboptimally applied by defining artificial total orders (e.g., topological order of categorical labels for nominal attributes) [25] that, due to their lack of semantic coherence, may severely hamper the utility of the protected outcomes.

3. Semantic rank swapping methods

In this section, we propose semantically-grounded rank swapping mechanisms for nominal attributes. Because our approach is general, it accommodates several variations depending on how the swapping ranges are built. The first one intuitively adapts the idea of "swapping interval" of the standard rank swapping method to nominal data. In a second approach, we propose to dynamically build the swapping ranges to minimize the (semantic) information loss associated to each swap. Finally, we extend the method so that multivariate data sets are managed as non-independent attributes. The multivariate method constitutes in fact the core offering of our work, since it is able to offer *ex ante* privacy guarantees while better preserving the correlation between attributes than the univariate counterparts do.

Let assume that the input data X is a de-identified data set with nominal values. $X = (X^1, \ldots, X^m)$ is represented as a table of n rows, each one corresponding to a record describing an individual (e.g., the clinical record of a patient), and m columns, each one containing the nominal values of the attributes of the individuals (e.g., diseases, treatments,

etc.). We use $r_i = (x_i^1, \ldots, x_i^m)$ to refer to the record contributed by the individual *i* and x_i^a to refer to the value of that individual for attribute X^a. The output data set generated by applying the semantic rank swapping method is denoted by $X^* = (X^{1^*}, \ldots, X^{m^*})$.

Because our methods aim at preserving the semantics of the data as much as possible, in the following, we detail how we capture and manage the semantics underlying nominal values by exploiting the formal semantics modeled an ontology. Then, we discuss how to rank data coherently with their semantics, and describe the different *semantic rank swapping* algorithms we propose.

3.1 Semantic management of nominal data

To semantically handle nominal data during the permutation process the values of the nominal attributes must be mapped on an ontology *O*, which formalizes the semantics underlying to the domain of the attributes. An ontology is a formal and structured knowledge base that explicitly and consensually represents the concepts and the semantic interrelations of a domain of knowledge [26]. From a structural perspective, an ontology can be considered as a graph whose nodes represent the concepts of a domain of knowledge and the edges correspond to the relationships between concepts [27]. Semantic relationships between concepts can be classified as taxonomic relationships, e.g., hyponymy and hypernymy (*is-a* links), or non-taxonomic relationships, e.g., meronymy and holonymy (*part-of* links); in this work, we focus on taxonomic relationships because they are available in any ontology and constitute the backbone of the knowledge structure that ontologies provide [28]. For example, in a medical ontology such as SNOMED-CT [29], the concepts can be types of diseases, medical procedures or clinical findings; i.e., single units of thought with a distinct clinical meaning, which are organized into several taxonomies recursively specializing the semantics of the different types of diseases, procedures, etc. In a taxonomy, a concept c_j is a specialization or a subsumed concept of another concept c_i, i.e., $c_j \subseteq c_i$, if and only if, every instance of c_j is also an instance of c_i.

We assume that the *n* values of each attribute in *X* have been univocally associated to concepts *c* modeled in *O*. This process, named *conceptual mapping*, can be carried out manually or by lexically matching the strings of nominal values and concept labels, as done in [30].

Below, we define the taxonomy associated with a nominal attribute $X^a \in X$ whose values have been mapped in an ontology *O*.

Definition 1. Let $S(X^a)$ be the set of subsumers of an attribute X^a mapped in an ontology *O*. The *least common subsumer* of X^a, denoted by $LCS(X^a)$, is the most specific concept in $S(X^a)$.

$$\begin{aligned} S(X^a) &= \left\{ c_i \in O \mid \forall c_j \in X^a : \ c_j \subseteq c_i \right\} \\ LCS(X^a) &= \left\{ c \in S(X^a) \mid \forall c_i \in S(X^a) : \ c \subseteq c_i \right\} \end{aligned} \tag{1}$$

That is, $LCS(X^a)$ is the most specific ancestor in *O* that subsumes all concepts in X^a. For example, if X^a refers to any type of disease, then, $LCS(X^a)$ in SNOMED-CT will be the concept *Disease (disorder)*.

Definition 2. The *taxonomy associated with a nominal attribute* X^a mapped in an ontology *O*, denoted by $\tau(X^a)$, is the concept hierarchy extracted from *O* that includes all concepts that are taxonomic specializations of $LCS(X^a)$, including itself.

$$\tau(X^a) = \left\{ c_i \in O \mid c_i \subseteq LCS(X^a) \right\} \tag{2}$$

Note that $LCS(X^a)$ is also the *root* concept of $\tau(X^a)$. If there exist several paths between a mapped concept x_i^a and $LCS(X^a)$, all of them are included in $\tau(X^a)$.

Once the values of a nominal attribute have been mapped to concepts in an ontology and its associated taxonomy has been obtained, we need a measure to quantify the semantic distance between concept pairs. Specifically, the *semantic distance*, *sd*: $c_1 \times c_2 \rightarrow \Re$, is a function mapping a pair of concepts to a real number that quantifies the differences between the meanings of two concepts according to the semantic evidence gathered from one or several knowledge sources (in our case, the taxonomy associated to the attribute domain). Different ontology-based distance measures have been proposed in the literature. These can be classified according to the theoretical principle they use to assess the semantic distance (or similarity) [31]: *edge-counting measures*, which measure the semantic distance as the number of taxonomic links separating the two concepts, *feature-based measures*, which consider the number of common and non-common ontological features between the concepts to compare, and *information content-based measures*, which assess the semantic similarity between the two concepts according to the amount of information they share (encompassed by their LCS).

In the methods we propose below, any ontology-based semantic distance measure can be used. The selection of the specific measure should be done according to its accuracy, computational cost and the type of knowledge available to assess the distance (e.g. taxonomic and/or non-

taxonomic relationships). To enforce our methods in practice, we used the well-known measure proposed by Wu and Palmer [32] because it is computationally efficient, reasonably accurate, only relies on taxonomic relationships and has been previously used in semantically-grounded data protection algorithms [13, 15, 19]. The semantic distance formulation of Wu and Palmer's measure is defined as follows:

$$sd_{wp}(c_1,c_2) = 1 - \frac{2 \times depth(LCS(c_1,c_2))}{2 \times depth(LCS(c_1,c_2)) + path(c_1,LCS(c_1,c_2)) + path(c_2,LCS(c_1,c_2))}, \quad (3)$$

where c_1 and c_2 are concepts in the taxonomy; $LCS(c_1,c_2)$ 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 $path(c_2,LCS(c_1,c_2))$.

3.2 Order relation on nominal data

Rank swapping requires establishing an order relation on the values of the input attributes to sort them and to perform rank-distance swaps. In this way, it is possible to maintain each permuted value within a certain distance from its original position, thus restricting and controlling the information loss associated to each swap. Therefore, to enforce this method on nominal data, it is in principle necessary to define an order relation that allows ranking all nominal values of the attribute. However, because nominal data are finite, discrete, textual and non-ordinal, a priori, it is not possible to carry out the required sorting operation. In this section, we discuss this issue and propose a suitable solution.

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

categorical attribute may be *color*, where the different categories may be ranked on basis of their wave lengths.

Nonetheless, beyond artificial orders such as the alphabetical order, nominal data may be ranked while considering their semantics by applying statements such as "*x* is a *y*" or "*x* is part of *y*", as formalized in a background ontology. By relying on the subsumption statement "*x* is a *y*", we can determine if a concept specializes another one. For example, if we consider the fragment of the medical ontology SNOMED-CT shown in Figure 1, whose edges represent *is-a* relationships, we can say that *coma* is a *disorder of nervous system*. Formally, the binary relation "is-a"(and, similarly, for "*x* is part of *y*") applied on a nominal attribute X^a mapped on an ontology *O*, denoted by $\subseteq^{X^a}$, is an order relation that holds the following properties for all x_i^a, x_j^a and x_l^a in X^a:

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

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

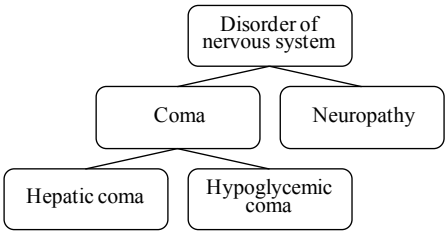


Figure 1. Example of taxonomy associated to the domain *Disease*, extracted from the SNOMED-CT medical ontology.

Because we should be able to rank all the elements of a sample to implement the rank swapping method, we require a binary relation that fulfills the totality property as an alternative to the partial order $\subseteq^{X^a}$. In an attempt to construct a totally ordered set from a partial order that permits to sort the non-comparable data, Torra [25] proposes defining a total topological order consistent with the partial order. Subsequently, a cumulative function of the frequency of the nominal values in the attribute defined on this topological order is used to rank the attribute. However, if this order relation is applied on the partial order $\subseteq^{X^a}$, the ranked attribute would lack semantic coherence because the result is partially determined by the attribute frequency distribution, rather than the semantics of the values; that is, the fact that two nominal attributes are equally frequent in a sample, does not imply that they have equal or even similar meanings.

To obtain a semantically-coherent result, we propose an order relation based on the notion of closeness to a reference point. Given a reference value x^a_{ref} in X^a, a value x^a_i is less than or equal to a value x^a_j, if x^a_i is closer to x^a_{ref} than x^a_j. Because the closeness of a nominal value to another of reference is determined by the difference between the semantics they convey, we can use the notion of *semantic distance* [31] discussed in Section 3.1. For example, if we want to rank the nominal values {*disorder of nervous system*, *coma*, *neuropathy*, *hepatic coma*, *hypoglycemic coma*} mapped into the taxonomy of Figure 1, firstly, we must select a value in the set as the reference point for the order relation. Secondly, we must calculate the semantic distance between each value of the set and that reference point. With this, the values of the set can be coherently ranked according to the computed distances. In the example, if the reference point was the concept *coma*, the ranked set would follow the sequence shown in Figure 2. As we can see, if we change the reference point, we obtain a different rank sequence, as shown in Figure 3. Therefore, this order relation generates as many rank sequences as different values the data set has.

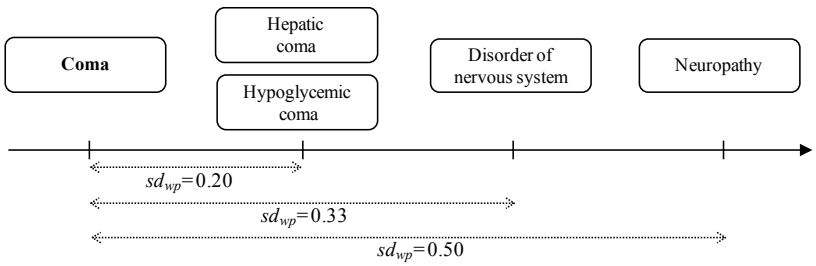


Figure 2. Example of ascending rank when the reference point is *Coma*. Semantic distances are calculated with the Wu and Palmer measure detailed in Section 3.1.

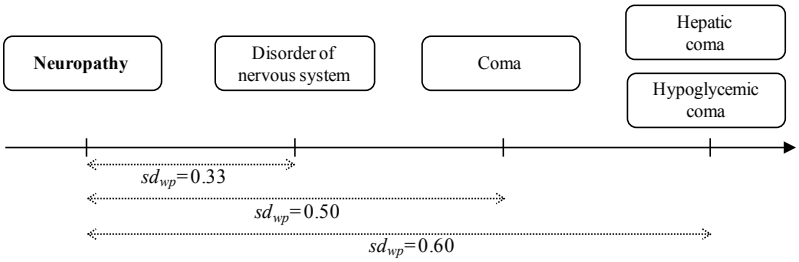


Figure 3. Example of ascending rank when the reference point is *Neuropathy*. Semantic distances are calculated with the Wu and Palmer measure detailed in Section 3.1.

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

Definition 3. The *order relation* on a nominal attribute X^a, given a reference value x^a_{ref} in X^a, denoted by $\leq_{x^a_{ref}}$, is defined as a binary relation where a value x^a_i is less than or equal to a value x^a_j, i.e., $x^a_i \leq_{x^a_{ref}} x^a_j$, if and only if the *semantic distance*, $sd(\cdot,\cdot)$, between x^a_i and x^a_{ref} is less than or equal to the *semantic distance* between x^a_j and x^a_{ref}.

$$\leq_{x^a_{ref}} = \left\{ x^a_i, x^a_j \in X^a : \quad x^a_i \leq_{x^a_{ref}} x^a_j \mid sd(x^a_i, x^a_{ref}) \leq sd(x^a_j, x^a_{ref}) \right\} \tag{4}$$

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

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

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

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

By applying this to our problem, the sequence of values of X^a ranked in ascending order according to $\leq_{x^a_{ref}}$ is represented by $\overrightarrow{X^a}^{\leq} = \langle x^a_{(1)}, \ldots, x^a_{(n)} \rangle$. Note that x^a_i and $x^a_{(i)}$ represent the i^{th}-

unordered and i^{th}-ordered value of X^a, respectively, i.e. $rank(x^a_{(i)})=i$. Obviously, the first value in the ranking is x^a_{ref}, i.e. $x^a_{(1)} = x^a_{ref}$. If there are tied values w.r.t. x^a_{ref}, these will be placed in contiguous positions in the ranking.

3.3 Semantic univariate rank swapping method

By means of the total preorder relation discussed in Section 3.2, which enables us to semantically sort nominal values, we have the mean to adapt rank swapping to the semantic domain of nominal data. In this section, we propose a semantically-grounded univariate rank swapping method for individual nominal attributes that pursues a twofold objective:

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

To generate the permuted version X^* of the original data set X, the semantic univariate rank swapping method must be applied independently on each nominal attribute. Let $X^a = (x^a_1,\ldots,x^a_i,\ldots,x^a_n)$ be an attribute in X. Like the numerical method, the records of X must be ranked in ascending order by values x^a_i of the attribute X^a. To do this, we use the total preorder relation $\leq_{x^a_{ref}}$ considering as reference point the boundary of the attribute, i.e. the most semantically-distant value from X^a. To find the most distant value from a set of nominal data in a semantically-coherent way, we use the notion of *marginality* [33]. Specifically, the marginality of a nominal value of a sample shows how outlying is that element w.r.t. the remaining values of the sample according to the aggregation of semantics distances. On this basis, we propose the following definition of the *most semantically-distant value of a nominal attribute*:

Definition 4. The *most semantically-distant value of a nominal attribute* X^a, denoted by $MostDistantValue(X^a)$, is the value x^a from X^a that maximizes the sum of the semantic distances w.r.t. all x^a_i in X^a.

$$MostDistantValue\left(X^a\right) = \arg\max_{x^a \in X^a}\left(\sum_{x^a_i \in X^a} sd\left(x^a, x^a_i\right)\right) \quad (5)$$

On the other hand, in order to offer a configurable level of permutation, and thus, to satisfy the first objective of the method, we should allow the user to define the length of the swapping interval or range. Similar to the numerical rank swapping method, this is accomplished through an input parameter *k*, which represents the length of the swapping interval in number of records. Notice that related works use *p* as input parameter, which specifies the percentage of the records in X in the interval, i.e., $k=p.n/100$. In our case, by setting *k*, the rank of two swapped values cannot differ by more than *k* records, which provides a clearer privacy guarantee than the use of percentages (see Section 3.5 for further details).

The semantic univariate rank swapping method is formalized in Algorithm 1. Firstly, the taxonomy $\tau(X^a)$ associated to the attribute X^a is obtained from the ontology O by following the procedure detailed in Section 3.1. Secondly, in line 2, the reference point x^a_{ref} used to rank X^a is computed by using eq. (5); this value is the most semantically-distant value of X^a. According to x^a_{ref}, the values in X^a are ranked in ascending order in line 3 by using the total preorder relation defined in eq. (4). Then, in line 4, the values of the ranked attribute $\overrightarrow{X^a} = \langle x^a_{(1)},\ldots, x^a_{(n)}\rangle$ are labeled as *unswapped*. In lines 5-12, each unswapped value $x^a_{(i)}$ is permuted by another unswapped value randomly chosen within a restricted range through the procedure *swap_value*. This swapping range is composed of the *k* values following to $x^a_{(i)}$ in the ranking $\langle x^a_{(1)},\ldots, x^a_{(n)}\rangle$, i.e., the interval $\left[x^a_{(i+1)}, x^a_{(i+k)}\right]$. The size of the interval is kept in *k* values, except when the index $i+k$ is greater than n, i.e., $i+k$ is greater than the size of the attribute. For this reason, the upper limit of the interval is the lower value of $\{i+k, n\}$, i.e., $\left[x^a_{(i+1)}, x^a_{(\min\{i+k,n\})}\right]$. After each swap, in lines 22 and 24, the processed values are labeled as *swapped*.

Algorithm 1. *Semantic univariate rank swapping* method.

Input :

X^a : nominal attribute with n records

O: ontology

k: length of the swapping interval in number of records

Output :

X^{a*} : rank-swapped nominal attribute

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

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

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

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

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

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

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

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

9: $interval \leftarrow$ obtain_SwappingInterval($\overrightarrow{X^a}, lower\_limit, upper\_limit$) //$interval = \left[x^a_{(i+1)}, x^a_{(\min\{i+k,n\})} \right]$

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

11: **end if**

12: **end for**

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

14: **procedure** swap_value ($swappingInterval, x^{attr}_{ref}$)

15: $\overline{swappingInterval} \leftarrow$ obtain_UnswappedValues($swappingInterval$) //$\overline{swappingInterval}$ is the set of the unswapped values in $swappingInterval$

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

17: $x^{attr}_{swa} \leftarrow$ select_SwappingValue($\overline{swappingInterval}$) //random selection

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

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

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

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

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

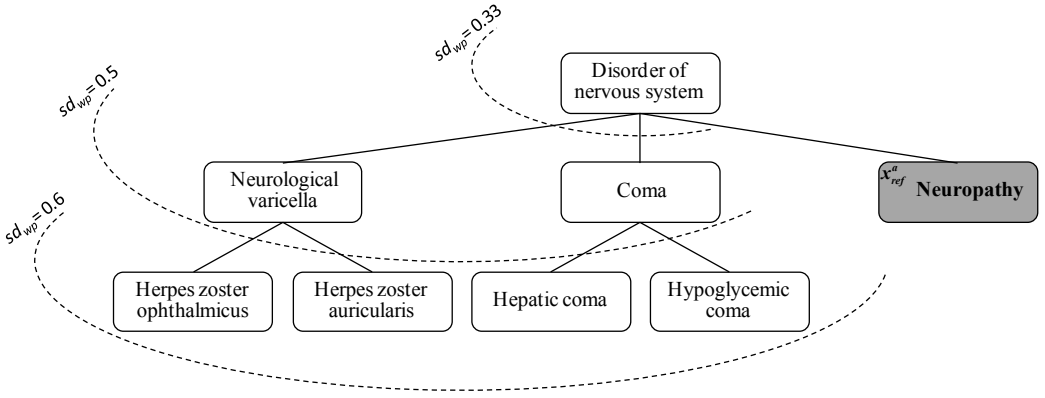
23: **end if**

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

25: **end procedure**

In Algorithm 1, the attribute is only ranked once at the beginning of the process. As discussed in Section 3.2, the total preorder relation $\leq_{x^a_{ref}}$ yields as many order sequences as different values the attribute has. This means that the ranking obtained at the beginning of the process is suitable

(w.r.t. bounding the maximum semantic permutation resulting from each swap) to build the swapping interval of the first treated value, but not to build the intervals of the remaining values. Because the ranking of the attribute is not quite suitable for those remaining values, very different values may be swapped from the second swap and upwards, thus incurring in an information loss much higher than that expected from the permutation level *k*. This issue is illustrated in Figure 4 for an attribute X^a={*Disorder of nervous system*, *Neurological varicella*, *Coma*, *Neuropathy*, *Herpes zoster ophthalmicus*, *Herpes zoster auricularis*, *Hepatic coma, Hypoglycemic coma*} and *k* =2. After ranking the values of X^a w.r.t. the most semantically-distant value of X^a, *Neuropathy*, we can see that the established ranking is suitable to build the swapping interval of $x^a_{(1)} = Neuropathy$ because the resulting interval is composed of the two most semantically-similar values to *Neuropathy* (i.e, *Disorder of nervous system* and *Coma*). However, for $x^a_{(6)} = Hepatic\ coma$, the swapping interval includes values that are much more semantically distant than expected from the value of *k* (e.g., *Herpes zoster ophthalmicus*, whose semantic distance w.r.t. *Hepatic coma* is sd_{wp}= 0.67).



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

Figure 4. Example of swapping intervals in an ascending ranked attribute w.r.t. *MostDistantValue*(X^a) = *Neuropathy*. Semantic distances are calculated with the Wu and Palmer measure detailed in Section 3.1.

To solve this issue, we propose a variation of Algorithm 1 that better preserves data semantics by swapping the original values with others within a semantic distance coherent with the input parameter *k*, thereby satisfying the objective 2(ii). The difference of this approach w.r.t. the previous one lies in the way of generating the swapping intervals during the permutation process. In this new approach, we propose re-ranking the attribute for each value x^a_i to swap by using x^a_i as reference point, i.e., $x^a_{ref} = x^a_i$. In this way, once the attribute has been ranked in

ascending order w.r.t. x_i^a, the swapping interval will be formed by the set of the *k* semantically-closest values to x_i^a. This swapping interval is formally defined bellow.

Definition 5. The *swapping interval* associated to a reference point x_{ref}^a from a nominal attribute X^a, denoted by $I_{x_{ref}^a}$, is defined as the set of the *k* semantically-closest values to x_{ref}^a in X^a, i.e. the *k* first values in the ranked attribute $\overrightarrow{X^a} = \langle x_{(1)}^a, \ldots, x_{(n)}^a \rangle$, excluding $x_{(1)}^a$.

$$I_{x_{ref}^a} = \left[x_{(2)}^a, x_{(1+k)}^a \right], \quad 1 \leq k < n \tag{6}$$

Because a new interval $I_{x_{ref}^a}$ needs to be generated every time a new value x_i^a must be swapped, we call this process *dynamic building of swapping intervals*.

To prioritize the permutation of the most marginal values, which are those whose swaps entail more information loss, we propose building the swapping intervals *at opposite ends*: each time a new swapping interval must be generated, it must be as far as possible from the previously generated interval. In other words, the reference points (i.e., the values to be swapped) of two consecutive swapping intervals must *maximize* their semantic distance.

To do this, our method starts by building the swapping interval at the boundary of X^a; that is, the first element to swap and, thus, the reference point $x_{ref}^{a'}$ for the first interval, will be the *most distant value* of X^a (eq. (5)). Then, to select the second element to swap, and thus, the second reference point $x_{ref}^{a''}$, we look for the value in X^a that is semantically-farthest from $x_{ref}^{a'}$, as formalized in Definition 6. Note that, $x_{ref}^{a''}$ must be selected among the still unswapped values in X^a.

Definition 6. The most semantically-distant value to a reference value x_{ref}^a in a nominal attribute X^a, denoted as *MostDistantValue*(X^a, x_{ref}^a), is the value x^a from X^a that maximizes the semantic distance with x_{ref}^a.

$$MostDistantValue\left(X^a, x_{ref}^a\right) = \underset{x^a \in X^a}{\arg\max} \left(sd\left(x^a, x_{ref}^a\right)\right) \tag{7}$$

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

Once a reference point x^a_{ref} has been selected (by using eq. (7)), the corresponding swapping interval is built according to eq. (6). Finally, x^a_{ref} is swapped with a value from the interval randomly chosen among those that still have not been swapped. In this way, like Algorithm 1, the rank of two swapped values cannot differ by more than *k* records.

Algorithm 2 formalizes the above-described procedure. As stated above, the first element to swap and, thus, the first reference point, is the most semantically-distant value of X^a, selected by using eq. (5). Then, in line 5, the swapping interval is built around the reference point by using eq. (6). In line 6, through the procedure *swap_value* of Algorithm 1, the value x^a_{ref} is swapped with another unswapped value randomly chosen within the interval. Finally, in line 7, the next reference point and, thus, the next value to swap, is chosen by applying the strategy of dynamically building the swapping intervals at opposite ends (eq. (7)). This process is repeated until all values in X^a are swapped.

Algorithm 2. *Semantic univariate rank swapping* method based on the strategy of dynamically building the swapping intervals at opposite ends.

Input :
X^a : nominal attribute with *n* records
O: ontology
k: length of the swapping interval in number of records
Output :
X^{a*} : rank-swapped nominal attribute
1: $\tau(X^a) \leftarrow$ obtain_taxonomy(X^a, O)
2: $\overline{X^a} \leftarrow X^a$ // $\overline{X^a}$ is the set of unswapped values in X^a
3: $x^a_{ref} \leftarrow$ obtain_MostDistantValue($\overline{X^a}$) // $x^a_{ref} = \arg\max_{x^a \in \overline{X^a}} \left(\sum_{x^a_i \in \overline{X^a}} sd(x^a, x^a_i) \right)$
4: **while** X^a has unswapped values **do**
5: $I_{x^a_{ref}} \leftarrow$ obtain_SwappingInterval(X^a, x^a_{ref}, k) // $I_{x^a_{ref}}$ is the set of the k semantically-closest values to x^a_{ref} in X^a
6: swap_value($I_{x^a_{ref}}, x^a_{ref}$)
7: $x^a_{ref} \leftarrow$ obtain_MostDistantValue($\overline{X^a}, x^a_{ref}$) // $x^a_{ref} = \arg\max_{x^a \in \overline{X^a}} \left(sd(x^a, x^a_{ref}) \right)$
8: **end while**
9: **return** X^{a*}

Regarding the objective 2(i), because the values in the permuted attribute are the same as those in the original attribute but swapped, by definition, the univariate features of the attributes (mean, variance, etc.) are perfectly preserved.

Regarding the computational complexity of Algorithm 2, due to the need to build total preorders for each value to be swapped, Algorithms 2 executes $n/2$ quasi-linear sorting operations for each attribute. Thus, its complexity is $O(n^2 \log n)$ for each attribute.

3.4 Semantic multivariate rank swapping method

The rank swapping methods presented in the previous section preserve, by construction, the univariate features of the attributes and incur in an information loss proportional to the desired level of permutation. However, because they are independently applied to each attribute of the data set, the potential correlation among attributes is likely to be significantly hampered. To solve this issue, we propose a semantic rank swapping method that, in addition to fulfilling the objectives of the univariate versions, is also capable of reasonably preserving the correlation (i.e., the semantic dependence in the nominal domain) among non-independent attributes (e.g., a *disease type* and the *medical procedure* used to treat it).

Let X be a data set with m nominal attributes and n records, such that $X=(X^1,\ldots,X^m)=\left\{r_i=(x_i^1,\ldots,x_i^m):\quad i=1,\ldots,n\right\}$, where the record $r_i=(x_i^1,\ldots,x_i^m)$ represents the value of the m attributes for individual i.

In order to preserve, as much as possible, the correlation among the m non-independent attributes of X during the permutation process, we propose considering each record as a unit, which thus conveys the relationship between attribute values. In this way, a *swapping interval* will be composed of the k semantically-closest records to a reference record. Like Algorithm 2, the swapping intervals will be dynamically built at opposite ends to minimize the information loss, but now they will encompass records rather than individual attributes. After obtaining the swapping interval of a reference record, each value in the reference record is swapped with another value of the same attribute, randomly chosen among those in the interval that still have not been swapped. Because the swap is independently carried out for each attribute value of the reference record, the resulting records will be different from those in the original data set, thus preventing re-identification. Nonetheless, because the swapping range is delimited by semantically similar records, attribute values within the swapping range will be both semantically similar within each attribute (which minimizes information loss) and semantically interrelated with the values of the other attributes (which contributes to preserve the attribute correlation).

Below, Definitions 3-6 are adapted to work with nominal records rather than individual attributes.

Definition 7. The order relation on a data set X of m nominal attributes, given a reference record r_{ref} in X, denoted by $\leq_{r_{ref}}$, is defined as a binary relation where a record r_i is less than or equal to a record r_j, i.e., $r_i \leq_{r_{ref}} r_j$, if and only if the semantic distance between r_i and r_{ref} is less than or equal to the semantic distance between r_j and r_{ref}, for all r_i and r_j belonging to X.

$$\leq_{r_{ref}} = \left\{ r_{ref}, r_i, r_j \in X : \quad r_i \leq_{r_{ref}} r_j \mid sd(r_i, r_{ref}) \leq sd(r_j, r_{ref}) \right\}, \tag{8}$$

where the *semantic distance* between a pair of records r_i and r_j is computed as the aggregation of pairwise semantic distances between attribute values:

$$sd\left(r_i, r_j\right) = \frac{1}{m} \sum_{attr=1}^{m} sd\left(x_i^{attr}, x_j^{attr}\right) \tag{9}$$

Definition 8. The most semantically-distant record in a data set X of m nominal attributes, denoted by *MostDistantRecord*(X), is the record r from X that maximizes the sum of the semantic distances respect to all r_i in X.

$$MostDistantRecord\left(X\right) = \arg\max_{r \in X} \left(\sum_{r_i \in X} sd(r, r_i) \right) \tag{10}$$

Definition 9. The swapping interval associated to a reference record r_{ref} in a data set X of m nominal attributes, denoted by $I_{r_{ref}}$, is defined as the set of the k semantically-closest records to r_{ref} in X, i.e. the k first records in the ranked set $\left\langle r_{(1)}, \ldots, r_{(n)} \right\rangle$, excluding $r_{(1)}$.

$$I_{r_{ref}} = \left[r_{(2)}, r_{(k+1)} \right], \quad 1 \leq k < n \tag{11}$$

Definition 10. The most semantically-distant record from a reference record r_{ref} in a data set X of m nominal attributes, denoted by *MostDistantRecord*(X,r_{ref}), is the record r from X, that maximizes the semantic distance with r_{ref}.

$$MostDistantRecord\left(X, r_{ref}\right) = \arg\max_{r \in X} \left(sd(r, r_{ref}) \right) \tag{12}$$

The method for m attributes is formalized in Algorithm 3. First, in line 4, all records in the input data set X are labeled as unswapped in $\overline{X}$. In line 5, the most distant record from X is obtained by using the eq. (10). Similar to the Algorithm 2, this record is the first to swap and, thus, the

first reference point $r_{ref} = (x_{ref}^1, \ldots, x_{ref}^m)$. Then, in line 7, the interval is built around the reference record r_{ref} by using the eq. (11). In lines 8-10, through the procedure *swap_value* of Algorithm 1, the swaps are independently undertaken for each attribute. Each value x_{ref}^{attr} in r_{ref} is swapped by another unswapped value belonging to the same attribute randomly chosen within the interval. Finally, in line 11, the next reference point is chosen by applying the idea of dynamically building the swapping intervals at opposite ends (eq. (12)). This process is repeated until all records in X are swapped. A record is considered swapped when all its values have been swapped.

Algorithm 3. *Semantic multivariate rank swapping* method based on dynamically building the swapping intervals at opposite ends.

Input :

X : nominal data set with m attributes and n records // $X=(X^1,\ldots,X^m)=\left\{r_i=(x_i^1,\ldots,x_i^m):\ i\ =\ 1,\ldots,\ n\right\}$

O: ontology

k: length of the swapping interval in number of records

Output :

X^* : rank-swapped nominal data set // $X^*=(X^{1*},\ldots,X^{m*})$

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

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

3 : **end for**

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

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

5 : $r_{ref} \leftarrow$ obtain_MostDistantRecord($\overline{X}$) //$r_{ref} = (x_{ref}^1,\ldots,x_{ref}^m) = \arg\max_{r\in\overline{X}}\left(\sum_{r_i\in\overline{X}} sd(r,r_i)\right)$

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

7 : $I_{r_{ref}} \leftarrow$ obtain_SwappingInterval(X, r_{ref}, k) //$I_{r_{ref}}$ is the set of the k semantically-closest records to r_{ref} in X

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

9 : swap_value($I_{r_{ref}}[attr], x_{ref}^{attr}$)

10 : **end for**

11 : $r_{ref} \leftarrow$ obtain_MostDistantRecord($\overline{X}, r_{ref}$) //$r_{ref} = \arg\max_{r\in\overline{X}}\left(sd(r,r_{ref})\right)$

12 : **end while**

13 : **return** X^*

Because Algorithm 3 manages data on record basis, it just performs $n/2$ quasi-linear sorting operations. Thus, its complexity is $O(n^2 \log n)$, which is in line with other multivariate data protection algorithms, such as data microaggregation [34].

3.5 On the privacy guarantees of semantic rank swapping

When applied to *quasi-identifiers*, rank swapping methods protect data against re-identification (i.e., *identity disclosure*). Either by applying the univariate methods to each quasi-identifying attribute or the multivariate method to all the quasi-identifiers at once, the tuples of quasi-identifying values can no longer be unequivocally associated to a known individual. As a result, the data set is anonymized because an attacker cannot know for sure if the individual he tries to re-identify appears in the data set.

In our methods, the k parameter defines the size of the swapping interval, which affects both the degree of protection and the preservation of data utility. When k increases, the dissimilarity or semantic distance between the original value (*before-swap value*) and the permuted value (*after-swap value*) tends to increase; this makes re-identification harder, but deteriorates the permuted data quality because the information loss associated to each swap tends to increase. Specifically, by setting k, the rank of two swapped values cannot differ by more than k records. For the multivariate method, this guarantees that the probability that an attacker infers the original values will not be greater than $1/k$. This constitutes an *ex ante* privacy guarantee that is known as *probabilistic k-anonymity* [35].

Probabilistic k-anonymity is a privacy model that provides the same level of protection against re-identification as *k-anonymity* [36, 37]: the probability of re-identifying individuals in the protected data set is, at most, $1/k$. *k-Anonymity* achieves this by demanding that each combination of values of quasi-identifying attributes is shared by, at least, k records; that is, records are indistinguishable w.r.t. their quasi-identifiers in groups of k. The indistinguishability requirements is enforced in practice by generalization and suppression [36] or microaggregation [38], which necessarily entails data variability and granularity loss and, therefore, substantial information loss. On the other hand, *probabilistic k-anonymity* does not require that records are indistinguishable. By relaxing the indistinguishability requirement, *probabilistic k-anonymity* can be satisfied by a wider set of mechanisms and with less information loss than *k-anonymity* [35]. Specifically, our multivariate rank swapping method allows attaining the required limit in the probability of record re-identification while preserving all univariate features of the data set and, as we discuss in the next section, by reasonably preserving the correlations among attributes.

4. Empirical analysis

In this section, we evaluate the semantic rank swapping methods we propose in Section 3 with several nominal data sets and w.r.t. different evaluation metrics. We also compare their results with those obtained by several baselines.

4.1 Evaluation data and utility metrics

As evaluation data, we used a structured database containing patient discharge data provided by the California Office of Statewide Health Planning and Development (OSHPD)[2], which were collected from licensed hospitals in California in 2009. Each record of the database details the healthcare discharge of a patient and, among others, it contains several nominal attributes stating the *diagnoses* and the *medical procedures* applied to the patient, which we selected to evaluate our methods. These data are especially suitable to illustrate the need for our methods due to their highly sensitive nature and their interest in research. In this respect, secondary uses of the health care data must guarantee the confidentiality of the patients to which the data refer, as stated in several regulations and guidelines (the Health Insurance Portability and Accountability Act (HIPAA) [39] or the Medical Information Security and Privacy Protection Guidelines [40]).

Diagnoses and procedure codes have been mapped to healthcare concepts in the SNOMED-CT medical ontology [29], which is especially well-suited to assist semantic distance assessments of medical-related data because of its large size and fine grained taxonomic detail: it contains more than 311,000 unique concepts organized in 18 overlapping hierarchies with more than 1.36 million relationships.

To assess the utility of the outcomes resulting from the permutation process from a semantic perspective, and, therefore, to evaluate the suitability of our methods, we need to measure up to which level the semantic features of the data have been preserved. To measure these features in a semantically and mathematically coherent way, we use the semantic versions of the *mean*, *variance*, *covariance* and *correlation* measures we defined in [19, 41], which constitute semantic counterparts of the statistical measures used to quantify the (numerical) utility in data anonymization [7]. In the same way as the numerical measures rely on the arithmetical difference between values, the semantic counterparts use the notion of *semantic distance* introduced in Section 3.2 to quantify the differences between the *meaning* of nominal values, as follows.

[2] http://www.oshpd.ca.gov/HID/Data_Request_Center/PUF.html

The *semantic mean* of a nominal attribute X^a is defined (discretized) as the concept c in $\tau(X^a)$ that minimizes the sum of the semantic distances w.r.t. all x_i^a in X^a, that is, the most central value to the sample.

$$s\mu(X^a) = \arg\min_{c \in \tau(X^a)} \left(\sum_{x_i^a \in X^a} sd\left(c, x_i^a\right) \right) \tag{13}$$

To assess the semantic dependence between nominal attribute pairs (e.g., diseases and their treatments), we use the *semantic distance covariance* and the *semantic distance correlation*, which are semantic adaptations of the statistical dependence measures proposed in [42].

The *semantic distance covariance* between two nominal attributes X^a and X^b is defined as the square root of the arithmetic mean of the products $\delta_{ij}^{X^a} \delta_{ij}^{X^b}$:

$$sdCov\left(X^a, X^b\right) = \frac{1}{n} \sqrt{\sum_{i,j=1}^{n} \delta_{ij}^{X^a} \delta_{ij}^{X^b}}, \tag{14}$$

where $\delta_{ij}^{X^a}$, and similarly for $\delta_{ij}^{X^b}$, is the value of the $(i,j)^{\text{th}}$ element of the $(n \times n)$ double centered semantic distance matrix of X^a, denoted by $\left(\delta_{ij}^{X^a}\right)_{i,j=1}^{n}$ and computed as follows:

$$\left(\delta_{ij}^{X^a}\right)_{i,j=1}^{n} = \left(sd(x_i^a, x_j^a) - \overline{sd}_{i.}^{X^a} - \overline{sd}_{.j}^{X^a} + \overline{sd}_{..}^{X^a} \right)_{i,j=1}^{n} \tag{15}$$

Each element $\delta_{ij}^{X^a}$ of $\left(\delta_{ij}^{X^a}\right)_{i,j=1}^{n}$ is calculated through the $(n \times n)$ semantic distance matrix $\left(sd(x_i^a, x_j^a)\right)_{i,j=1}^{n}$ of the attribute X^a, such that $sd(x_i^a, x_j^a)$ is the value of the $(i,j)^{\text{th}}$ element, $\overline{sd}_{i.}^{X^a}$ is the mean of i^{th} row, $\overline{sd}_{.j}^{X^a}$ is the mean of j^{th} column and $\overline{sd}_{..}^{X^a}$ is the mean of all values that compose the matrix. The semantic distance covariance satisfies $sdCov\left(X^a, X^b\right) \geq 0$ and it is 0 if and only if X^a *and* X^b are independent.

Like the standard numerical correlation, the *semantic distance correlation* of two nominal attributes X^a and X^b is the nonnegative number computed by dividing their semantic distance covariance, $sdCov\left(X^a, X^b\right)$, by the product of their *semantic distance standard deviations*.

$$sdCor(X^a, X^b) = \begin{cases} \dfrac{sdCov(X^a, X^b)}{\sqrt{sdVar(X^a) \times sdVar(X^b)}}, & sdVar(X^a) \times sdVar(X^b) > 0 \\ 0, & sdVar(X^a) \times sdVar(X^b) = 0 \end{cases}, \quad (16)$$

where $sdVar(X^a)$ and $sdVar(X^b)$ are the *semantic distance variances* of attributes X^a and X^b, which are a particular case of semantic distance covariance where the two attributes are identical (i.e., $sdVar(X^a) = sdCov(X^a, X^a)$).

The *semantic distance correlation* is bounded in the [0..1] range, and it is 0 if and only if X^a and X^b are independent. As in the numerical case, values close to zero evince a weak semantic association between attributes, while larger values evince a stronger association.

By relying on these semantically grounded measures, we quantify the loss of utility resulting from the permutation process and, thus, the suitability of our algorithms, according to the following metrics:

1. The *semantic mean* of the original attribute X^a, $s\mu(X^a)$, and of the permuted attribute X^{a^*}, $s\mu(X^{a^*})$, by using eq. (13), and the *semantic distance* between both, $sd(s\mu(X^{a^*}), s\mu(X^a))$. A distance value of 0 indicates that the mean has been perfectly preserved after the swapping process.
2. The *semantic distance variance* of original and permuted attributes, $sdVar(X^a)$ and $sdVar(X^{a^*})$, and the absolute difference between both values. A difference of 0 indicates that the variance has been perfectly preserved after the swapping process.
3. The *root mean square error* (RMSE), measured as the root average square *semantic distance* between original and permuted value pairs, $RMSE(X^a, X^{a^*})$. It measures the overall loss of information in terms of semantics. Small values indicate low information loss, and thus, permuted data with better quality.
4. The *semantic distance correlation* of original and permuted attribute pairs, $sdCor(X^a, X^b)$, and $sdCor(X^{a^*}, X^{b^*})$ by using eq. (16), and the absolute difference between the actual *semantic distance correlation* of pairs of permuted attributes and original attributes, i.e., $|sdCor(X^{a^*}, X^{b^*}) - sdCor(X^a, X^b)|$. A difference of 0 indicates that the correlation has been perfectly preserved after the swapping process.

4.2 Comparison of algorithms

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

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

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

We have evaluated the two versions of the univariate method (Algorithm 1 and Algorithm 2, Section 3.3) and the multivariate method (Algorithm 3, Section 3.4).

Tables 2, 3 and 4 depict the semantic features and the evaluation metrics of the results provided by Algorithms 1, 2 and 3, respectively, for several values of the input parameter k={2, 5,10, 20, 50, 100}.

Table 2. *Dataset1*: evaluation metrics of rank-swapped attributes values (X^a = *principal diagnosis,* X^b = *medical procedure*) with the Algorithm 1.

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

Table 3. *Dataset1*: evaluation metrics of rank-swapped attributes values (X^a = *principal diagnosis,* X^b = *medical procedure*) with the Algorithm 2.

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

Table 4. *Dataset1*: evaluation metrics of rank-swapped attributes values (X^a = *principal diagnosis,* X^b = *medical procedure*) with the Algorithm 3.

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

First, we can see the *semantic mean* and the *semantic variance* are perfectly preserved for all the attributes. Because attribute values in the permuted outcome are the same as those in the original data set but swapped, by definition, univariate features (e.g., mean, variance, min/max values) are perfectly preserved for each individual attribute. This contrasts with other data protection mechanisms, such as data microaggregation [34] or generalization [36], which protect data by making them more homogeneous and, therefore, reduce the variance and/or the granularity of the original data; or noise addition, which alters the dispersion of the sample [43].

Second, because the *k* parameter determines the swapping range, the larger the *k*, the larger the permutation and, thus, the larger the RMSE. Specifically, RMSEs in Tables 2-4 show that our methods are able to proportionally distort the outcomes according to the desired level of protection. In addition, Algorithms 2 and 3 provide significantly lower RMSE values than Algorithm 1. The differences observed between the former and the latter quantify the positive influence of the *dynamic building the swapping intervals at opposite ends* strategy we detailed in Section 3.3, which minimizes the information loss resulting from the permutation process by

i) limiting the swapping range of the original value/record to the *k* semantically-closest values/records in the data set, and ii) prioritizing the permutation of those values whose swaps entail more information loss. Algorithm 2 provides slightly better RMSEs for individual attributes than Algorithm 3 because the former is able to optimize the swapping ranges for individual attributes, whereas the latter does it for complete records, which is likely suboptimal for individual attributes.

Finally, as expected, correlations between attributes are preserved by the multivariate method significantly better than by any of the univariate algorithms, because the former constraints value swappings towards maintaining the dependence between the attributes. Regarding univariate methods, we can see that Algorithm 2 reasonably preserves correlations because, for two correlated attributes, it is reasonable to assume that, if a value of the first attribute is closely related to another value of the second attribute, values semantically similar to the former one (resulting from the swap) will also be related to values semantically similar to the latter. This behavior also explains why Algorithm 1 provides such poor correlation results, especially for low values of *k*: correlation differences are proportional to the also large RMSEs. So, we can conclude that Algorithm 2 is still valid when maintaining attribute correlations is not a priority or when dealing with non-dependent attributes, because it is able to minimize per-attribute errors better than Algorithm 3.

To contextualize the results of our methods against those of related works, in Figures 5-7, we compare the outcomes of our algorithms with those provided by two non-semantic swapping mechanisms. The first one is a basic *data swapping* method that does not employ any criterion to compare nominal values. Due to its impossibility of using total preorders on nominal attributes, it randomly permutes the values of the input attribute without restricting the swapping range; that is, the values to swap are chosen randomly and these may be swapped by any other value of the attribute. The second one uses a distributional criterion to map nominal values to numbers corresponding to their frequencies of appearance in the data set. This mapping is based on the notion of *distributional semantics* that, in absence of a better semantic background, quantifies the similarities between linguistic items based on their distributional properties in a sample [44]. In this way, this *distributional rank swapping* mechanism is able to rank nominal values according to their frequency, and swap them within restricted intervals of the *k* most similar values w.r.t. their distributions; as a result, e.g., common (rarer) diagnoses will tend to be replaced by also common (rarer) diagnoses.

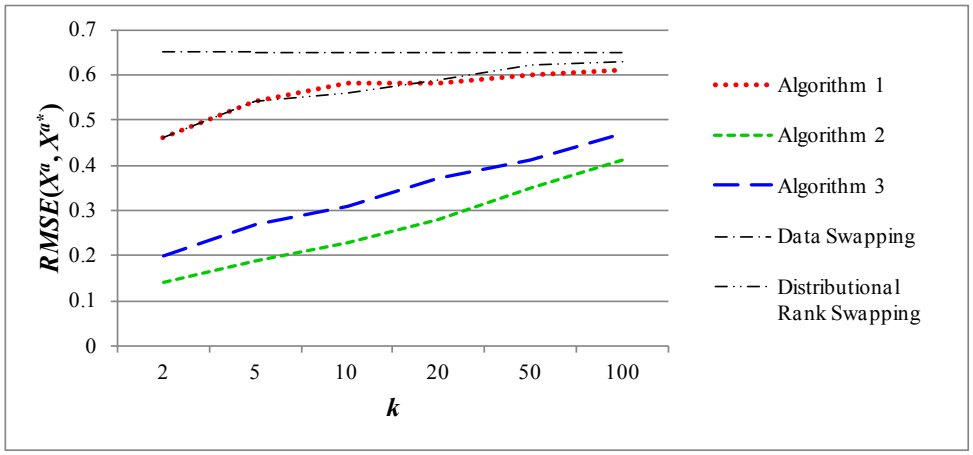


Figure 5. RMSE of attribute X^a with *data swapping*, *distributional rank swapping* and *semantic rank swapping* (Algorithms 1, 2 and 3) with *Dataset1*.

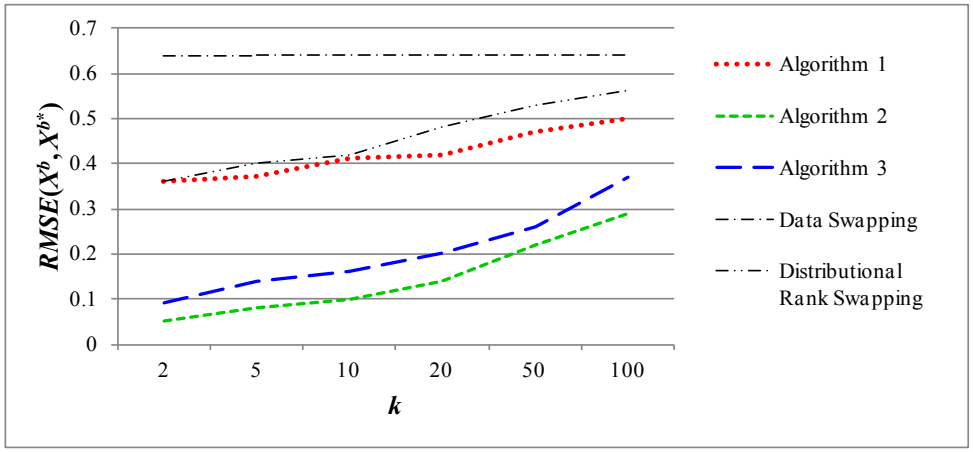


Figure 6. RMSE of attribute X^b with *data swapping*, *distributional rank swapping* and *semantic rank swapping* (Algorithms 1, 2 and 3) with *Dataset1*.

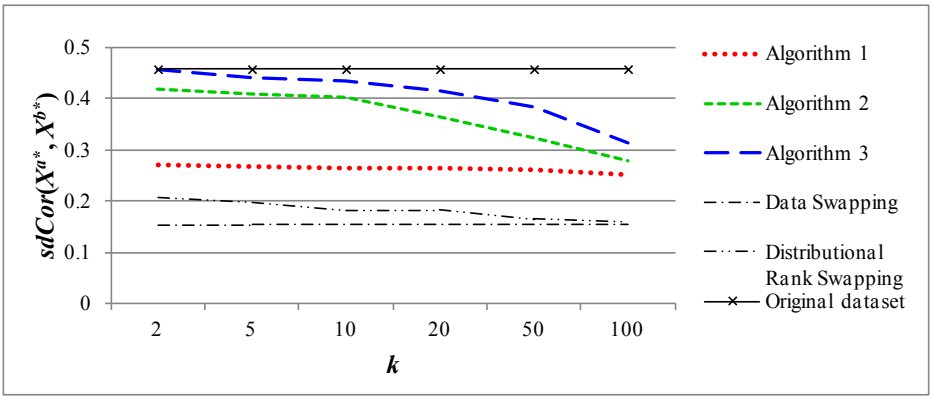


Figure 7. Semantic distance correlation for *data swapping*, *distributional rank swapping* and *semantic rank swapping* (Algorithms 1, 2 and 3) with *Dataset1*.

As expected, the random swaps produced by the *data swapping* mechanism result in a large perturbation that is also non-configurable and, moreover, largely breaks the correlation between

the attributes. On the other hand, *distributional rank swapping* behaves significantly better and it even competes with our semantic *Algorithm 1* w.r.t. the RMSE. We can therefore conclude that the "fixed" semantic ranking implemented by *Algorithm 1* offers little benefit over the distributional criterion used by the *distributional rank swapping*. Even though, the attribute correlation shows a different picture: it is significantly worse for the two non-semantic mechanisms. In any case, our more semantically accurate methods, *Algorithms 2* and *3*, drastically improve the RMSEs and attribute correlations, both w.r.t. the non-semantic mechanisms and w.r.t. *Algorithm 1*. This shows the benefits of a proper and accurate management of the semantics conveyed by nominal data.

4.3 On the preservation of attribute correlations

To test the generality of our multivariate method (Algorithm 3), a second experiment has been carried out with another data set of 1,012 patients, named *Dataset2*. In this case, the attributes X^a =*principal diagnosis* and X^b = *medical procedure* present a stronger correlation than in *Dataset1*, $sdCor(X^a, X^b)$ =0.6392. This stronger correlation implies that records are more homogenous and the frequencies of attribute categories are higher: X^a = 55 different categories (average of 18.4 records per category) and X^b = 94 different categories (average of 10.8 records per category). Table 5 depicts the RMSEs and semantic correlation metrics of the results provided by Algorithm 3.

Table 5. *Dataset2*: evaluation metrics of rank-swapped attributes values (X^a = *principal diagnosis,* X^b = *medical procedure*) with the Algorithm 3.

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

As we can see, for k=2, the RMSE of both attributes is zero, which means that the swaps of attribute values have not resulted in values different from the original ones. This behavior is consistent with the frequency distribution of both attributes in *Dataset2:* because the cardinality of attribute values, and also of records of the two attributes, is larger than 2, values within the swapping range are equal. From the perspective of k-anonymity [35, 37], this means that the original data set was already indistinguishable for sets of k=2 records; that is, it is already *probabilistically-2-anonymous* and, also, *2-anonymous*. In this case, for k=2, the original data

set does not need to be modified to achieve the desired level of protection, and so does our algorithm, which enforces *probabilistic k-anonymity*; otherwise, unnecessary information loss would occur.

For k=5 or 10, which are still below the average frequency of attribute categories, we obtain very small (albeit not null) errors, whereas for k ≥ 20, which exceed the average frequencies, differences are more noticeable. In all cases, the semantic features of the data (in particular the strong attribute correlation) are preserved proportionally to the desired permutation level.

Moreover, we also tested how Algorithm 3 behaves with more than two attributes. For such purpose, we added a third attribute (X^c = *secondary diagnosis*) to *Dataset2*. The resulting data set, named *Dataset3*, presents the following correlations between attribute pairs: $sdCor(X^a, X^b)$ =0.6392, $sdCor(X^a, X^c)$ =0.4229 and $sdCor(X^b, X^c)$ =0.3703. Evaluation metrics are depicted in Table 6.

Table 6. *Dataset3*: evaluation metrics of three rank-swapped attributes (X^a = *principal diagnosis*, X^b = *medical procedure*, X^c = *secondary diagnosis*) with the Algorithm 3.

Metric	***k*=2**	***k*=5**	***k*=10**	***k*=20**	***k*=50**	***k*=100**
$RMSE\ (X^a, X^{a*})$	0	0.0106	0.0322	0.0866	0.2412	0.3951
$RMSE\ (X^b, X^{b*})$	0	0.0309	0.0890	0.1653	0.2409	0.3640
$RMSE\ (X^c, X^{c*})$	0.0011	0.0579	0.1649	0.2948	0.4174	0.5035
$sdCor(X^{a*},X^{b*})$	0.6392	0.6373	0.6319	0.6185	0.5276	0.3753
$\lvert sdCor(X^{a*},X^{b*}) - sdCor(X^a,X^b)\rvert$	0	0.0019	0.0073	0.0207	0.1116	0.2639
$sdCor(X^{a*},X^{c*})$	0.4229	0.4220	0.4191	0.4031	0.3090	0.2291
$\lvert sdCor(X^{a*},X^{c*}) - sdCor(X^a,X^c)\rvert$	0	0.0009	0.0038	0.0198	0.1136	0.1938
$sdCor(X^{b*},X^{c*})$	0.3703	0.3644	0.3577	0.3432	0.2828	0.2118
$\lvert sdCor(X^{b*},X^{c*}) - sdCor(X^b,X^c)\rvert$	0	0.0059	0.0126	0.0271	0.0875	0.1585

Even though the number of attributes to manage is larger, we can see that their correlations are largely preserved by Algorithm 3, and that the results (RMSEs and correlations) for attributes X^a and X^b are similar, albeit slightly worse, to those yielded when just two attributes are considered (Table 5). Only when we increase the k parameter to very large values (above 50), we observe more significant differences because, with more attributes to consider, records in the swapping intervals tend to be more heterogeneous on attribute basis.

5. Conclusions and future work

In this paper, we have presented semantically-grounded alternatives to the standard rank swapping mechanism that are capable of protecting nominal data while better preserving their semantic features. In particular, we have proposed solutions to protect individual nominal attributes and multivariate data sets. The multivariate solution, which constitutes the core offering of our work, is of great interest for practitioners and data analysts, because it offers *ex ante* privacy guarantees on the data protection (*probabilistic k-anonymity*), while reasonably preserving the semantic association among the attributes. In this way, the inferences extracted from the semantic analysis of non-independent attributes protected with our method will be similar to those drawn from the original data.

The empirical study carried on real patient data consisting on several non-independent nominal attributes has shown that our methods are capable of permuting values consistently with the desired level of protection, and incurring in an information loss much lower than non-semantic swapping methods. Another strength of our proposal is that the multivariate solution is able to largely preserve the semantic correlation between attributes for the typical swapping ranges, while this is totally broken by non-semantic swapping. These benefits, together with the preservation of all univariate features, such as the mean, variance, frequency distribution, outlying values, granularity and cardinality, make our method yields protected data that are useful for a larger spectrum of data analyses in comparison with other well-known perturbative data protection mechanisms, such as data microaggregation [34] (which make data more homogeneous and, therefore, reduce their variance and/or granularity), or noise addition [43] (which alters the dispersion of the sample).

As future work, thanks to the mathematical consistence of our approach, we plan to integrate our methods with the standard numerical ones, so that we can treat heterogeneous data involving numerical and nominal attributes in an integrated way. Finally, other semantic distance measures [45] exploiting one or several knowledge sources [46] may be considered to better capture the semantics of nominal attributes.

Acknowledgements and disclaimer

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”).

References

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

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

[3] B. Malin, L. Sweeney, How (not) to protect genomic data privacy in a distributed network: using trail re-identification to evaluate and design anonymity protection systems, Journal of Biomedical Informatics 37 (3) (2004) 179-192.

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

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

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

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

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

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

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

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

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

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

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

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

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

[18] M. Rodriguez-Garcia, M. Batet, D. Sánchez, Semantic Noise: Privacy-Protection of Nominal Microdata through Uncorrelated Noise Addition, in: Proceedings of the 27th IEEE International Conference on Tools with Artificial Intelligence, ICTAI 2015, Vietri sul Mare, Italy, 2015, pp. 1106-1113.

[19] M. Rodriguez-Garcia, M. Batet, D. Sánchez, A Semantic Framework for Noise Addition with Nominal Data, Knowledge-Based Systems 122 (C) (2017) 103-118.

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

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

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

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

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

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

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

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

[28] 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, Washington, D.C., USA, 2004, pp. 652-659.

[29] 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.

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

[32] 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.

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

[34] J. Domingo-Ferrer, J.M. Mateo-Sanz, Practical Data-Oriented Microaggregation for Statistical Disclosure Control, IEEE Transactions on Knowledge and Data Engineering 14 (1) (2002) 189-201.

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

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

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

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

[39] HIPAA. Health Insurance Portability and Accountability Act, 2004. http://www.hhs.gov/ocr/hipaa/.

[40] C.M. Yang, H.C. Lin, C. Polun, W.S. Jian, Taiwan's perspective on electronic medical records' security and privacy protection: Lessons learned from HIPAA, Computer Methods and Programs in Biomedicine 82 (3) (2006) 277-282.

[41] M. Rodriguez-Garcia, D. Sánchez, M. Batet, Perturbative Data Protection of Multivariate Nominal Datasets, in: Privacy in Statistical Databases. UNESCO Chair in Data Privacy, Proceedings of International Conference, PSD 2016, Dubrovnik, Croatia, Springer International Publishing, 2016, pp. 94-106.

[42] 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.

[43] 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.

[44] B.B. Rieger, On Distributed Representations in Word Semantics, Forschungsbericht TR-91-012, International Computer Science Institute (ICSI) (1991).

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

[46] 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