Semantic Noise: Privacy-protection of Nominal Microdata through Uncorrelated Noise Addition

Mercedes Rodriguez-Garcia*, Montserrat Batet† and David Snchez*

*UNESCO Chair in Data Privacy
Department of Computer Engineering and Mathematics
Universitat Rovira i Virgili (Tarragona, Spain)
Email: mercedes.rodriguez@uca.es, david.sanchez@urv.cat

†Internet Interdisciplinary Institute (IN3)
Universitat Oberta de Catalunya (Castelldefels, Spain)
Email: montserrat.batet@urv.cat

***Abstract*—Personal data are of great interest in statistical studies and to provide personalized services, but its release may impair the privacy of individuals. To protect this privacy, in this paper, we present the notion and practical enforcement of semantic noise, a semantically-grounded version of the numerical uncorrelated noise addition method, which is capable of masking textual data while properly preserving their semantics. Unlike other perturbative masking schemes, our method can work with both datasets containing information of several individuals and single data. Empirical results show that our proposal provides semantically-coherent outcomes preserving data utility better than non-semantic perturbative mechanisms.**

***Keywords*-data privacy; statistical disclosure control; noise addition; ontologies; nominal microdata**

I. Introduction

Data of individuals arising from surveys or electronic records are of great interest for public and private organizations. The publication of this information allows conducting a variety of statistical studies, for instance, on health, education, trade preferences, living conditions or employability.

This information can be released in two main ways [1]: as *macrodata*, which consist of aggregated values, or as *microdata*, where each record details the attributes of a single individual. Unlike macrodata, microdata confer flexibility to perform a personalized analysis. However, the publication of microdata may compromise the individuals' privacy. Government agencies and current legislation on data protection emphasize the need of protecting personal data from disclosure. In this direction, a de-identification process is used to generate non-identifiable datasets. The simplest strategy of de-identification consists in removing identifying attributes, such as identity numbers or names, before releasing microdata. However, some studies [1] [2] show that combinations of certain non-identifying attributes, known as quasi-identifiers (e.g. occupation, sex, ZIP), may be linked with external data sources (e.g. voter registration) to re-identify individuals. Nowadays, the re-identification process allows data brokers to compile and aggregate individuals information, make inferences about their habits or preferences and share the outcomes with third parties [3].

A. Related work

To protect quasi-identifiers, different methods have been proposed within the discipline of Statistical Disclosure Control (SDC) [4]. Among them, perturbative methods are the most widespread, which include noise addition, microaggregation, rank swapping or data shuffling. These mechanisms generate a modified version of the original dataset by distorting quasi-identifying attribute values. Ideally, the masked dataset should retain its analytical utility as much as possible without disclosing confidential information or jeopardizing the privacy of the individuals. Therefore, masking methods pursue a twofold objective:

- Minimize the risk of disclosure. The masking method must properly distort the quasi-identifiers to prevent linking them with external datasets.
- Minimize the information loss to maximize the analytical utility of the data. In this way, a statistical analysis on the masked dataset should not differ significantly from the same analysis on the original dataset.

Many perturbative methods achieve an anonymous version of the dataset by creating groups of indistinguishable records [4] according to some anonymization property. For example, to fulfill k-anonymity [5], masking methods build clusters of k indistinguishable records by replacing the original values of each record with the representative value of the cluster to which it belongs (e.g. cluster mean). These methods should receive a homogenous record set as input, because the larger the heterogeneity between records, the larger the information loss resulting from making them indistinguishable.

On the other hand, noise addition methods, which distort original values with random noise, are able to deal with records individually. This feature is very useful in certain scenarios, as in the online anonymization of transac-tional data, especially if data are generated individually via streaming. A representative example is a user performing queries to a Web Search Engine (WSE), which profiles her according

to such queries to provide personalized search services, but also for marketing purposes. In this scenario, the user desires to protect the privacy of her profile with respect to the WSE while not impairing the WSE functionalities (e.g., query disambiguation [6], query suggestion and refinement [7]). Because the generated profiles may fully characterize the personal features of the users [8] [9], it is desirable to add some uncertainty to the user's queries. In this regard, noise addition could create fake but plausible queries from the original ones in a controlled way. This would help to hide the real user details while preserving, as much as possible, the WSE functionalities.

Noise addition has also gained relevance in recent years, thanks to the popularization of the ε-differential privacy model [10], whose enforcement usually relies on Laplacian noise. In this model, the outcomes are protected by making them insensitive (via random noise addition) to changes in one input record, with a probability depending on ε.

Another important consideration is the data type on which a perturbative masking method can operate. Data types can be categorized as:

- Continuous. A datum is continuous if it is numerical and admits arithmetical operations, e.g., age.
- Ordinal categorical. A datum is ordinal categorical if even though it is textual, it admits order relationships, e.g., clothing textual size. Note that arithmetic operations do not make sense with this data type.
- Nominal categorical. A datum is nominal categorycal if it is textual and does not admit neither order relationships nor arithmetic operations. Much of the information used by data brokers for categorizing individuals is of nominal type [3], e.g., occupation, education or personal interests.

Most perturbative masking methods have been designed to deal with continuous data and, in some cases, with ordinal categorical data [4]. Unlike the previous data types, nominal categorical data are finite, discrete, textual and non-ordinal. In this scenario, it is generally not possible to carry out the arithmetical data transformations required in the masking process. Moreover, as the utility of nominal data is closely related to the preservation of their semantics [11], data transformations require from operators that consider the meaning of words [12].

In this sense, only the microaggregation method has been adapted to work with nominal data and obtain semantically-coherent outcomes [13] [14]. The proposed solutions are based on exploiting the semantic knowledge modeled in ontologies. Ontologies are structures that formally describe concepts of a domain of knowledge and the relationships between them. Some ontologies are of general purpose, as WordNet [15] that tries to model knowledge of the world, and others are of specific domains, as MeSH [16] that models clinical knowledge.

Because of their mathematical roots, noise addition methods are inherently focused on numerical data [17]. However, within the context of differential privacy, some mechanisms have been proposed to deal with discrete data (either discrete numbers or categorical values): the geometric mechanism (which offers a discrete probability distribution alternative to the continuous Laplace distribution) [18], and the exponential mechanism (which probabilistically chooses the output of a discrete function according to the input dataset and a quality criterion while preserving ε-differential privacy) [19]. However, both of them rely on the data distribution rather than on the actual semantics of the values. This makes them more suitable for discrete numerical values, rather than nominal categorical ones. From a semantic perspective, [20] suggests the adequacy of a noise addition to protect individual textual documents, but does not specify its calculation.

B. Contributions and plan

In this paper, we present the notion and practical enforcement of *semantic noise*, a semantically-grounded version of the numerical noise addition method, which is capable of masking nominal data while properly preserving their semantics. The contributions of our work are:

- The exploitation of the formal knowledge modeled in ontologies in order to properly capture and manage the semantics of the values to be masked during the noise addition process.
- An adaptation of the statistical operators used in the standard noise addition mechanism to the semantic domain, so that data perturbation is done in coherency with data semantics.
- A semantically-grounded algorithm to add uncorrelated noise to individual attributes, with a specific heuristic to better preserve the meaning of the data.
- A set of empirical experiments with a reference dataset and a comparison with random methods regarding the preservation of the analytical utility (semantic).

The rest of the paper is organized as follows. Section II provides the background on uncorrelated noise addition. Section III describes our proposal. Section IV details the experiments and discusses the empirical results. Section V contains the conclusions and provides some lines of future research.

II. Background on uncorrelated noise addition

Uncorrelated noise addition is a perturbative SDC method that is a priori only suitable to mask numerical data. The initial idea was proposed by Conway [21] and tested thoroughly by Spruill [22] and Kim [23]. This scheme is based on adding sequences of normally distributed random noise to attributes from an input dataset. The outcome is a masked dataset where each anonymized attribute has a mean roughly equal to the original one and a configurable proportional variance.

Following the notation used by Brand in a comprehensive survey about noise addition [17], the input dataset *X* is treated as a set of *p* attributes (or variables), each one corresponding to a different feature of the described individual:

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

where $X_j = \{x_{1j}, \ldots, x_{ij}, \ldots, x_{nj}\}$ is the j-th attribute (or j-th variable) of the dataset and x_{ij} is the value of the attribute *j* corresponding to the individual/record *i*.

For masking 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 masked attribute, X_j is the original attribute and $\varepsilon_j = \{\epsilon_{1j}, \ldots, \epsilon_{ij}, \ldots, \epsilon_{nj}\}$ is the noise sequence. $X_j \sim (\mu_j, \sigma_j^2)$ is a vector with mean μ_j and variance σ_j^2 and $\varepsilon_j \sim N(0, \sigma_\varepsilon^2)$ is a vector of normally distributed random errors with mean zero and variance σ_ε^2. The error variance σ_ε^2 is proportional to the original attribute variance as follows:

$$\sigma_\varepsilon^2 = \alpha\sigma_j^2, \quad \alpha > 0 \tag{3}$$

The factor α determines the amount of applied noise, whose value usually ranges between 0.1 and 0.5 [24]. The higher the α, the higher the masking level, and thus of privacy protection; but also, the lower the data utility. Thus, the factor α defines the trade-off between privacy protection and utility preservation. Note that if $\alpha > 0.5$, more than 50% of the variation in the masked data is due to the added noise and, as a consequence, data tend to become marginal.

From the foregoing it follows that the method preserves the mean of original data and keeps the variance proportional in a factor 1+α:

$$\mu_z = \mu_j + \mu_\varepsilon = \mu_j \tag{4}$$

$$\sigma_z^2 = \sigma_j^2 + \sigma_\varepsilon^2 = (1+\alpha)\sigma_j^2$$

In the case of masking multiple attributes, Tendick [24] and Muralidhar [25] state that, given the uncorrelated character of the method, the noise must be applied to each attribute independently. Accordingly, the method will perturb a different variable at each step without considering the noise applied to previous variables. For this reason,

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

that is, the covariance between any two different noise variables ε_t (added to the attribute X_t) and ε_l (added to the attribute X_l) is zero, which means that correlations between masked variables are not preserved. As a result, the method is suitable for statistical analysis over attributes but not over records.

III. Semantic uncorrelated noise addition method

In this section, we propose a semantically-grounded method to mask nominal attributes of a dataset. The method uses uncorrelated noise addition in combination with the semantic knowledge provided by an ontology. Its operation is based on replacing the original values (i.e., textual terms) of each attribute by other concepts from the same taxonomic domain, which are as semantically distant as the random noise calculated from a specific normal distribution. In this manner, original terms are replaced by semantically similar ones, whose similarity is proportional to the desired privacy protection level. In our approach, data utility is preserved because the semantically-oriented mean and variance measures we define and use during the noise addition process carefully consider the meaning of the data. In Section III.A we present the notion of nominal domain based on an underlying ontology and discuss which semantic distance measures are suited to compare terms. Section III.B defines a semantic version of the statistical operators used in uncorrelated noise addition. Section III.C proposes an algorithm to mask nominal datasets through uncorrelated noise addition.

A. Ontology-based domain

Unlike numerical data, the domain of nominal data is finite, discrete and non-ordinal. This domain can be expressed either as an unstructured term list or as a hierarchically structured set of concepts in an ontology. The former case omits data semantics and, as a consequence, the masking process over nominal attributes may produce outcomes with a significant information loss. The latter case takes into account the meaning of nominal data thanks to the semantic knowledge provided by an ontology. An ontology is a structured knowledge source that explicitly and consensually represents the concepts and the semantic interrelations of a domain [26]. Its structure is a directed graph in which concepts are interrelated mainly by means of taxonomic links (is-a) and, in some cases, non-taxonomic links (as part-of) [27] [28]. By relying on ontologies, operations performed over nominal attributes can exploit the modeled semantic relationships between concepts to provide results that are semantically coherent with the original terms and, thus, better preserve the utility of the masked data [29].

Our proposal uses ontologies to capture the underlying semantics of nominal data. In order to ensure the generality of the method, we only consider taxonomical relations because they are available in any ontology and constitute the backbone of its knowledge structure [30]. In this context, the ontology is seen as a taxonomic tree in which concepts (nodes) are interrelated by means of is-a links (edges).

Before applying noise, our method requires of mapping terms of the input dataset to concepts in a taxonomy. Let *X* be a dataset with *n* records and *p* nominal attributes whose terms have been modeled in a taxonomy τ. Following the

notation of Section II, we represent the attribute X_j from the dataset X as $X_j = \{x_1, \ldots, x_i, \ldots, x_n\}$, where x_i is the value of the individual i mapped to a concept in τ.

By applying semantic noise to mask the terms of a given attribute, it is possible to obtain new concepts from τ different from the original ones. To ensure the semantic coherence of the results, the new concepts must belong to the domain of the attribute, e.g., if the original term is a disease, the masked term must also be a disease. Thus, it is necessary to define the domain of an attribute and specify what concepts of the taxonomy τ are candidates to participate in the masking process of that attribute.

Definition 1. The *domain of an attribute* X_j, denoted by $D(X_j)$, is defined as the set of all concepts belonging to the category of X_j.

$$D(X_j) = \{c \in Category(X_j)\} \tag{6}$$

e.g., if the category of the attribute X_j is disease, its domain $D(X_j)$ is the list of all the possible diseases.

On the other hand, we define $\tau(D(X_j))$ as the minimum hierarchy extracted from τ that includes all terms in $D(X_j)$. Formally:

Definition 2. The *taxonomy associated to the domain* $D(X_j)$ is the hierarchy $\tau(D(X_j))$ from τ created by the union of all the branches between each concept c_i in $D(X_j)$ and the *Least Common Subsumer* of $D(X_j)$.

$$\tau(D(X_j)) = \bigcup_{c_i \in D(X_j)} \{branch(c_i, LCS(D(X_j)))\} \tag{7}$$

where $branch(c_i, LCS(D(X_j)))$ is the set of concepts from τ between c_i and $LCS(D(X_j))$ connected by *is-a* links, including themselves, and $LCS(D(X_j))$ is the deepest ancestor from τ that subsumes all terms of $D(X_j)$. A concept c_i subsumes a concept c_j, i.e., $c_i \geq c_j$, if c_i is a generalization/taxonomic ancestor of c_j or c_i and c_j are the same concept.

Many of the operations carried out in the noise addition process need to semantically compare two terms, e.g. for assessing how distinct the masked term must be from the original one according to the amount of noise that must be added. For this purpose, we use the notion of semantic distance, a function that quantifies the semantic differences between terms modeled in a taxonomy. Different functions to measure the semantic distance have been proposed in the literature [27]. A suitable semantic distance to be applied in the noise addition scenario should: i) output values normalized in the range [0..1], where the boundary value 0 represents the minimum distance, i.e. the terms are perfect synonyms or the same, and the boundary value 1 represents the maximum distance, ii) perform a linear assessment of the distance, which is suitable for the uniform distribution used to generate noise values, and iii) be computationally efficient. The ontology-based semantic similarity measure proposed by Wu&Palmer [31] fulfills the above features, provides high accuracy and presents low computational cost in comparison with other measures that deal with corpora or consider all the hyponyms of a concept [27]. According to Wu&Palmer the semantic similarity between two concepts modeled in a taxonomy is defined as follows,

$$sim_{wp}(c_1, c_2) = \frac{2 \times depth(LCS(c_1, c_2))}{depth(c_1) + depth(c_2)} \tag{8}$$

where the *depth* of a given term *c* is the number of concepts linked between *c* and *root*, including themselves, and *root* is the top node of the taxonomy.

In the context of our proposal, the taxonomy is limited to $\tau(D(X_j))$. Thus, c_1, c_2, $LCS(c_1, c_2) \in \tau(D(X_j))$ and *root* is the top node of $\tau(D(X_j))$. According to definition 2, *root* coincides with $LCS(D(X_j))$. As sim_{wp} evaluates the similarity between concepts, we formulate sd_{wp} to compute the desired semantic distance, as follows:

$$sd_{wp}(c_1, c_2) = 1 - sim_{wp}(c_1, c_2) \tag{9}$$

B. Semantic statistical operators

The masking process through uncorrelated noise addition requires the computation of two statistical measures: the mean and the variance of each attribute. In our method, these measures must be adapted to nominal data for two reasons:

- Standard arithmetical operations cannot be directly applied on nominal data.
- Measures should capture the semantics of nominal data in order to truly preserve the data utility.

For the computation of the mean, we rely on the notions of semantic marginality [29] and centroid [12]. Basically, these works define that the mean of a set of concepts is the least marginal/distant concept of the set. In the context of semantic noise, we propose the following definition:

Definition 3. The *mean of a nominal attribute* X_j is the concept c from $\tau(D(X_j))$ that minimizes the marginality with respect to X_j.

$$\mu_{X_j} = \underset{c \in \tau(D(X_j))}{\arg\min} \; (m(c, X_j)) \tag{10}$$

where $m(c, X_j)$ is the marginality of the term *c* with respect to the set X_j, and is computed as the sum of the semantic distances between *c* and each term x_i in X_j,

$$m(c, X_j) = \sum_{x_i \in X_j} sd(c, x_i) \tag{11}$$

Note that any term in $\tau(D(X_j))$ may be the *mean* of the attribute, i.e., μ_{X_j} does not necessarily have to match a term from X_j. In this manner, the set of mean candidates increases and it is possible to obtain a better approximation of the mean that, due to the inherent nature of nominal data, is discrete.

Similarly to the arithmetic variance, the variance of a nominal dataset should take into account the differences

between each term of the set and the mean. From a semantic perspective, these differences are computed using semantic distances [32].

Definition 4. The *variance of a nominal attribute* X_j is the average of squared semantic distances between each concept x_i in X_j and the mean μ_{X_j}.

$$\sigma^2_{X_j} = \frac{\sum_{x_i \in X_j} sd(x_i, \mu_{X_j})^2}{n} \tag{12}$$

where *n* is the number of terms in X_j.

C. Semantic noise addition

Thanks to the availability of a semantic distance measure that fulfills our requirements (Section III.A) and the semantically-grounded versions of the mean and variance measures (Section III.B), we can now adapt the noise addition method to nominal data. By doing so, our method aims the following purposes in order to minimize the loss of semantics of the masking process: i) to preserve, as much as possible, the mean of the original data, ii) to obtain a data dispersion proportional to the variance of the original data and the noise magnitude, and iii) to replace original values by masked terms within a semantic distance coherent with the desired distortion level.

In the numerical domain, arithmetical noise/error represents a magnitude to be added/subtracted from the original values. Thus, this error represents the numerical distance between the original and masked values. Likewise, in the semantic domain, error values should correspond to semantic distances. These distances are used to replace the original terms by other concepts in the underlying taxonomy that are as semantically distant as defined by the error magnitude. To preserve the analytical utility of the data, after adding noise to an attribute, the mean shall remain the same. In the numerical domain, if a positive error is added to an original value greater (lower) than the mean, the new value will get away from (closer to) the mean in the same magnitude; on the contrary, if the error is negative, the new value will get closer to (away from) the mean. Since the error is normally distributed around zero, the magnitude of the accumulated additions and subtractions with respect to the mean will compensate each other, thus keeping the value of the mean. However, nominal data presents an issue: it lacks a total order, i.e., there are as many orders as reference points, which could be any concept in the taxonomy. As a consequence, if we move away a certain distance from a concept, we cannot guarantee that we will be also closer to or farther from the mean concept with the same distance. Thus, if we use the original terms as reference points to apply the error values/semantic distances, we will respect the absolute errors regarding such values, but we cannot ensure that the mean will be preserved.

To solve this issue, we propose an interpretation of the error sign that guides the replacement of terms in the masking process towards the preservation of the mean:

- If the error ϵ_i is positive, the concept c in $\tau(D(X_j))$ that will replace the original term x_i must be farther from the mean than x_i, i.e., $sd(c, \mu_{X_j}) > sd(x_i, \mu_{X_j})$.
- If the error ϵ_i is negative, the concept c in $\tau(D(X_j))$ that will replace the original term x_i must be closer to the mean than x_i, i.e., $sd(c, \mu_{X_j}) < sd(x_i, \mu_{X_j})$.

The idea behind this strategy is to balance the number of movements towards and away from the mean. Since the magnitude of the positive and negative errors should be equivalent, this strategy will tend to preserve the mean.

As it was stated in Section II, to mask a multivariate dataset of *p* nominal attributes through uncorrelated noise, the noise addition method must be applied to each attribute independently.

Formally, the data masking algorithm is shown in Algorithm 1. First, the taxonomy $\tau(D(X_j))$ associated to the domain of the attribute X_j is obtained from τ following the procedure detailed in Section III.A. After that, the terms of X_j are mapped to concepts of $\tau(D(X_j))$. In lines 4 and 5, the semantic mean μ_{X_j} and the variance $\sigma^2_{X_j}$ of X_j are computed by using (10) and (12). Then, according to Section II, we generate the noise sequence consisting on $n = |X_j|$ random numbers $\varepsilon_j = \{\epsilon_1, \ldots, \epsilon_i, \ldots, \epsilon_n\}$ that follows a normal distribution with mean 0 and variance $\sigma^2_\varepsilon = \alpha\sigma^2_{X_j}$, where α determines the desired degree of semantic noise.

Algorithm 1 Data masking with semantic noise

Input: X: original dataset with p attributes,
τ: taxonomy, α: semantic noise level
Output: Z: masked dataset

1: **for all** X_j in X **do**
2: $\tau(D(X_j)) \leftarrow$ obtain_taxonomy$(D(X_j), \tau)$
3: $X_j \leftarrow$ map$(X_j, \tau(D(X_j)))$
4: $\mu_{X_j} \leftarrow$ compute_mean$(X_j, \tau(D(X_j)))$
5: $\sigma^2_{X_j} \leftarrow$ compute_variance$(X_j, \mu_{X_j}, \tau(D(X_j)))$
6: $\sigma^2_\varepsilon \leftarrow \alpha\sigma^2_{X_j}$
7: $\varepsilon_j \leftarrow$ generate_noise_vector(σ^2_ε) // $\varepsilon_j \sim N(0, \sigma^2_\varepsilon)$
8: **for all** x_i in X_j **do**
9: **if** $\epsilon_i = 0$ **then**
10: $z_i \leftarrow x_i$
11: **else if** x_i matches the mean μ_{X_j} **then**
12: $z_i \leftarrow \underset{c \in \tau(D(X_j))}{\arg\min} \{sd(c, x_i) | sd(c, x_i) \geq |\epsilon_i|\}$
13: **else if** ϵ_i is positive **then**
14: $z_i \leftarrow \underset{c \in \tau(D(X_j))}{\arg\min} \binom{sd(c,x_i) | sd(c,x_i) \geq |\epsilon_i| \wedge}{sd(c,\mu_{X_j}) > sd(x_i,\mu_{X_j})}$
15: **else if** ϵ_i is negative **then**
16: $z_i \leftarrow \underset{c \in \tau(D(X_j))}{\arg\min} \binom{sd(c,x_i) | sd(c,x_i) \geq |\epsilon_i| \wedge}{sd(c,\mu_{X_j}) < sd(x_i,\mu_{X_j})}$
17: **end if**
18: **end for**
19: **end for**
20: **return** Z

In order to compute the masked terms z_i, we add the noise to each original term x_i by replacing them by a concept c in $\tau(D(X_j))$, whose semantic distance computed by using (9) ideally matches the error magnitude $|\epsilon_i|$, i.e. $sd(c, x_i) = |\epsilon_i|$, and gets closer to or away from the mean μ_{X_j} according to the sign of ϵ_i. However, as nominal values are discrete, it may happen that there is not a concept at the exact required distance. In such case, the method selects the concept that, while exceeding the error magnitude, minimizes its distance with x_i. At this step, the error sign interpretation proposed above is used (line 14 when ϵ_i is positive and line 16 when ϵ_i is negative). Obviously, if ϵ_i is zero, the masked term z_i is exactly x_i. Finally, when x_i matches μ_{X_j}, the masked term z_i is just the concept c in $\tau(D(X_j))$ that minimizes its distance with x_i (line 12).

In any case, if a concept c in $\tau(D(X_j))$ with $sd(c, x_i) \geq |\epsilon_i|$ does not exist (i.e., we cannot get farther enough within $\tau(D(X_j))$), then we select the term that best approximates the condition. Because of this and due to the need to discretize error values, the accuracy of the noise-added outcomes would depend on the size and granularity of the underlying taxonomy.

IV. Experiments

In this section, we evaluate the semantic noise addition method proposed in Section III and compare its results with two non-semantic methods based on data randomization and data distribution:

- A *naïve randomization*, in which original values are randomly replaced by other values of the same dataset.
- A *probabilistic randomization*, in which the probability of selecting a value as a replacement corresponds to the probability of appearance of that value in the sample. Because the distribution of the data is considered during the randomization, the outcome will better preserve the statistical features of the data.

In the experiments, we have used the nominal attribute *occupation* from the *Adult Census* dataset, which is publicly available in the UCI repository [36]. The attribute describes the occupation of a set of 30,242 individuals, after removing records with missing values. The 14 textual values of the attribute have been mapped to concepts from the WordNet ontology [33]. After mapping, the distribution for the attribute *occupation* is $X_j = \{$*protector*(644), *functionary*(3295), *salesperson*(3584), *technician*(912), *carrier*(1572), *farmer*(989), *cleaner*(1350), *clerk*(3721), *executive*(3992), *craftsman*(4030), *specialist*(4038), *serviceman*(6), *factory worker*(1966), *housekeeper*(143)$\}$. According to the semantic statistical measures proposed in Section III.B, the mean of the sample is μ_{X_j} = *employee* (since values are quite balanced among the different categories, the mean of the sample tends to be a general concept of the taxonomy) and the variance is $\sigma^2_{X_j} = 0.23$.

To quantify the accuracy of the noise-added masked results, we have considered the following semantic metrics:

1) The semantic distance (9) between the mean of the masked sample and of the original sample. A value of 0 indicates that the mean has been perfectly preserved during the masking process.
2) The absolute difference between the actual variance of the masked sample and the expected variance after adding noise (4). In our method, the latter is a function of the original variance $\sigma^2_{X_j}$ and the parameter α.
3) The root mean square error (RMSE), measured as the root average square semantic distance between original and masked value pairs. This measures the overall loss of semantics in the masked sample, which should be similar to the target error resulting from the desired magnitude of noise to be added.

Results for these metrics are presented by our method for different values of α in Table I. Since α determines the amount of applied noise, the higher the α, the higher the RMSE, and thus the masking level. In all cases, the mean of the masked dataset is preserved regardless of the value of α. The difference between the variance of the masked attribute and the expected variance is maintained around a 20-30% of the parameter α. For nominal data, it would be in general difficult to achieve a null difference because of the need to discretize noise-added values to concepts in the underlying taxonomy and the limited scope of the taxonomy, which may result in truncated values. On the other hand, the *actual RMSEs* show that our method is able to appropriately adapt the distortion process to the configured magnitude of the error (*target RMSEs*, that are, the mean errors of the noise sequences) and thus, to the desired privacy protection level (i.e., in all cases, the *actual RMSE* is greater or equal to the *target RMSE*). The small difference between *actual* and *target RMSEs* are caused again by the need to discretize error values. This difference tends to be higher for small values of α because, when the error components ϵ_i are small, the relative effect of the discretization is more apparent over the absolute magnitude.

In Table II, we compare the accuracy of our approach with the random methods introduced above. For the purpose of a fair comparison, we set the error magnitude for our method to the maximum reasonable value ($\alpha = 1$), trying to match the degree of perturbation added to the values by random methods. However, we can see that the random methods tend to add a significantly larger amount of noise, which is also non-configurable. From the evaluation metrics, we can see that the naïve randomization provides the worst results, with a significant perturbation of the mean of the masked sampled. The probabilistic randomization, on the other hand, shows a behavior that is more similar to that of our method. In this case, the small spectrum of values in the dataset (14 occupation categories) and the large and even balance

Table I
EVALUATION METRICS FOR A SAMPLE OF 100 RECORDS OF THE ADULT OCCUPATION DATASET WITH SEMANTIC NOISE

Metric	α					
	0.1	**0.2**	**0.3**	**0.4**	**0.5**	**1**
μ_{Z_j}	employee					
$sd(\mu_{Z_j}, \mu_{X_j})$	0					
$\lvert\sigma^2_{z_j}\text{-}(1+\alpha)\sigma^2_{x_j}\rvert$	0.03	0.05	0.07	0.10	0.11	0.23
Actual RMSE	0.19	0.23	0.26	0.28	0.30	0.35
Target RMSE	0.12	0.17	0.21	0.24	0.27	0.35

Table II
EVALUATION METRICS FOR NAÏVE RAND., PROBABILISTIC RAND. AND SEMANTIC NOISE FOR THE ADULT OCCUPATION DATASET

Metric	**Naïve**	**Probabilistic**	**Semantic** (α=1)
μ_{Z_j}	skilled worker	employee	employee
$sd(\mu_{Z_j}, \mu_{X_j})$	0.33	0	0
Actual RMSE	0.56	0.54	0.35

of repetitions among the categories configure a favorable scenario for methods based on data distributions.

In order to evaluate the methods with a relatively finer grained and less balanced dataset, we extracted a small sample of 100 records from the attribute *occupation* with the following distribution $X_j = \{$*craftsman*(46), *cleaner*(20), *farmer*(12), *technician*(22)$\}$. The mean is μ_{X_j} = *craftsman* (i.e., a specific value rather than a general concept) and the variance is $\sigma^2_{X_j} = 0.13$. Evaluation results are depicted in Tables III and IV. From Table III, we now observe a greater variability with regard to the preservation of the mean, according to the value of α. In any case, the difference between the original and masked means is small for our method. On the other hand, from Table IV, the nave and probabilistic methods show a significantly larger discrepancy for the mean. Moreover, now, the *actual RMSEs* of the naïve and probabilistic methods are very similar to that of our method; i.e., all the methods are introducing a similar degree of distortion/protection to the data but our method is able to better preserve the semantic features of the sample.

Table III
EVALUATION METRICS FOR A SAMPLE OF 100 RECORDS OF THE ADULT OCCUPATION DATASET WITH SEMANTIC NOISE

Metric	α					
	0.1	**0.2**	**0.3**	**0.4**	**0.5**	**1**
μ_{Z_j}	crafts-man	crafts-man	skilled worker	crafts-man	skilled worker	skilled worker
$sd(\mu_{Z_j}, \mu_{X_j})$	0	0	0.14	0	0.14	0.14
$\lvert\sigma^2_{z_j}\text{-}(1+\alpha)\sigma^2_{x_j}\rvert$	0	0.01	0.03	0.03	0.04	0.07
Actual RMSE	0.14	0.18	0.20	0.23	0.24	0.35
Target RMSE	0.09	0.13	0.15	0.19	0.20	0.29

Table IV
EVALUATION METRICS FOR NAÏVE RAND., PROBABILISTIC RAND. AND SEMANTIC NOISE FOR THE ADULT OCCUPATION DATASET

Metric	**Naïve**	**Probabilistic**	**Semantic** (α=1)
μ_{Z_j}	laborer	technician	skilled worker
$sd(\mu_{Z_j}, \mu_{X_j})$	0.56	0.25	0.14
Actual RMSE	0.37	0.35	0.35

V. CONCLUSIONS AND FUTURE WORK

We have presented a method to mask nominal data with semantic noise, which offers a semantically-grounded alternative to the classic uncorrelated noise addition. Unlike other perturbative methods, our method is able to deal with records individually, which is especially useful in the online anonymization of transactional data originated via streaming. In comparison with noise addition methods based on data distributions, our proposal is able to better preserve the data utility by exploiting the formal semantics modeled in ontologies, and by replacing the original concepts by semantically similar ones according to a controllable (i.e., parameterized) level of protection. It is also important to note that the core of our method (i.e., the mapping of values to ontological concepts and the semantically grounded replacement of noise added values) is not linked to a specific noise distribution or privacy model and thus, it can also be applied to other noise-based mechanisms, such as Laplace noise, which is widely used to enforce ε-differential privacy.

As future work, we plan to further develop the heuristic that guides the masking process so that we can either optimize the preservation of a particular statistic (e.g., the average error or the mean), in case data utility strongly depends on that statistic, or to achieve the best balance between all of them. For correlated multivariate datasets, we also plan to adapt other noise addition mechanisms that are able to preserve the correlations between attributes. Finally, we also plan to use the Laplace probability distribution instead of the Normal one, so that we can offer a semantically-coherent differential privacy enforcing mechanism, embrace its robust and a priori privacy guarantees, and compare our results with the geometric and exponential mechanisms.

ACKNOWLEDGMENT

The authors are solely responsible for the views expressed in this paper, which do not necessarily reflect the position of UNESCO nor commit that organization.

This work was partly supported by the European Commission under the H2020 project CLARUS, by the Spanish Government (through projects ICWT TIN2012-32757, CO-PRIVACY TIN2011-27076-C03-01 and SmartGlacis) and by the Government of Catalonia through grant 2014 SGR 537.

This work was also made possible through the support of a grant from Templeton World Charity Foundation. The opinions expressed in this paper are those of the authors

and do not necessarily reflect the views of Templeton World Charity Foundation.

REFERENCES

[1] V. Ciriani, S. Vimercati, S. Foresti, and P. Samarati, "Microdata protection," in *Secure Data Management in Decentralized Systems*. Springer, 2007, pp. 291–321.

[2] P. Samarati, "Protecting respondents' identities in microdata release," *IEEE Transactions on Knowledge and Data Engineering*, vol. 13, no. 6, pp. 1010–1027, 2001.

[3] E. Ramirez, *et al.*, "Data brokers: A call for transparency and accountability," Federal Trade Commission, Tech. Rep., May 2014.

[4] A. Hundepool, *et al.*, "Microdata," in *Statistical Disclosure Control*. Wiley, 2012, pp. 23–130.

[5] P. Samarati and L. Sweeney, "Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression," Tech. Rep., 1998.

[6] C. Makris, Y. Plegas, and S. Stamou, "Web query disambiguation using pagerank," *JASIST*, vol. 63, no. 8, pp. 1581–1592, 2012.

[7] X. Shi and C. C. Yang, "Mining related queries from web search engine query logs using an improved association rule mining model," *JASIST*, vol. 58, no. 12, pp. 1871–1883, 2007.

[8] A. Viejo and D. Sánchez, "Profiling social networks to provide useful and privacy-preserving web search," *JASIST*, vol. 65, no. 12, pp. 2444–2458, 2014.

[9] D. Sánchez, J. Castellà-Roca, and A. Viejo, "Knowledge-based scheme to create privacy-preserving but semantically-related queries for web search engines." *Information Sciences*, vol. 218, pp. 17–30, 2013.

[10] in *Automata, Languages and Programming*, 2006, vol. 4052.

[11] in *Data Privacy Management and Autonomous Spontaneous Security*, 2011, vol. 6514.

[12] S. Martínez, A. Valls, and D. Sánchez, "Semantically-grounded construction of centroids for datasets with textual attributes," *Knowledge-Based Systems*, vol. 35, pp. 160 – 172, 2012.

[13] S. Martínez, D. Sánchez, and A. Valls, "Semantic adaptive microaggregation of categorical microdata," *Computers& Security*, vol. 31, no. 5, pp. 653 – 672, 2012.

[14] M. Batet *et al.*, "Semantic anonymisation of set-valued data," in *Proc. International Conference on Agents and Artificial Intelligence (ICAART'14)*, vol. 1, 2014, pp. 102–112.

[15] C. Fellbaum, *WordNet: an electronic lexical database*. MIT Press, 1998.

[16] S. Nelson, D. Johnston, and B. L. Humphreys, "Relationships in medical subject headings," in *Relationships in the Organization of Knowledge*. Kluwer Academic, 2001, pp. 171–184.

[17] R. Brand, "Microdata protection through noise addition," in *Proc. Inference Control in Statistical Databases, From Theory to Practice*, 2002, pp. 97–116.

[18] A. Ghosh *et al.*, "Universally utility-maximizing privacy mechanisms," in *Proc.Annual ACM Symposium on Theory of Computing (STOC'09)*, 2009, pp. 351–360.

[19] F. McSherry and K. Talwar, "Mechanism design via differential privacy," in *Proc. Annual IEEE Symposium on Foundations of Computer Science (FOCS'07)*, 2007, pp. 94–103.

[20] D. Abril, G. Navarro-Arribas, and V. Torra, "On the declassification of confidential documents," in *Modeling Decision for Artificial Intelligence*. Springer, 2011, vol. 6820, pp. 235–246.

[21] R. W. Conway and D. Strip, "Selective partial access to a database," Cornell University, Tech. Rep., 1976.

[22] N. Spruill, "Protecting confidentiality of business microdata by masking," in *CRC 523*. Public Research Institute, 1984.

[23] J. Kim, "A method for limiting disclosure in microdata based on random noise and transformation," in *Proc. of the Section on Survey Research Methods*, 1986, pp. 370–374.

[24] P. Tendick, "Optimal noise addition for preserving confidentiality in multivariate data," *Journal of Statistical Planning and Inference*, vol. 27, no. 3, pp. 341 – 353, 1991.

[25] K. Muralidhar and R. Sarathy, "Security of random data perturbation methods," *ACM Transactions on Database Systems*, vol. 24, pp. 487–493, 2000.

[26] M. Batet and D. Sánchez, "Review on semantic similarity," in *Encyclopedia of Information Science and Technology*, 2014, pp. 7575–7583.

[27] D. Sánchez *et al.*, "Ontology-based semantic similarity: A new feature-based approach," *Expert Systems with Applications*, vol. 39, no. 9, pp. 7718 – 7728, 2012.

[28] A. Weichselbraun *et al.*, "Discovery and evaluation of non-taxonomic relations in domain ontologies," *International Journal of Metadata, Semantics and Ontologies*, vol. 4, no. 3, pp. 212–222, 2009.

[29] J. Domingo-Ferrer, D. Sánchez, and G. Rufian-Torrell, "Anonymization of nominal data based on semantic marginality," *Information Sciences*, vol. 242, pp. 35 – 48, 2013.

[30] D. Sánchez and M. Batet, "Semantic similarity estimation in the biomedical domain: An ontology-based information-theoretic perspective," *Journal of biomedical informatics*, vol. 44, no. 5, pp. 749–759, 2011.

[31] Z. Wu and M. Palmer, "Verbs semantics and lexical selection," in *Proc. of the Annual Meeting of the Association for Computational Linguistics*, 1994, pp. 133–139.

[32] M. Batet and D. Sánchez, "A semantic approach for ontology evaluation," in *Proc.IEEE International Conference on Tools with Artificial Intelligence (ICTAI'14)*, 2014, pp. 138–145.

[33] ICS, "Uci knowledge discovery in databases archive," http://kdd.ics.uci.edu/ 2005.