Perturbative data protection of multivariate nominal datasets

Mercedes Rodriguez-Garcia[1], David Sánchez[1], and Montserrat Batet[2]

[1]UNESCO Chair in Data Privacy
Dep. of Computer Engineering and Maths, Universitat Rovira i Virgili, Tarragona, Spain
mercedes.rodriguez@uca.es,david.sanchez@urv.cat
[2]Internet Interdisciplinary Institute (IN3), Universitat Oberta de Catalunya, Castelldefels, Spain
montserrat.batet@urv.cat

Abstract. Many of the potentially sensitive personal data produced and compiled in electronic sources are nominal and multi-attribute (e.g., personal interests, healthcare diagnoses, commercial transactions, etc.). For such data, which are discrete, finite and non-ordinal, privacy-protection methods should mask original values to prevent disclosure while preserving the underlying semantics of nominal attributes and the (potential) correlation between them. In this paper we tackle this challenge by proposing a semantically-grounded version of numerical correlated noise addition that, by relying on structured knowledge sources (ontologies), is capable of perturbing/masking multivariate nominal attributes while reasonably preserving their semantics and correlations.

Keywords: privacy · data protection · semantics · ontologies.

1 Introduction

Personal data are crucial both for business and research purposes. *Microdata* (multivariate records, each one detailing the attributes of an individual) are of especial interest because, unlike *macrodata*, they do not restrict the type and granularity of data analyses. However, publishing *microdata* for secondary use may compromise the privacy of the individuals. To protect privacy while ensuring that data are analytically useful, data protection methods [1, 2] aim at balancing the trade-off between disclosure risk and data utility preservation.

Even though most protection methods focus on numerical data, nowadays, *nominal* microdata (categorical values or textual responses) account a significant amount of the personal data available from social networks, polls, healthcare records or web browsing [3]. Unlike numbers, nominal data are finite, discrete and, in general, non-ordinal. Thus, it is not possible to apply the arithmetic and statistical operators required for data masking. Moreover, whereas the utility of numerical data is a function of their statistical properties, nominal data utility depends on the preservation of semantics [4]. In this

paper, we tackle this challenging scenario and specifically focus on multivariate nominal datasets, which are especially relevant for research.

1.1 Related Work

Many data protection methods have been proposed within the Statistical Disclosure Control [1] and Privacy Preserving Data Publishing [2] disciplines. For nominal data, non-perturbative methods based on replacing values by generalizations are the most widespread ones. These, however, suffer from several drawbacks that may hamper data utility [5]. First, generalization produces a loss of granularity because input values are replaced by a reduced set of categories/generalizations. Moreover, if outlier values are present in the sample, the need to generalize them may result in very coarse generalizations and, thus, a high loss of information/utility.

Perturbative methods such as data microaggregation, rank swapping, data shuffling or noise addition are free from these drawbacks since they mask data while maintaining their granularity. From these, only microaggregation has been adapted to semantically mask nominal data [6, 7]. Microaggregation is usually employed to enforce *k*-anonymity [8] by building clusters of *k* indistinguishable records and replacing original values with a representative value of the cluster (e.g., the mean). To minimize the information loss associated to the replacement, microaggregation requires static and homogenous data.

By contrast, noise addition is able to mask records individually, which is useful when data are generated as a stream and should be protected on the fly [9]. Moreover, unlike methods based on a homogenous data aggregation, noise addition allows defining per-individual perturbation/masking levels, which is useful to accommodate heterogeneous privacy needs. Finally, noise addition has recently gained more relevance due to the ε-differential privacy model [10], whose enforcement relies on Laplacian noise addition.

Because of its mathematical roots, noise addition has been rarely applied to categorical/nominal data. In [11] the authors propose changing the values of each attribute according to a probability distribution. In [12], it is suggested dividing the categorical attributes into sub-attributes that present natural orders, which are needed for noise addition. In the context of differential privacy, some mechanisms have been proposed to perturb discrete data: the geometric mechanism (a discrete probability distribution alternative to the Laplace one) [13] and the exponential mechanism (which probabilistically chooses the output of a discrete function) [14]. However, both of them rely on the data distribution rather than on semantics; this fact makes them more suitable for discrete numerical values rather than nominal data. In [15] the use of "semantic noise" is suggested to mask text, but its calculation is not specified.

1.2 Contributions

In [16] we presented the notion of *semantic noise* as a semantically-grounded version of numerical uncorrelated noise addition. With this method we were able of masking individual nominal attributes while preserving their semantics better than with perturbation mechanisms based only on data distributions.

In this paper we extend this work to support correlated noise, which is needed to protect non-independent attributes in multivariate datasets while preserving their correlations. As in [16], we exploit ontologies to capture and manage the underlying semantics of nominal values and employ semantic operators alternative to the arithmetical ones used for numerical noise addition (i.e., distance, mean, variance and covariance). Finally, we propose a semantic algorithm to add user-defined correlated noise to pairs of nominal attributes.

The rest of the paper is organized as follows. Section 2 provides background on (numerical) noise addition. Section 3 describes the semantic management of nominal values, and details our semantic correlated noise addition method. Section 4 reports the results of several empirical experiments with nominal multivariate datasets. Section 5 depicts the conclusions and provides some lines of future research.

2 Background

Uncorrelated noise [17] perturbs individual attributes by adding normally distributed random noise. Let us have a dataset X, consisting on a ($n \times p$) matrix with p attributes of n records, where $X_j = \{x_{1j}, \ldots, x_{ij}, \ldots, x_{nj}\}$ is the j^{th} attribute and x_{ij} is the value of the attribute j corresponding to the individual/record i. To perturb the attribute X_j, each value x_{ij} is replaced by a noisy version z_{ij}:

$$Z_j = X_j + \varepsilon_j, \tag{1}$$

where $\varepsilon_j = \{\epsilon_{1j}, \ldots, \epsilon_{ij}, \ldots, \epsilon_{nj}\}$ is the noise sequence. Being $X_j \sim N(\mu_{Xj}, \sigma_{Xj}^2)$ a vector with mean μ_{Xj} and variance σ_{Xj}^2, $\varepsilon_j \sim N(0, \sigma_{\varepsilon j}^2)$ is a vector of normally distributed random errors with mean zero and variance $\sigma_{\varepsilon j}^2$. The error variance $\sigma_{\varepsilon j}^2$ is proportional to the original attribute variance as follows:

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

The factor α determines the amount of noise, whose value usually ranges between 0.1 and 0.5 [18]. This method preserves the mean of the attributes and keeps the variance proportional in a factor of $1+\alpha$; but, since noise is independently applied to each attribute, the covariance between noise-added attributes is zero and, thus, the attribute correlations are not preserved.

To solve this issue, which is crucial for research because most datasets are multivariate and correlated and many data analyses are performed over records rather than attributes, correlated noise addition has been proposed [19]. The noise applied to multiple attributes of a dataset X follows matrix notation:

$$Z = X + \varepsilon, \tag{3}$$

where the $(n \times p)$ data matrix $X \sim N(\mu_X, \Sigma_X)$ follows a multivariate normal distribution with mean the p-dimensional vector μ_X and covariance matrix the $(p \times p)$ matrix Σ_X. The diagonal elements of Σ_X are the variances of the attributes and the off-diagonal elements are the covariances between the attributes. Similarly $\varepsilon \sim N(0, \Sigma_\varepsilon)$ is a $(n \times p)$ matrix with p noise sequences of n error values. The noise matrix ε also follows a multivariate normal distribution with mean the p-dimensional vector 0 and covariance matrix the $(p \times p)$ matrix Σ_ε. The covariance matrix of errors is proportional to the covariance of the data:

$$\sum_\varepsilon = \alpha \sum_X, \quad \alpha > 0 \tag{4}$$

Correlated noise addition preserves the mean of each attribute, keeps the covariance matrix proportional in a factor $1+\alpha$ to the covariance matrix of the original data and maintains the correlation between the attributes.

3 Nominal Data Perturbation with Semantic Correlated Noise

We propose a (perturbative) correlated noise addition method for masking multivariate nominal datasets that exploits ontologies to capture and manage data semantics. The idea is to replace original values by other semantically similar concepts, whose similarity is proportional to the desired magnitude of noise (which states the level of masking to be applied to the data).

3.1 Semantic Management of Nominal Data

To manage the semantics of nominal data we first define the *semantic domain* of the attributes according to the concepts modelled in an ontology τ, which we exploit as knowledge base. An ontology can be viewed as a tree/graph whose nodes depict the concepts from an area of knowledge and the edges detail the semantic relationships between them. Thus, our method assumes the availability of an appropriate ontology for the domain of interest.

Let $A = \{a_1, \ldots, a_i, \ldots, a_n\}$ an attribute of a dataset X of n records, where a_i is the value of the individual i mapped to a concept in τ. The *semantic domain* of an attribute A is defined as the set of concepts belonging to the category of A (e.g. if A contains diseases, $D(A)$ is the set of all diseases).

$$D(A) = \{c \in Category(A)\} \tag{5}$$

Then, we extract the *taxonomy of the semantic domain* $D(A)$ from the ontology τ, as the hierarchy $\tau(D(A))$ including of all concepts that are taxonomic specializations of the Least Common Subsumer of $D(A)$, $LCS(D(A))$; that is, the most specific ancestor in τ that subsumes all concepts in $D(A)$).

$$\tau(D(A)) = \bigcup_{c_i \in \tau} \{c_i \mid LCS(D(A)) \geq c_i\} \tag{6}$$

Contrary to the numerical domain, $D(A)$ is finite, discrete and, in general, non-ordinal. To manage it, in [16] we defined semantic versions of the arithmetic operators needed for uncorrelated noise addition (i.e. distance, mean and variance). Because they are also used for correlated noise addition, we briefly introduce them here.

The most basic operator is the one that measures the *distance* between two values. In the semantic domain, in which nominal values should be compared according to the semantics of the concepts to which they refer, the dissimilarity between the meaning of two concepts c_i and c_j can be computed through their semantic distance, $sd(c_i,c_j)$. In [16] the distance version of the well-known Wu and Palmer (W&P) semantic similarity measure [20] was selected to compute $sd(c_i,c_j)$ because it provides values normalized in the range [0..1], which is coherent with a Normal noise distribution, it provides a non-logarithmic and non-exponential assessment of the distance, which avoids the concentration of values in the high or low zones of the output range, and it is computationally efficient.

By relying on the semantic distance $sd(\cdot,\cdot)$, we can define the *semantic mean of a nominal attribute A* as the concept c form $\tau(D(A))$ that minimizes the sum of the semantic distances respect to all a_i in A.

$$s\mu(A) = \arg\min_{c \in \tau(D(A))} \left(\sum_{a_i \in A} sd(c, a_i) \right) \tag{7}$$

On the other hand, the *semantic variance of a nominal attribute A* can be defined as the average of squared semantic distances between each concept a_i in A and the semantic mean $s\mu(A)$.

$$s\sigma^2(A) = \frac{\sum_{a_i \in A} \left(sd(a_i, s\mu(A)) \right)^2}{n} \tag{8}$$

As stated in section 2, for correlated noise addition, we also require two additional operators measuring the linear dependence between attribute pairs:

covariance and *correlation*. In the numerical domain, the standard arithmetical covariance is positive when the greater values of one attribute mainly correspond with the greater values of the other attribute and the same holds for the lesser values. To distinguish great values from small values a total order over the attribute domains must exist, as it is the case of numbers. However, the domains of nominal attributes are usually non-ordinal and, thus, we need a measure of statistical dependence that does not rely on total orders.

In [21], Székely defined the *distance covariance* and the *distance correlation*; they measure the statistical dependence between variables by relying on the pairwise distance (or dissimilarity) between the values of each variable. In this paper, we adapt these measures to the nominal domain by relying on the above mentioned semantic distance $sd(\cdot,\cdot)$.

Let $A=\{a_1,\ldots, a_n\}$ and $B=\{b_1,\ldots, b_n\}$ be two nominal attributes of a sample of n individuals, where a_i and b_i represent the values for individual i. According to the definition of *distance covariance*, we must first compute the n by n distance matrices of each attribute. To do so, we must compute all pairwise distances of each attribute that, in the semantic domain, correspond to the semantic distances *sd* between nominal value pairs. Distance matrices are used to obtain the n by n double centered distance matrices that, for attribute A, is computed as follows:

$$\left(\delta_{ij}^{A}\right)_{i,j=1}^{n} = \left(sd(a_i,a_j) - \delta_{i.}^{A} - \delta_{.j}^{A} + \delta_{..}^{A}\right)_{i,j=1}^{n}, \tag{9}$$

where $sd(a_i,a_j)$ is the value of the $(i,j)^{\text{th}}$ element of the distance matrix of attribute A, $\delta_{i.}^{A}$ is the mean of i^{th} row of the distance matrix, $\delta_{.j}^{A}$ is the mean of j^{th} column of the distance matrix and $\delta_{..}^{A}$ is the mean of all values of the distance matrix (i.e. the grand mean). The notation is similar for attribute B.

Then, the *semantic distance covariance* between two nominal attributes A and B is the square root of the arithmetic mean of the products $\delta_{ij}^{A}\delta_{ij}^{B}$. It satisfies $sdCov(A,B) \geq 0$ and it is 0 if and only if A and B are independent [21].

$$sdCov(A,B) = \frac{1}{n}\sqrt{\sum_{i,j=1}^{n}\delta_{ij}^{A}\delta_{ij}^{B}} \tag{10}$$

As for the numerical correlation, the *semantic distance correlation* of two nominal attributes A and B can be computed as the nonnegative number obtained by dividing their distance covariance, $sdCov(A,B)$, by the product of their distance standard deviations.

$$sdCor(A,B) = \begin{cases} \dfrac{sdCov(A,B)}{\sqrt{sdVar(A)\times sdVar(B)}}, & sdVar(A)\times sdVar(B) > 0 \\ 0, & sdVar(A)\times sdVar(B) = 0 \end{cases}, \tag{11}$$

where $sdVar(A)$ and $sdVar(B)$ are the *semantic distance variances* of attributes A and B, which are a special case of distance covariance when the two attributes are identical (i.e., $sdVar(A) = sdCov(A,A)$). The *semantic distance correlation* results are bounded in the [0..1] range, and it is 0 if and only if A and B are independent. As in the numerical case, values close to zero suggest a weak semantic association between attributes, while larger values suggest a stronger association.

3.2 Semantic Noise for Multivariate Datasets

In this section we adapt the standard correlated noise addition method to the semantic domain of nominal data. The goal of our method is twofold: i) to mask original nominal values by replacing them with terms within a semantic distance coherent with the desired level of protection, and ii) to preserve, as much as possible the semantic features of the data. Regarding the latter, we specifically aim at preserving, as much as possible, the semantic mean of each attribute; obtaining a per-attribute dispersion proportional to the variance of the original data and the noise magnitude; obtaining a pairwise attribute dispersion proportional to the covariance the original data and the noise magnitude; and preserving, as much as possible, the semantic correlation between the attributes.

Unlike uncorrelated noise, the sequences of correlated noise must consider the degree of correlation between attributes in order to preserve their level of association (see Section 2). For simplicity of the following explanation, let us assume that the dataset X has only two nominal attributes A and B with n records (for datasets with more than two attributes, the process depicted below would be applied for attribute pairs). By relying on the operators detailed above, the generated *semantic correlated noise* consists on a (n×2) matrix of random numbers $\varepsilon_{A,B} = \{(\epsilon_{a1}, \epsilon_{b1}), \ldots, (\epsilon_{ai}, \epsilon_{bi}), \ldots, (\epsilon_{an}, \epsilon_{bn})\}$ that follows a multivariate normal distribution $\varepsilon_{A,B} \sim N(0, \alpha\Sigma_{A,B})$, with mean the vector 0 and covariance the matrix $\alpha\Sigma_{A,B}$, being α the desired level of semantic noise.

$$\Sigma_{A,B} = \begin{pmatrix} sdVar(A) & sdCov(A,B) \\ sdCov(B,A) & sdVar(B) \end{pmatrix} \tag{12}$$

In the numerical domain, the noise magnitude can be directly and coherently added/subtracted to/from the original values. Specifically, if a positive noise is added to an original value greater (lower) than any other value in the attribute sample, the masked value will get away from (closer to) the other values in the sample in the same magnitude; on the contrary, if the noise is negative, the masked value will get closer to (away from) the other values. Since the error is normally distributed around zero, the total added/subtracted magnitudes are compensated so that the aggregated relative differences are

maintained in the masked outcome, thus preserving the statistical features of the data like the mean.

In the semantic domain, we propose replacing original nominal values by other concepts in the underlying taxonomy $\tau(D(\cdot))$ whose semantic distance is, ideally, equal to the noise magnitude. However, since nominal data are not ordinal, if we replace an original value by another concept at a certain distance, we cannot guarantee that the new concept is also closer to or farther from the other values at the same distance.

To solve this issue we propose a heuristic that guides the replacement of values according to the noise sign and that aims at improving the preservation of the correlation between attributes. In general, to better preserve the correlation between the two attributes A and B, if a positive noise is added to an original value a_i greater (lower) than its pair b_i, the new value should get away from (closer to) b_i at the same magnitude; on the contrary, if the error is negative, the new value should get closer to (away from) b_i. Thus, the magnitude of the accumulated additions and subtractions between the pair will compensate each other.

In order to balance the number of "movements" between the attribute value pairs and preserve the attribute correlation, the heuristic we propose interprets the noise sign with respect to a reference point defined by the pair corresponding to the value that is being replaced: if the noise is positive, the concept c in $\tau(D(A))$ that will replace a_i must be farther from its pair b_i than a_i, that is, $sd(c, b_i) > sd(a_i, b_i)$, and vice versa; if the noise is negative, the concept c must be closer to b_i than a_i, that is, $sd(c, b_i) < sd(a_i, b_i)$, and vice versa. Understandably, both attributes must belong to the same semantic domain, that is, $\tau(D(A)) = \tau(D(B))$.

Algorithm 1 formalizes the semantic correlated noise addition process according to the proposed heuristic. First, the taxonomies $\tau(D(A))$ and $\tau(D(B))$ associated to the domains of attributes A and B are obtained as detailed in Section 3.1. Then, the values of each attribute are mapped to concepts of $\tau(D(A))$ and $\tau(D(B))$. In lines 4 and 5, the (2×2) covariance matrix $\Sigma_{A,B}$ is built and the (n×2) noise matrix $\varepsilon_{A,B} \sim N(0, \alpha\Sigma_{A,B})$ is generated. Finally, in lines 6 to 14, attribute A is masked according to the desired magnitude of correlated noise and by following the proposed heuristic. The same process is applied for attribute B.

It is important to note that, to provide a priori privacy guarantees, our method should be applied in the context of a noise-based privacy model, such as ε-differential privacy [10]. Through a specially tailored noise distribution, this model guarantees that the outputs are insensitive (up to a factor dependent on ε) to modifications of one input record. In this way, the participation of one individual in a survey (or any specifics that he has contributed to the survey) will not be jeopardized by more than a $1+\varepsilon$ factor.

Algorithm 1. Semantic correlated noise addition for two attributes A and B.

Input : A, B : nominal attributes of n records; τ: taxonomy; α: semantic noise level
Output : A^*, B^* : noise-added masked nominal attributes

1: $\tau(D(A)) \leftarrow$ obtain_taxonomy$(D(A), \tau)$
2: $A \leftarrow$ map$(A, \tau(D(A)))$
3: Apply lines 1-2 to the attribute B
4: $\Sigma_{A,B} \leftarrow$ compute_CovarianceMatrix(A, B)
5: $\varepsilon_{A,B} \leftarrow$ generate_NoiseMatrix$(\alpha, \Sigma_{A,B})$ $\quad // \varepsilon_{A,B} = \left\{\left(\epsilon_{a_1}, \epsilon_{b_1}\right), \ldots, \left(\epsilon_{a_n}, \epsilon_{b_n}\right)\right\} \sim N(0, \alpha \Sigma_{A,B})$
6: **for all** a_i in A **do**
7: **if** $\epsilon_{a_i} = 0$ **then**
8: $a_i^* \leftarrow a_i$
9: **else if** ϵ_{a_i} is positive **then**
10: $a_i^* \leftarrow \underset{c \in \tau(D(A))}{\arg\min} \left\{ sd(c, a_i) \mid sd(c, a_i) \geq \left|\epsilon_{a_i}\right| \wedge sd(c, b_i) > sd(a_i, b_i) \right\}$
11: **else if** ϵ_{a_i} is negative **then**
12: $a_i^* \leftarrow \underset{c \in \tau(D(A))}{\arg\min} \left\{ sd(c, a_i) \mid sd(c, a_i) \geq \left|\epsilon_{a_i}\right| \wedge sd(c, b_i) < sd(a_i, b_i) \right\}$
13: **end if**
14: **end for**
15: Apply lines 6-14 to attribute B
16: **return** A^*, B^*

4 Experiments

In this section we evaluate the algorithm proposed in Section 3 with two nominal datasets with significantly different degrees of attribute correlation, and compare them with the semantic uncorrelated noise addition method we presented in [16]. As evaluation data, we use a patient discharge dataset provided by the California Office of Statewide Health Planning and Development, where each record describes a patient and contains two -correlated- nominal attributes: the principal diagnosis and the secondary diagnosis. SNOMED-CT [22], a healthcare knowledge base modeling all the diagnoses in the dataset, has been used as the ontology that provides the data semantics.

In the first experiment we have used a sample of 1,350 records with a high *semantic distance correlation* (0.94) between the two nominal attributes (i.e., *A=principal diagnosis* and *B=secondary diagnosis*). The dataset has been masked with the algorithm we propose in Section 3 and also with the uncorrelated noise addition method presented in [16], for the usual values of the noise parameter α=[0.1..0.5]. Figure 1 shows the semantic distance correlation (*sdCor*, equation (11)) of the noise-added attributes for the different methods and values of α. As expected, the results show that *sdCor* is preserved by the correlated noise addition algorithm better than by the uncorrelated method [16], a difference that is more noticeable for large values of α. Notice that the

semantic correlation (and, as we discuss below, any other feature of the data) cannot be perfectly preserved when adding noise to nominal data because of two reasons. First, in the semantic domain, value replacements are discrete and noise magnitudes rarely match to exact concept distances in the taxonomy τ (especially for coarse grained taxonomies); thus, some approximation errors will be accumulated during the noise-based replacement of values. Second, because the size of the taxonomy $\tau(D(A))$ is limited, there will be cases in which we cannot find a replacement concept that is as far as stated by the magnitude of the noise; in these cases, the farthest concept will be used as replacement, thus truncating the noise magnitude to the distance of that most distant concept.

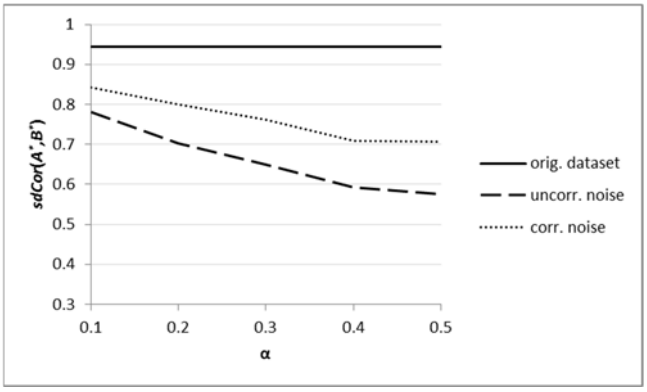

Fig. 1. Semantic distance correlation (*sdCor*) for a sample of 1,350 records with strongly correlated attributes (0.94).

In addition to the attribute correlation, noise addition methods should also preserve other features of the data. In [16] it was shown that the uncorrelated noise method was able to reasonably preserve the semantic mean ($s\mu$, equation (7)) of individual attributes and to maintain a data dispersion ($s\sigma^2$, equation (8)) proportional. The semantic correlated noise addition method we present here should also preserve these features, in addition to maintain a proportional semantic distance covariance (*sdCov*, equation 10).

For the evaluation dataset, the mean of attribute A is $s\mu(A)$=*Furuncle of chest wall* and of attribute B is $s\mu(B)$=*Viral hepatitis with hepatic coma* and the variances are $s\sigma^2(A)$= 0.22 and $s\sigma^2(B)$=0.24. Finally the covariance between A and B is $sdCov(A,B)$=0.2553. In Table 1 we depict how these features have been preserved in the noise-added attributes for the different methods and values of α. Specifically, we measured the semantic distance between the mean concept of each noise-added attribute and that of the original sample, and the absolute difference between the actual variances/covariance of the masked attributes and the expected variances/covariance after adding noise

with α={0.2, 0.4}. In all cases a value of 0 states that the means, variances or covariance have been perfectly controlled. In addition, we also measured the actual root mean square error (RMSE), computed as the root average square semantic distance between original and masked value pairs; this measures the overall loss of semantics in the masked dataset, which should be similar to the target RMSE associated to the noise sequences to be added for each α.

Table 1. Evaluation metrics for noise-added attributes with the semantic uncorrelated method from [16] and the semantic correlated method proposed here.

Metric	Semantic uncorrelated noise addition		Semantic correlated noise addition	
	α=0.2	α=0.4	α=0.2	α=0.4
$sd(s\mu(A), s\mu(A^*))$	0.20	0.20	0	0.20
$\lvert s\sigma^2(A^*) - (1+\alpha)\, s\sigma^2(A)\rvert$	0.02	0.07	0.01	0.04
$Actual\_RMSE(A)$	0.28	0.36	0.30	0.39
$Target\_RMSE(A)$	0.21	0.31	0.23	0.34
$sd(s\mu(B), s\mu(B^*))$	0	0	0.18	0.18
$\lvert s\sigma^2(B^*) - (1+\alpha)\, s\sigma^2(B)\rvert$	0.05	0.09	0.05	0.08
$Actual\_RMSE(B)$	0.24	0.34	0.28	0.37
$Target\_RMSE(B)$	0.21	0.31	0.24	0.34
$\lvert sdCov(A^*,B^*) - (1+\alpha)\, sdCov(A,B)\rvert$	0.18	0.26	0.17	0.25

Because α determines the magnitude of the added noise, the larger the α, the larger the RMSE (and the stricter the masking). Table 1 shows that the actual RMSE is similar to the target RMSE (i.e., the methods are able to controllably perturb data), even though not identical. As discussed above, the differences between target and actual RMSEs are due to the need to discretize and truncate noise magnitudes to the concepts in the taxonomy. The mean of the masked attributes is preserved up to a similar degree for the two methods, and we cannot observe significant differences between values of α. Note that, although the correlated noise addition algorithm is more focused in preserving the correlation, it is still capable of preserving the mean up to a similar or even better degree than the uncorrelated method. Table 1 also shows that the absolute difference between the variances and covariance of the masked sample and the expected ones are nearly 0 for the two methods. This means that the correlated noise addition algorithm, in addition to better preserve correlations, also maintain dispersions as controlled as the uncorrelated method.

In a second experiment we extracted a sample of 1,049 records with weakly correlated diagnoses ($sdCor(A,B)$=0.24). As above, Figure 2 shows the *semantic distance correlation* for the different methods and values of α. In this case, the differences between methods are much smaller. This suggests that, for weakly correlated attributes, the correlated method may not be worth it.

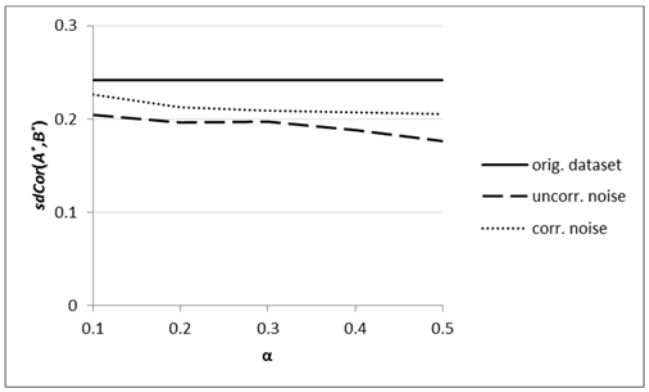

Fig. 2. Semantic distance correlation (*sdCor*) for a sample of 1,049 records with weakly correlated attributes (0.24).

5 Conclusions

In this paper, we have presented a semantically-grounded version of the numerical correlated noise addition method, which is capable of masking multivariate nominal datasets while reasonably preserving their semantics and attribute correlations. Unlike other perturbative data protection methods based on clustering records, our proposal is able to protect records individually, which is useful when protecting dynamic data or data streams, or when we need to accommodate per-individual privacy needs. Moreover, unlike the non-perturbative methods usually employed to mask nominal values (i.e., based on generalizations), our approach does not coarsen the granularity of the data. Finally, our proposal to noise addition is general and can support the usual noise distributions (e.g., Normal, Laplace) employed in noise-based data protection schemas and privacy models (such as ε-differential privacy).

As future work, we plan define other heuristics to assist the value replacement process so that we are able to preserve better a particular feature of the data, in case the posterior data analysis strongly depends on that. Moreover, we also plan to compare our approach with non-semantic ones, such as the discrete but distributional-oriented geometric and exponential mechanisms employed by ε-differential privacy for categorical data.

Acknowledgements. This work was supported by the EU Commission under the H2020 project "CLARUS", by the Spanish Government through projects TIN2014-57364-C2-R "SmartGlacis", TIN2011-27076-C03-01 "Co-Privacy" and TIN2015-70054-REDC "Red de excelencia Consolider ARES" and by the Government of Catalonia under grant 2014 SGR 537. M. Batet is supported by a Postdoctoral grant from Ministry of Economy and Competitiveness (MINECO) (FPDI-2013-16589).

References

1. Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Nordholt, E.S., Spicer, K., Wolf, P.-P.: Microdata. In: Wiley (ed.) Statistical Disclosure Control, pp. 23-130 (2012)
2. Domingo-Ferrer, J., Sánchez, D., Soria-Comas, J.: Database Anonymization: Privacy Models, Data Utility, and Microaggregation-based Inter-model Connections. Morgan & Claypool Publishers (2016)
3. Ramirez, E., Brill, J., Ohlhausen, M., Wright, J., Mc-Sweeny, T.: Data brokers: A call for transparency and accountability. Federal Trade Commission, Tech. Rep. (May 2014)
4. Sánchez, D., Batet, M.: C-sanitized: A privacy model for document redaction and sanitization. Journal of the Association for Information Science and Technology 67, 148-163 (2016)
5. Soria-Comas, J., Domingo-Ferrer, J., Sánchez, D., Martínez, S.: t-Closeness through Microaggregation: Strict Privacy with Enhanced Utility Preservation. IEEE Transactions on Knowledge and Data Engineering 27, 3098-3110 (2015)
6. Martínez, S., Sánchez, D., Valls, A.: Semantic adaptive microaggregation of categorical microdata. Computers & Security 31, 653-672 (2012)
7. Batet, M., Erola, A., Sánchez, D., Castellà-Roca, J.: Utility preserving query log anonymization via semantic microaggregation. Information Sciences 242, 49-63 (2013)
8. Samarati, P., Sweeney, L.: Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression. Computer Science Laboratory, SRI International (1998)
9. Krempl, G., Zliobaite, I., Brzezinski, D., Hüllermeier, E., Last, M., Lemaire, V., Noack, T., Shaker, A., Sievi, S., Spiliopoulou, M., Stefanowski, J.: Open Challenges for Data Stream Mining Research. ACM SIGKDD Explorations Newsletter 16, 1-10 (2014)
10. Dwork, C.: Differential privacy. Automata, Languages and Programming 4052, 1-2 (2006)
11. Kooiman, P., Willenborg, L., Gouweleeuw, J.: Pram: A method for disclosure limitation of microdata. Research Paper 9705, Statistics Netherlands, P.O. Box 4000, 2270 JM Voorburg, The Netherlands (1997)
12. Giggins, H., Brankovic, L.: Protecting privacy in genetic databases. Proceeding of the 6th Engineering Mathematics and Applications Conference (EMAC 2003), vol. 2, pp. 73-78, Sydney, Australia (2003)
13. Ghosh, A., Roughgarden, T., Sundararajan, M.: Universally utility-maximizing privacy mechanisms. Proceeding of the ACM Symposium on Theory of Computing (STOC'09), pp. 351-360 (2009)
14. McSherry, F., Talwar, K.: Mechanism design via differential privacy. Proceeding of Annual IEEE Symposium on Foundations of Computer Science (FOCS'07), pp. 94-103 (2007)
15. Abril, D., Navarro-Arribas, G., Torra, V.: On the declassification of confidential documents. Modeling Decision for Artificial Intelligence 6820, 235-246 (2011)
16. Rodriguez-Garcia, M., Batet, M., Sánchez, D.: Semantic Noise: Privacy-Protection of Nominal Microdata through Uncorrelated Noise Addition. Proceeding of the 27th IEEE International Conference on Tools with Artificial Intelligence, ICTAI 2015, pp. 1106-1113, Vietri sul Mare, Italy (2015)
17. Conway, R., Strip, D.: Selective partial access to a database. Cornell University, Tech. Rep. (1976)
18. Tendick, P.: Optimal noise addition for preserving confidentiality in multivariate data. Journal of Statistical Planning and Inference 27, 341-353 (1991)
19. Kim, J.: A method for limiting disclosure in microdata based on random noise and transformation. Proceeding of the ASA Section on Survey Research Methods, pp. 370-374 (1986)
20. Wu, Z., Palmer, M.: Verbs semantics and lexical selection. Proceeding of the Annual Meeting of the Association for Computational Linguistics, pp. 133-139 (1994)

21. Székely, G.J., Rizzo, M.L., Bakirov, N.K.: Measuring and testing dependence by correlation of distances. Annals of Statistics 35, 2769-2794 (2007)
22. Spackman, K.A.: SNOMED CT milestones: endorsements are added to already-impressive standards credentials. Healthcare informatics: the business magazine for information and communication systems 21, 54-56 (2004)