Published version:
doi.org/10.1016/j.ins.2013.04.021

Anonymization of Nominal Data Based on Semantic Marginality

Josep Domingo-Ferrer, David Sánchez[1], Guillem Rufian-Torrell

UNESCO Chair in Data Privacy, Department of Computer Science and Mathematics
Universitat Rovira i Virgili, Av. Països Catalans 26, E-43007 Tarragona, Catalonia

Abstract

Nominal attributes are very common in data sets about individuals, specifically medical data like patient healthcare records. Attributes of this type tend to be sensitive due to their personal nature. If public-use data sets need to be released, *e.g.* for clinical research purposes, data should be first anonymized. However, since most anonymization methods omit data semantics when dealing with nominal attributes (*e.g.* in a medical data set diagnosis is a nominal attribute), anonymization results in unnecessary information loss for such attributes, which is especially serious given their analytical importance. In this paper, we present a knowledge-based numerical mapping for nominal attributes that captures and quantifies their underlying semantics. Using this mapping, we show how to compute semantically and mathematically coherent mean, variance and covariance functions for nominal attributes; we also propose a distance measure between records containing numerical and nominal attributes. Thus, the proposed mapping allows adapting to nominal data some Statistical Disclosure Control anonymization methods originally designed for numerical attributes. Evaluation results obtained for one of these methods applied to real patient discharge data shows that the use of our mapping retains better the semantics of original data and, hence, it yields anonymized data with better utility for clinical research.

Keywords: Anonymization, Data privacy, Data semantics, Medical Ontologies, Statistical disclosure control

1. Introduction

The analysis of patient data is of utmost importance in modern medicine since such data capture the healthcare experience, which is the base to improve

[1]Corresponding author. Address: Departament d'Enginyeria Informàtica i Matemàtiques. Universitat Rovira i Virgili. Avda. Paisos Catalans, 26. 43007. Tarragona. Catalonia. Tel.:+034 977 559657; Fax: +034 977 559710. E-mail: david.sanchez@urv.cat

future patient assessments and treatments [14]. Hence, the publication of these data for secondary use is crucial for medical research.

Patient data are at the same time highly sensitive by definition, because they may expose the patient's identity together with confidential outcomes (*e.g.* diagnosis). The disclosure of this information in published data sets may cause a serious damage to both individual patients and medical professionals who are responsible for the patient's privacy. In fact, legal regulations, like the UK Data Protection Act (DPA, 1998), consider patient data as "sensitive personal data" and state that they cannot be released without the patient's consent. Exceptions are made for medical research, allowing *non-identifiable* patient data to be released (DPA, Section 39).

To produce non-identifiable patient data sets, the U.S. Health Insurance Portability and Accountability Act (HIPAA) defines 18 identifying data elements, including names, geography and dates, which must be removed or coarsened prior to publication. However, several studies [14, 26, 22] have raised awareness that disclosure may still happen even when identifier attributes have been removed. Given the amount of gathered patient data, identities can still be revealed through statistical inference applied to published data. Scarce (or even unique) combinations of attribute values may enable re-identifying individuals. A well-known example of patient disclosure from published tabular data was the re-identification of a late abortion case thanks to the low amount of late abortions occurred in the area in which both the patient and the responsible clinician were located [14]. For published individual data (microdata) re-identification may even be easier, because the level of detail is greater.

To enable the publication of truly privacy-preserving but still analytically useful data, statistical disclosure control (SDC, [19, 18, 12, 36]), a.k.a. data anonymization or privacy-preserving data mining, can be used. SDC methods aim at making possible the publication of statistical data in such a way that individual responses of specific users cannot be inferred from the published data and background knowledge available to intruders. Since published data should still be useful for medical research (which is the main motivation for releasing data), SDC methods are intended to optimize the trade-off between disclosure risk and information loss resulting from the anonymization process. This is a major difference between SDC and data encryption or differential privacy [13], which are approaches targeted only at minimizing disclosure. To achieve their goal, SDC methods mask original data (via perturbation or detail reduction) or generate synthetic data preserving statistical features of original data.

1.1. Background on SDC methods

Most SDC methods have been designed to deal with numerical data. Numbers are easy to treat because arithmetical functions can be applied to them to perform the comparison and transformation operations required for data anonymization. However, sensitive nominal values (such as diagnoses, procedures, treatments, etc.), which take values in a finite set of categories and for which arithmetical operations do not make sense, are very common and of the utmost importance in the medical domain.

Applying existing data anonymization methods to nominal attributes is not straightforward. Following [19], we classify anonymization methods for microdata into perturbative —methods which distort original data— and non-perturbative —methods which, instead of perturbing data, rely on reducing their detail or partially suppressing them—. *Perturbative anonymization methods* applicable to nominal data suffer from a common shortcoming: they perturb categorical values without taking the hierarchy of categories into account, that is, while ignoring the semantics of the categories. This causes utility loss, but it may also bring scalability problems. Specifically:

- Several perturbative methods replace each nominal attribute by as many binary 0-1 attributes as the number of possible attribute categories; such is the case of multiply-imputed synthetic data [30] and data shuffling [25]. This approach soon yields unmanageable data sets (*e.g.* diseases, as modeled in the ICD-9 taxonomy [20] can take over 12,000 different categories).

- PRAM [15, 18] is an anonymization technique designed for nominal attributes. It certainly does not need binary attributes, but it requires as a control parameter a Markov transition matrix, whose size grows quadratically in the number of nominal categories.

- Microaggregation is a family of perturbative SDC methods originally defined for numerical data [6, 9]. First, original records are partitioned into groups in such a way that records in the same group are *similar* to each other and so that the number of records in each group is at least k. Then, an aggregation operator (typically the group centroid/mean) is computed for each group and is used to replace original records. As a result, each masked record becomes indistinguishable from, at least, $k-1$ other records, thereby achieving k-anonymity [31, 34]. In [35] and [11], extensions of microaggregation for categorical attributes were proposed: the former paper addressed only categorical ordinal attributes and proposed the median as an aggregation operator; the latter paper also considered nominal attributes using the equality/inequality predicate and proposed the modal value as an aggregation operator for them. However, the modal value is a very coarse aggregation operator which may not even be uniquely defined, especially over a small group of values. We see that, while microaggregation is scalable, the way it is currently applied to nominal attributes causes substantial utility loss.

Thus, the above-mentioned perturbative methods incur a high complexity for anonymizing nominal data or they are coarse and cause substantial information loss. This is because they treat nominal data as flat categorical values, for which the only possible operator is binary comparison for equality [11]. This simplistic approach omits data semantics. Overlooking semantics decreases the utility of the anonymized data set since it fails to preserve the meaning of the original data. Semantically-grounded analyses would be desirable to better preserve the data utility.

On the other hand, *non-perturbative anonymization methods* usually do take semantics into account, as they rely on category generalization [17, 21, 31, 34, 5]. Categories are words or noun phrases referring to concepts (*e.g.* disease names) which capture their semantics, and semantics is a human-inherited feature. Hence, semantic analysis requires a human-tailored knowledge base that captures and structures the conceptualization of nominal attributes. For this purpose, structured thesauri, taxonomies or ontologies [16] can be used. Due to the importance of knowledge and terminology in clinical assessment, the medical community has been especially fertile at producing standard structured vocabularies that systematically model all known medical terms (*i.e.* diseases, symptoms, procedures, substances, devices, etc.), such as ICD-9 [20], MeSH [24] or SNOMED-CT [33]. The anonymization process consists of substituting original categories by more general ones obtained from a hierarchical structure derived from a knowledge base. This process reduces the number of distinct tuples in the data set and, therefore, increases the level of anonymity. The substitution is selected according to a metric that measures the information loss caused by each substitution compared to the original data. Since the utility of anonymized data depends on the suitability of available generalizations, these methods define *ad hoc* hierarchical structures (named Value Generalization Hierarchies (VGHs)) that are best suited for input data. If input categories change, VGHs must be modified accordingly. Moreover, VGHs usually offer a rough and overspecified knowledge model compared to fine-grained and general taxonomies/ontologies [23]. Even though these methods consider the semantics of input data, they usually cause a high information loss, because categories are replaced by *more general* versions that only retain data semantics in a partial way. This issue is especially evident for heterogeneous data, in which the need to generalize outliers results in coarser categories [23]. Hence, generalization usually results in a significant loss of granularity. Furthermore, for numerical attributes, generalization discretizes input numbers to numerical ranges and thereby changes the nature of data from continuous to discrete.

In summary, while perturbative methods ignore the semantics of nominal attributes, non-perturbative methods cause substantial granularity loss. Clearly, there is room for improvement in the way nominal attributes are treated by anonymization methods.

1.2. Contribution and plan of this article

The work in this paper is motivated by the following observations: i) most SDC methods omit data semantics during the anonymization of nominal data (which reduces utility due to the lack of a semantically-coherent anonymization) or are based on strict value generalizations (which usually cause a high information loss); ii) in the medical domain, nominal data are important for clinical research and, at the same time, sensitive by nature; iii) standard medical knowledge bases are available which offer structured conceptualizations of all medical nominal categories. In view of the above, in this paper we present a knowledge-based numerical mapping for nominal attributes that captures and quantifies their underlying semantics. By means of this mapping, we show that

it is possible to compute semantically and mathematically coherent *mean*, *variance* and *covariance* functions for nominal data. Based on these functions, we also propose a *distance measure* to manage and compare records containing numerical and nominal attributes. The proposed functions and measures can be directly plugged into numerically-oriented SDC methods, in order to transparently capture the semantic and statistical features of nominal data during their anonymization.

To test the benefits that the proposed mapping brings to existing SDC methods, we show how a method originally designed for numerical data can be easily adapted to perform a semantically-grounded anonymization of nominal data. Evaluation results obtained for *real* patient discharge data show that the use of our mapping retains better the semantics of original data and, hence, it yields anonymized data with better analytical utility.

Section 2 introduces the knowledge-based numerical mapping for nominal data. Section 3 shows how to plug the proposed mapping into a microaggregation SDC method and presents the evaluation results. Section 4 lists conclusions and future research issues.

2. Methods

In this section, we describe the proposed knowledge-based numerical mapping for nominal attributes.

Our objective is to associate to each nominal value a number (named *marginality*) that captures both its *semantic* and *distributional* features. Marginality can be understood as a measure of value centrality within a background ontology/taxonomy [29], that is, it attempts to determine the "middle" of the hierarchy and how far each nominal value lies from that middle. A distinctive feature of marginality is that it takes into account the position of the category the value comes from within the taxonomy, and also the frequency of the value in the data set: a value belonging to an extreme category becomes more central as its frequency increases, just like a distant district of a city would become central if most of the city's population moved to that district.

Other centrality measures used in anonymization are either solely based on sample frequency (*e.g.* by taking the mode as the central value [11]) or assume that the most concrete concept generalizing all sample values in a taxonomy [1] is the appropriate center. Since the former approach omits data semantics and the latter neglects the sample distribution, neither of them captures both dimensions of data as marginality does.

In our proposal, the semantics of nominal values is captured using a *semantic distance measure* that, based on the structure of the background ontology/taxonomy, quantifies the taxonomic resemblance between value pairs. In the next section, we present the measure and discuss its advantages from the semantic and mathematical perspectives.

2.1. Ontology-based semantic distance

The notion of semantic similarity/distance has been extensively studied and used in the past to quantify and compare the semantics of nominal data. Different approaches can be identified according to the techniques and knowledge bases used to perform the assessment [32]. Since our proposal is based on structured knowledge bases like taxonomies/ontologies, in the following we focus on ontology-based measures.

To exploit the semantics modeled in ontologies, these can be viewed as directed graphs in which taxonomic relations are represented as links between nodes that correspond to concepts. A straightforward way to estimate the semantic distance between concept pairs is to count the number of edges separating them. Several measures have been developed based on this principle [28, 37]. Even though edge-counting methods are easily applicable, they have been surpassed by other methods that exploit ontological knowledge more exhaustively. In [32], a state-of-the-art ontology-based measure is proposed that measures the distance $d(c_1, c_2)$ between two concepts c_1 and c_2 as a function of their number of non-common taxonomic ancestors divided (for normalization) by their total number of ancestors

$$d(c_1, c_2) = \log_2 \left(1 + \frac{|T(c_1) \cup T(c_2)| - |T(c_1) \cap T(c_2)|}{|T(c_1) \cup T(c_2)|} \right) \tag{1}$$

where $T(c_i)$ is the set of taxonomic ancestors of concept c_i in the background ontology, including itself.

From a semantic perspective, the distance specified by Expression (1) captures more taxonomic knowledge than edge-counting methods, since it implicitly considers all possible paths connecting two concepts (several paths may exist in case of ontologies with multiple inheritance, which are very common in the medical domain [3]). Thanks to the normalizing denominator, the above distance can differentiate concept pairs with the same amount of shared ancestors. As a result, it approximates human judgments of similarity better than other ontology-based measures, as demonstrated for several standard benchmarks including general terms [32] and biomedical ones [3].

In contrast with absolute distance values returned by edge-counting methods, Expression (1) yields positive normalized values in the $[0, 1]$ range, which are desirable to coherently compare distances obtained from different ontologies. As a result, this distance can be applied to scenarios in which concept pairs are spread in different ontologies [4]. Moreover, as demonstrated in [32] and [2], Expression (1) satisfies *non-negativity*, *reflexivity*, *symmetry* and *subadditivity*, thereby being a distance measure in the mathematical sense.

2.2. A marginality measure for nominal attributes

Consider a nominal attribute X whose values are modeled in an ontology/taxonomy. Let T_X be a sample of values of X. The *marginality* $m(\cdot)$ of each value x_j in T_X is computed as

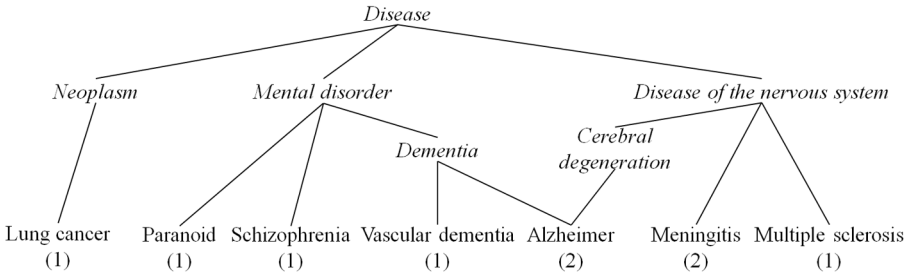

Figure 1: Example taxonomy of a sample of a *"Diagnosis"* attribute.

$$m(x_j) = \sum_{x_l \in T_X - \{x_j\}} d(x_j, x_l) \tag{2}$$

where $d(\cdot, \cdot)$ is the distance in Expression (1). Other distances could also be used [2], but the suitability of our choice is justified in Section 2.1. The greater $m(x_j)$, the more marginal (*i.e.* the less central) is x_j. Since $m(x_j)$ accumulates the semantic distances $d(x_j, x_l)$ from a given x_j to each other x_l in T_X, it captures both the "semantic centrality" of x_j in the taxonomy (*i.e.* the graph centrality of the concept representing x_j in the taxonomy) and also the sample frequencies (*i.e.* if the frequency of a value in T_X increases, the marginality of that value decreases since it becomes more "central").

Example 1. Assume a nominal attribute *"Diagnosis"*, for which a sample is available whose values can be taxonomically classified as shown in Figure 1. The sample has one element for each diagnosis category, except for "Alzheimer" and "Meningitis", for each of which there are two elements.

Label the elements in the sample as follows: x_1 (lung cancer), x_2 (paranoia), x_3 (schizophrenia), x_4 (vascular dementia), x_5 (first Alzheimer element), x_6 (second Alzheimer element), x_7 (first meningitis element), x_8 (second meningitis element) and x_9 (multiple sclerosis). The distance matrix between elements is given below, where component (j, l) represents the semantic distance $d(x_j, x_l)$

[2] A preliminary version of marginality based on an edge-counting distance was presented in the conference paper [7].

Table 1: Marginalities of elements in the "Diagnosis" sample of Figure 1

x_j	$m(x_j)$
x_1	$0 + 0.85 + 0.85 + 0.87 + 0.91 + 0.91 + 0.85 + 0.85 + 0.85 = 6.94$
x_2	$0.85 + 0 + 0.58 + 0.68 + 0.78 + 0.78 + 0.85 + 0.85 + 0.85 = 6.22$
x_3	$0.85 + 0.58 + 0 + 0.68 + 0.78 + 0.78 + 0.85 + 0.85 + 0.85 = 6.22$
x_4	$0.87 + 0.68 + 0.68 + 0 + 0.65 + 0.65 + 0.87 + 0.87 + 0.87 = 6.14$
x_5	$0.91 + 0.78 + 0.78 + 0.65 + 0 + 0 + 0.65 + 0.65 + 0.65 = 5.07$
x_6	$0.91 + 0.78 + 0.78 + 0.65 + 0 + 0 + 0.65 + 0.65 + 0.65 = 5.07$
x_7	$0.85 + 0.85 + 0.85 + 0.87 + 0.65 + 0.65 + 0 + 0 + 0.58 = 5.3$
x_8	$0.85 + 0.85 + 0.85 + 0.87 + 0.65 + 0.65 + 0 + 0 + 0.58 = 5.3$
x_9	$0.85 + 0.85 + 0.85 + 0.87 + 0.65 + 0.65 + 0.58 + 0.58 + 0 = 5.88$

as defined in Expression (1):

$$\begin{pmatrix} 0 & 0.85 & 0.85 & 0.87 & 0.91 & 0.91 & 0.85 & 0.85 & 0.85 \\ 0.85 & 0 & 0.58 & 0.68 & 0.78 & 0.78 & 0.85 & 0.85 & 0.85 \\ 0.85 & 0.58 & 0 & 0.68 & 0.78 & 0.78 & 0.85 & 0.85 & 0.85 \\ 0.87 & 0.68 & 0.68 & 0 & 0.65 & 0.65 & 0.87 & 0.87 & 0.87 \\ 0.91 & 0.78 & 0.78 & 0.65 & 0 & 0 & 0.65 & 0.65 & 0.65 \\ 0.91 & 0.78 & 0.78 & 0.65 & 0 & 0 & 0.65 & 0.65 & 0.65 \\ 0.85 & 0.85 & 0.85 & 0.87 & 0.65 & 0.65 & 0 & 0 & 0.58 \\ 0.85 & 0.85 & 0.85 & 0.87 & 0.65 & 0.65 & 0 & 0 & 0.58 \\ 0.85 & 0.85 & 0.85 & 0.87 & 0.65 & 0.65 & 0.58 & 0.58 & 0 \end{pmatrix}$$

The marginality $m(x_j)$ of element x_j can be obtained by adding all distances in the j-th row of the above matrix. Marginalities for all elements are shown in Table 1. It turns out that x_1 (lung cancer) is the most marginal element, which is consistent with the layout of the taxonomy in Figure 1, since it is the most outlying element. On the other hand, x_5 and x_6 (Alzheimer) are the least marginal elements, due to both their central position in the hierarchy (given that they belong to both the mental disorder and the disease of nervous system taxonomic branches) and the fact that there are two Alzheimer elements. This illustrates that marginality captures both the semantics modeled in the taxonomy and the distribution of the sample.

2.3. Statistical analysis of nominal data

In the previous section we have shown how a nominal value x_j can be associated a marginality measure $m(x_j)$. In this section, we show how this numerical measure can be used in statistical analyses and also to define an integrated distance measure between multi-attributed records of different types (numerical and nominal). This will enable anonymization methods to coherently compare and aggregate records with heterogeneous attribute types considering both the semantics and the distribution of nominal values.

2.3.1. Marginality-based approximate mean

The mean of a sample of nominal values cannot be computed in the standard sense, since it would be necessary to discretize the numerical mean to a nominal value. However, it can be reasonably approximated by the least marginal value, that is, by the sample centroid.

Definition 1 (Marginality-based approximate mean). *Given a sample T_X of a nominal attribute X, the marginality-based approximate mean is defined as*

$$Mean_M(T_X) = \arg\min_{x_j \in T_X} m(x_j) \tag{3}$$

if one wants the mean to be a nominal value, or

$$Num\_mean_M(T_X) = \min_{x_j \in T_X} m(x_j) \tag{4}$$

if one wants a numerical mean value.

Example 2. In Example 1, the nominal mean of the sample is Alzheimer since, as discussed above, it is the least marginal value. Consistently, the numerical mean is $m(Alzheimer) = 5.07$.

2.3.2. Marginality-based variance and covariance

The intuitive idea behind variance in a taxonomy is that a sample of nominal values belonging to categories which are all children of the same parent category has smaller variance than a sample with children from different parent categories. The average marginality of a sample turns out to capture this notion of variance.

Definition 2 (Marginality-based variance). *Given a sample T_X of n values drawn from a nominal attribute X, the marginality-based sample variance is defined as*

$$Var_M(T_X) = \frac{\sum_{x_j \in T_X} m(x_j)}{n} \tag{5}$$

Example 3. It can be seen from Table 1 that, for the sample of Example 1, the marginality-based variance is

$$\frac{6.94 + 6.22 + 6.22 + 6.14 + 5.07 + 5.07 + 5.3 + 5.3 + 5.88}{9} = 5.79$$

Definition 3 (Marginality-based covariance). *Given a bivariate sample $T_{(X,Y)}$ consisting of n ordered pairs of values $\{(x_1, y_1), \cdots, (x_n, y_n)\}$ drawn from the ordered pair of nominal attributes (X, Y), the marginality-based sample covariance is defined as*

$$Covar_M(T_{(X,Y)}) = \frac{\sum_{j=1}^{n} \sqrt{m(x_j)m(y_j)}}{n} \tag{6}$$

The above definition yields a non-negative covariance whose value is higher when the marginalities of the values taken by X and Y are positively correlated: as the values taken by X become more marginal, so become the values taken by Y.

Given a multivariate data set T containing a sample of d nominal attributes $X^1, \cdots, X^d$, using Definitions 2 and 3 yields a covariance matrix $\mathbf{S} = \{s_{jl}\}$, for $1 \leq j \leq d$ and $1 \leq l \leq d$, where $s_{jj} = Var_M(T_j)$, $s_{jl} = Covar_M(T_{jl})$ for $j \neq l$, T_j is the column of values taken by X^j in T and $T_{jl} = (T_j, T_l)$.

2.3.3. Marginality-based distance between records

Based on variances (whether plain numerical or marginality-based), we can define the following distance between records having attributes of different types (numerical and nominal).

Definition 4 (S-distance). *The S-distance between two records $\mathbf{x}_1$ and $\mathbf{x}_2$ in a data set with d attributes is*

$$\delta(\mathbf{x}_1, \mathbf{x}_2) = \sqrt{\frac{(S^2)^1_{12}}{(S^2)^1} + \cdots + \frac{(S^2)^d_{12}}{(S^2)^d}} \tag{7}$$

where $(S^2)^l_{12}$ is the variance of the l-th attribute over the group formed by $\mathbf{x}_1$ and $\mathbf{x}_2$, and $(S^2)^l$ is the variance of the l-th attribute over the entire data set.

It can be seen from Expression (7) that, for records consisting only of numerical attributes, the proposed distance is a normalized Euclidean distance. However, expressing the distance in terms of variances has the advantage that one can extend it to nominal attributes using Definition 2. In order to add variances of different attributes, we need to eliminate the influence of the attribute scales (units). To this end, the variance of each attribute over the pair of records is normalized by dividing it by the variance of the attribute over the entire data set. Such an integrated distance allows convenient handling of heterogeneous multi-attribute records.

We prove in the Appendix the following two theorems stating that the distance above satisfies the properties of a mathematical distance.

Theorem 1. *The S-distance based on the marginality-based variance as per Definition 2 and computed on multivariate records consisting of nominal attributes is a distance in the mathematical sense.*

Theorem 2. *The S-distance based on the usual numerical variance and computed on multivariate records consisting of ordinal or numerical attributes is a distance in the mathematical sense.*

By combining the proofs of Theorems 1 and 2, the next corollary follows.

Corollary 1. *The S-distance on multivariate records consisting of attributes of any type, where the marginality-based variance is used for nominal attributes and the usual numerical variance is used for numerical and ordinal attributes, is a distance in the mathematical sense.*

The above distance can be used for a variety of purposes, including clustering and microaggregation. Specifically, it allows microaggregating heterogeneous data [9, 11] in view of anonymization.

3. Results and discussion

Marginality converts nominal data into numbers that can be conveniently treated by numerical anonymization techniques without disregarding data semantics. In this section, we describe the adaptation of an anonymization method based on microaggregation to incorporate the marginality model. Afterwards, we present an empirical evaluation on two real clinical data sets.

We choose microaggregation because, as justified in Section 1.1, this anonymization method does not present scalability problems when dealing with nominal attributes with large categorical domains, and it does not incur granularity loss. The only problem of microaggregation is that it does not consider semantics, but this can be remedied by combining it with marginality.

3.1. Marginality-based microaggregation

As introduced in Section 1.1, microaggregation methods for numerical data (like MDAV [11]) partition a set of records into groups with, at least, k elements each and with high within-group similarity, that is, with low within-group variance. The aim is to produce k-anonymous data sets [31, 34, 11]. Partitioning uses mean records as reference points. Using the marginality-based definitions of mean, variance and distance given earlier (Definitions 1, 2 and 4), we can directly partition nominal data while maximizing their *semantic* similarity.

To avoid the variance reduction caused by the aggregation step of microaggregation (which typically involves replacing records in each group of the partition by the group centroid/mean record), we replace the records of a group by synthetic data generated to preserve means and covariances of the original data. This was proposed in [8] for numerical data and we generalize it as follows for any kind of data.

Algorithm 1 (C_i: group of records obtained in the partition).
For each record $\mathbf{x}_j$ in C_i and for each attribute X_l to be synthesized:

1. *Pick a random value x'_{jl}* among those that can be taken *by X_l such that*

$$\delta(x_{jl}, x'_{jl}) \leq \delta_{max,i}(x_{jl})$$

   where $\delta(\cdot,\cdot)$ corresponds to the S-distance (Expression (7)), x_{jl} is the original value of the attribute in $\mathbf{x}_j$ and $\delta_{max,i}(x_{jl})$ is the maximum distance between x_{jl} and the values taken by attribute X_l over C_i;

2. *Replace x_{jl} by x'_{jl}.*

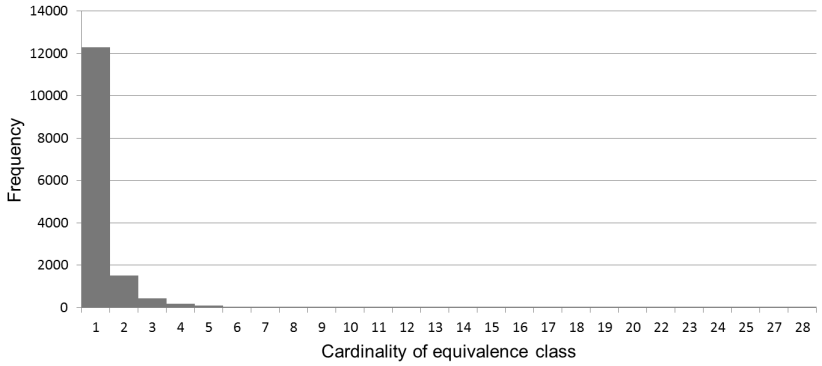

Figure 2: Distribution of the size of equivalence classes for DISCHARGE NEOPLASM data.

3.2. Evaluation data sets

To evaluate the improvements that the marginality model brings to data sets with nominal attributes, we applied the above-presented anonymization method to two real clinical data sets. We took the Patient Discharge Data for 2010 that can be obtained from California's Office of Statewide Health Planning and Development [27]. Within this, we selected the following numerical and nominal attributes for each patient: AGE_YRS (age in years), LOS (length of stay from admission to discharge in days), CHARGE (in dollars) and DIAG_P (principal diagnosis). DIAG_P is a nominal attribute coded according to the International Classification of Diseases (ICD-9-CM, [20]), which classifies diseases, symptoms, abnormal findings, complaints, social circumstances, and external causes of injury or disease. This standard classification, that models more than 12,000 concepts, is used by hospitals and healthcare centers worldwide to encode medical episodes and also to build electronic healthcare data sets that can be used for statistical analyses, medical research and decision support. From the entire data, we created two data sets by selecting the subsets of records for which DIAG_P was i) some form of neoplasm and ii) a kind of digestive disease. We deleted records with missing data and those for which CHARGE was $0 (a value 0 means that the charge for that discharge was unknown or invalid). For neoplasm-related data, we obtained 19,502 records; Figure 2 shows how many equivalence classes (groups of identical records) of each cardinality exist in the data set; for example, the first vertical bar of the figure states that 12,304 value tuples are unique (equivalence classes with cardinality 1 mean unique records), the second bar shows that 1,514 records are repeated twice, etc. We named the resulting data set DISCHARGE NEOPLASM. On the other hand, digestive-related data resulted in 143,472 records; Figure 3 shows how many equivalence classes of each cardinality exist in this data set, and in particular its first bar shows that it contains 142,897 unique records. We named this data set DISCHARGE DIGESTIVE. The large proportion of unique records in both data sets may facilitate identity disclosure to intruders.

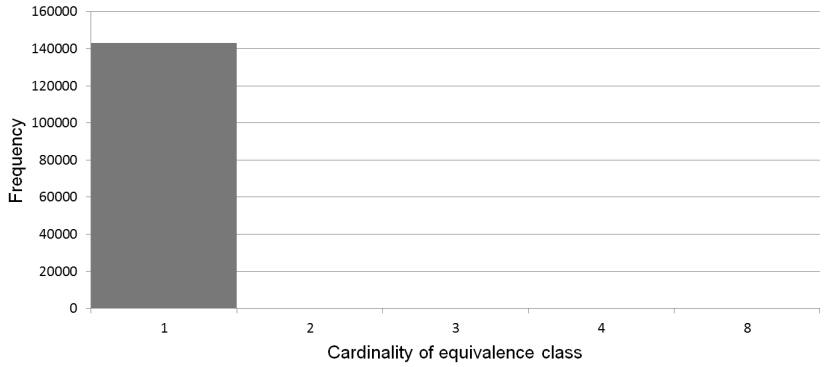

Figure 3: Distribution of the size of equivalence classes for DISCHARGE DIGESTIVE data.

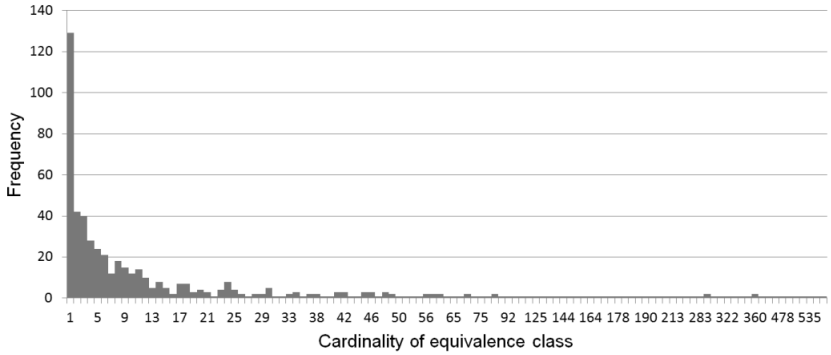

Figure 4: Distribution of the size of equivalence classes for DIAGNOSIS NEOPLASM data.

In addition to evaluating the marginality model with a combination of numerical and nominal attributes, we also tested its behavior when dealing solely with a nominal attribute, the DIAG_P attribute in the above data sets. We named the resulting single-attribute data sets DIAGNOSIS NEOPLASM and DIAGNOSIS DIGESTIVE, respectively. The distribution of the size of the equivalence classes of the DIAG_P attribute in both data sets (see Figures 4 and 5) shows a substantial number of unique values (129, that is, 0.66% of the total number of records) for DIAGNOSIS NEOPLASM; this gives an idea of the potential privacy risks of publishing diagnosis data "as is" [14]. Comparatively, the DIAGNOSIS DIGESTIVE data set includes only 37 unique values (a mere 0.025% of the total number of records), which represents a much lower risk of disclosure.

For both NEOPLASM data sets, the DIAG_P attribute included 542 different categories, which were covered by an ICD-9 taxonomic tree with 690 categories. For both DIGESTIVE data sets, the DIAG_P attribute included 409 different categories, which were covered by an ICD-9 taxonomic tree with 552 categories. These taxonomies configure the knowledge base on which the

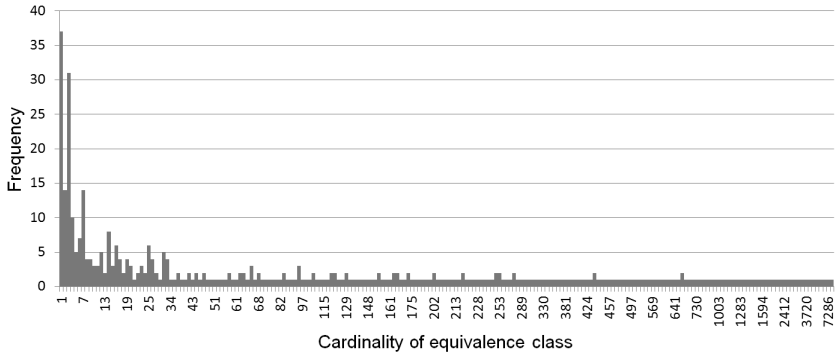

Figure 5: Distribution of the size of equivalence classes for DIAGNOSIS DIGESTIVE data.

marginality model relies.

3.3. Discussion on empirical results

We anonymized the above described data with two different versions of the algorithm introduced in Section 3.1: *marginality-based microaggregation* and *classic microaggregation*. In marginality-based microaggregation, records are compared and averaged according to the marginality-based definitions of variance, S-distance and mean; in turn, marginality for nominal values relies on the ontology-based semantic distance given in Expression (1), using the ICD-9 taxonomic tree as a knowledge base. In classic microaggregation [35, 11], the S-distance of Expression (7) is also used, but variances for nominal attributes are based on binary (rather than semantic) distances; that is, the distance between two nominal values is 0 if they are equal and 1 otherwise. Also, in classic microaggregation the average operator used for nominal values is the modal value (as in [11]) instead of the marginality-based mean. As a result, in this latter algorithm, nominal attributes are only evaluated according to their distribution of values, but not to their semantics. Note that, in both microaggregation versions, numerical attributes are treated in the same way by the S-distance expression, using standard arithmetic operators.

Comparing the information loss incurred by the two algorithms shows the benefits of the semantically-grounded marginality model. To measure the information loss caused by data microaggregation, the well-known Sum of Squared Errors (SSE) is usually employed in the literature [11, 9, 1, 10]. It is defined as the sum of squares of distances between original attribute values and their masked versions:

$$SSE(X_l) = \sum_{x_{jl} \in \tau(X_l)} (dist(x_{jl}, x'_{jl}))^2,$$

where $\tau(X_l)$ is the set of values taken by X_l in the data set, x_{jl} is the value of attribute X_l for the j-th record and x'_{jl} is its masked version; $dist(\cdot, \cdot)$ corresponds to the Euclidean distance if X_l is numerical and to the semantic distance

Table 2: SSE values obtained for each attribute of the DISCHARGE NEOPLASM data with marginality-based and classic microaggregation, respectively

	$SSE(AGE\_YRS)$		$SSE(LOS)$		$SSE(CHARGE)$		$SSE(DIAG\_P)$	
k	Classic	Marginal.	Classic	Marginal.	Classic	Marginal.	Classic	Marginal.
2	11430	11436	7664	7662	2.05E+12	2.05E+12	7589	7515
4	32740	32746	19154	19155	3.81E+12	3.81E+12	12121	10892
6	49978	49882	41234	41231	5.46E+12	5.46E+12	13722	11852
8	62009	61925	52172	52175	7.56E+12	7.56E+12	14517	12373
10	73136	73254	55748	55756	9.79E+12	9.79E+12	14989	12670
12	98734	98800	68512	68506	1.17E+13	1.17E+13	15308	12927
14	104671	104605	78060	78073	1.38E+13	1.38E+13	15550	13064
16	107409	107461	87869	87855	1.51E+13	1.52E+13	15757	13193
18	114864	114793	97136	97142	1.60E+13	1.60E+13	15896	13281
20	129845	129770	107016	106938	1.68E+13	1.68E+13	15981	13345

Table 3: SSE values obtained for each attribute of the DISCHARGE DIGESTIVE data with marginality-based and classic microaggregation, respectively

	$SSE(AGE\_YRS)$		$SSE(LOS)$		$SSE(CHARGE)$		$SSE(DIAG\_P)$	
k	Classic	Marginal.	Classic	Marginal.	Classic	Marginal.	Classic	Marginal.
2	34205	34205	20031	20031	4.51E+12	4.51E+12	51556	51556
4	96405	96405	64075	64075	3.05E+13	3.05E+13	88047	77478
6	144699	144699	109049	109049	4.52E+13	4.52E+13	101141	85827
8	192725	192725	159429	159429	5.73E+13	5.73E+13	107804	89896
10	237758	237758	187512	187512	7.04E+13	7.04E+13	111781	92616
12	257297	257297	203685	203685	7.67E+13	7.67E+13	114321	94442
14	294633	294633	217839	217839	8.64E+13	8.64E+13	116240	95873
16	322896	322896	248137	248137	9.43E+13	9.43E+13	117663	96894
18	339505	339505	327250	327250	9.44E+13	9.44E+13	118805	97642
20	377568	377568	358737	358737	9.93E+13	9.93E+13	119688	98321

given by Expression (1) over the ICD-9 taxonomy if X_l is nominal. Hence, the lower is SSE, the lower is information loss and the higher is data utility.

Considering the data distribution of evaluation data sets (see Section 3.2), we performed different anonymization tests by varying the k-anonymity level from $k = 2$ to 20. SSE values obtained for both microaggregation versions (marginality-based and classic) for the DISCHARGE NEOPLASM and DISCHARGE DIGESTIVE data are shown in Tables 2 and 3, respectively.

From the results we can see that, for both microaggregation algorithms, the information loss for numerical attributes (AGE_YRS, LOS and CHARGE) is almost identical for DISCHARGE NEOPLASM and identical for DISCHARGE DIGESTIVE for all k-anonymity levels. The information loss for the nominal attribute DIAG_P, on the contrary, is clearly reduced when using the marginality-based anonymization. The relative reduction against the non-semantic algorithm tends to grow as the anonymity parameter k grows. For DISCHARGE NEOPLASM an almost 10% reduction in information loss is obtained for $k = 4$ that improves up to a 16% reduction for $k = 20$, whereas figures for DISCHARGE DIGESTIVE are 12% and 18%, respectively. Two conclusions can be reached. First, the marginality model retains better the semantics of the nominal attribute, and hence, the utility of masked data; this illustrates the effectiveness of exploiting medical knowledge structures for anonymization of nominal attributes. Second, the information loss for nominal attributes is reduced without increasing the information loss for numerical attributes; this

Table 4: SSE values obtained for the DIAGNOSIS NEOPLASM data with marginality-based and classic microaggregation, respectively

	$SSE(DIAG\_P)$	
k	Classic	Marginality
2	242	150
4	858	280
6	1465	435
8	2018	545
10	2710	667
12	3154	785
14	3822	904
16	4012	1015
18	4570	1136
20	4846	1185

Table 5: SSE values obtained for the DIAGNOSIS DIGESTIVE data with marginality-based and classic microaggregation, respectively

	$SSE(DIAG\_P)$	
k	Classic	Marginality
2	103	103
4	310	206
6	512	323
8	701	439
10	912	560
12	1042	621
14	1197	708
16	1490	851
18	1642	945
20	1789	1038

demonstrates how the proposed S-distance (Expression 7) effectively integrates numerical and nominal attributes so that the global information loss at the record level can be minimized.

In order to make the preservation of the nominal semantics by the marginality model more evident, we re-ran the same tests above for the DIAGNOSIS NEOPLASM and DIAGNOSIS DIGESTIVE data sets, which contain only the nominal attribute (DIAG_P). Results are shown in Tables 4 and 5, respectively.

We first notice that information loss figures are lower than those in Tables 2 and 3 for the DIAG_P attribute. This is because partitioning and aggregating records in the DIAGNOSIS data sets need only be adjusted to the values of attribute DIAG_P, whereas in the DISCHARGE data sets those operations must take the values of all attributes into account, and hence, they are less adjusted to DIAG_P values. Moreover, in Table 4, the differences between the semantic and non-semantic anonymizations are more noticeable: marginality-

based anonymization reduces information loss from around 38% for $k = 2$ to more than 75% for $k = 20$ for NEOPLASM data, whereas figures for DIGESTIVE data go from 33% for $k = 4$ to 42% for $k = 20$. The smaller differences observed for the latter data set are caused by the lower number of unique or low-frequency values (as shown in Figure 5). To sum up, the exploitation of medical knowledge to partition and aggregate diagnosis data turns out to be crucial to produce a semantically coherent anonymization, which cannot be achieved by classic methods focused only on the distribution of data.

4. Conclusion and future research

Nominal attributes are common in healthcare data and they are often among the most important ones (for example, "Diagnosis" or "Treatment"). When used for secondary purposes, like clinical research, data containing nominal attributes must be anonymized, but, unfortunately, most existing anonymization techniques neglect nominal semantics, a circumstance that negatively affects the utility of anonymization results. In this study, we have addressed this issue.

We have described a knowledge-based numerical mapping for nominal attributes, called marginality. With this mapping, any anonymization procedure for numerical data can be employed to anonymize nominal data, as long as the anonymized nominal attributes take values in the same set of categories as the original nominal attributes. We have illustrated the application of this approach using microaggregation. An empirical evaluation performed using real medical data shows a noticeable reduction of the information loss during the anonymization of nominal data, thanks to the better preservation of their semantics.

Future research will involve enlarging the choice of numerical anonymization techniques that can be adapted for nominal anonymization using marginality. As an example, if noise addition is used (*e.g.* see [19] for a survey of noise addition methods), the marginality-converted original nominal attributes take the marginalities of original categories as values; then, anonymization adds noise to those marginalities, so that the noise-added marginalities no longer correspond to any original category. This makes it impossible to map the noise-added marginalities back to the original nominal categories. One could think of an approximate reverse mapping for methods which perturb input marginalities; that is, each numerical output marginality m could be mapped back to the original category having marginality closest to m. However, approximate reverse mapping can lead to gross distortion if there are categories very distant within the taxonomy that have similar marginalities, because they could be unduly swapped. Hence, blocking strategies or other mechanisms should be devised to avoid such undesirable effects.

Appendix: Proofs

Lemma 1. *Given non-negative* A, A', A'', B, B', B'' *such that* $\sqrt{A} \leq \sqrt{A'} + \sqrt{A''}$ *and* $\sqrt{B} \leq \sqrt{B'} + \sqrt{B''}$*, it holds that*

$$\sqrt{A+B} \leq \sqrt{A'+B'} + \sqrt{A''+B''} \tag{8}$$

Proof. Squaring the two inequalities in the lemma assumption, we obtain

$$A \leq (\sqrt{A'} + \sqrt{A''})^2$$

$$B \leq (\sqrt{B'} + \sqrt{B''})^2$$

Adding both expressions above, we get the square of the left-hand side of Expression (8)

$$A + B \leq (\sqrt{A'} + \sqrt{A''})^2 + (\sqrt{B'} + \sqrt{B''})^2$$

$$= A' + A'' + B' + B'' + 2(\sqrt{A'A''} + \sqrt{B'B''}) \tag{9}$$

Squaring the right-hand side of Expression (8), we get

$$(\sqrt{A' + B'} + \sqrt{A'' + B''})^2$$

$$= A' + B' + A'' + B'' + 2\sqrt{(A' + B')(A'' + B'')} \tag{10}$$

Since Expressions (9) and (10) both contain the terms $A' + B' + A'' + B''$, we can neglect them. Proving Inequality (8) is equivalent to proving

$$\sqrt{A'A''} + \sqrt{B'B''} \leq \sqrt{(A' + B')(A'' + B'')}$$

Suppose the opposite, that is,

$$\sqrt{A'A''} + \sqrt{B'B''} > \sqrt{(A' + B')(A'' + B'')} \tag{11}$$

Square both sides:

$$A'A'' + B'B'' + 2\sqrt{A'A''B'B''} >$$

$$(A' + B')(A'' + B'') = A'A'' + B'B'' + A'B'' + B'A''$$

Subtract $A'A'' + B'B''$ from both sides to obtain

$$2\sqrt{A'A''B'B''} > A'B'' + B'A''$$

which can be rewritten as

$$(\sqrt{A'B''} - \sqrt{B'A''})^2 < 0$$

Since a real square cannot be negative, the assumption in Expression (11) is false and the lemma follows. □

Theorem 3. *The S-distance based on the marginality-based variance as per Definition 2 and computed on multivariate records consisting of nominal attributes is a distance in the mathematical sense.*

Proof. We must prove that the S-distance is non-negative, reflexive, symmetrical and subadditive (*i.e.* it satisfies the triangle inequality).

Non-negativity. The S-distance is defined as a non-negative square root, hence it cannot be negative.

Reflexivity. If $\mathbf{x}_1 = \mathbf{x}_2$, then $\delta(\mathbf{x}_1, \mathbf{x}_2) = 0$. Conversely, if $\delta(\mathbf{x}_2, \mathbf{x}_2) = 0$, the variances are all zero, hence $\mathbf{x}_1 = \mathbf{x}_2$.

Symmetry. It follows from the definition of the S-distance.

Subadditivity. Given three records $\mathbf{x}_1$, $\mathbf{x}_2$ and $\mathbf{x}_3$, we must check whether

$$\delta(\mathbf{x}_1, \mathbf{x}_3) \stackrel{?}{\leq} \delta(\mathbf{x}_1, \mathbf{x}_2) + \delta(\mathbf{x}_2, \mathbf{x}_3)$$

By expanding the above expression using the definition of S-distance, we obtain

$$\sqrt{\frac{(S^2)^1_{13}}{(S^2)^1} + \cdots + \frac{(S^2)^d_{13}}{(S^2)^d}} \stackrel{?}{\leq}$$

$$\sqrt{\frac{(S^2)^1_{12}}{(S^2)^1} + \cdots + \frac{(S^2)^d_{12}}{(S^2)^d}} + \sqrt{\frac{(S^2)^1_{23}}{(S^2)^1} + \cdots + \frac{(S^2)^d_{23}}{(S^2)^d}} \tag{12}$$

Let us start with the case $d = 1$, that is, with a single attribute, *i.e.* $\mathbf{x}_i = x_i$ for $i = 1, 2, 3$. To check Inequality (12) with $d = 1$, we can ignore the variance in the denominators (it is the same on both sides) and we just need to check

$$\sqrt{S^2_{13}} \stackrel{?}{\leq} \sqrt{S^2_{12}} + \sqrt{S^2_{23}} \tag{13}$$

We have

$$S^2_{13} = Var(\{x_1, x_3\}) = \frac{m(x_1) + m(x_3)}{2}$$

$$= \frac{d(x_1, x_3)}{2} + \frac{d(x_3, x_1)}{2} = d(x_1, x_3) \tag{14}$$

Similarly $S^2_{12} = d(x_1, x_2)$ and $S^2_{23} = d(x_2, x_3)$. Therefore, Expression (13) is equivalent to subadditivity for $d(\cdot, \cdot)$ and the latter is proven in [2]. Let us now make the induction hypothesis for $d-1$ and prove subadditivity for any d. Call now

$$A := \frac{(S^2)^1_{13}}{(S^2)^1} + \cdots + \frac{(S^2)^{d-1}_{13}}{(S^2)^{d-1}}$$

$$A' := \frac{(S^2)^1_{12}}{(S^2)^1} + \cdots + \frac{(S^2)^{d-1}_{12}}{(S^2)^{d-1}}$$

$$A'' := \frac{(S^2)^1_{23}}{(S^2)^1} + \cdots + \frac{(S^2)^{d-1}_{23}}{(S^2)^{d-1}}$$

$$B := \frac{(S^2)^d_{13}}{(S^2)^d};\ B' := \frac{(S^2)^d_{12}}{(S^2)^d};\ B'' := \frac{(S^2)^d_{23}}{(S^2)^d}$$

Subadditivity for d amounts to checking whether

$$\sqrt{A+B} \stackrel{?}{\leq} \sqrt{A'+B'} + \sqrt{A''+B''} \tag{15}$$

which holds by Lemma 1 because, by the induction hypothesis for $d-1$, we have $\sqrt{A} \leq \sqrt{A'} + \sqrt{A''}$ and, by the proof for $d = 1$, we have $\sqrt{B} \leq \sqrt{B'} + \sqrt{B''}$. □

Theorem 4. *The S-distance based on the usual numerical variance and computed on multivariate records consisting of ordinal or numerical attributes is a distance in the mathematical sense.*

Proof. Non-negativity, reflexivity and symmetry are proven in a way analogous as in Theorem 3. As to subaddivity, we just need to prove the case $d = 1$, that is, the inequality analogous to Expression (13) for numerical variances. The proof for general d is the same as in Theorem 3. For $d = 1$, we have

$$S^2_{13} = \frac{(x_1 - x_3)^2}{2};\ S^2_{12} = \frac{(x_1 - x_2)^2}{2};\ S^2_{23} = \frac{(x_2 - x_3)^2}{2}$$

Therefore, Expression (13) obviously holds with equality in the case of numerical variances because

$$\sqrt{S^2_{13}} = \frac{x_1 - x_3}{\sqrt{2}} = \frac{(x_1 - x_2) + (x_2 - x_3)}{\sqrt{2}} = \sqrt{S^2_{12}} + \sqrt{S^2_{23}}.$$

□

Disclaimer and acknowledgments

This work was partly supported by the Government of Catalonia under grant 2009 SGR 1135, by the Spanish Government through projects TSI2007-65406-C03-01 "E-AEGIS", TIN2011-27076-C03-01 "CO-PRIVACY", TIN2012-32757 "ICWT", IPT2012-0603-430000 "BallotNext" and CONSOLIDER INGENIO 2010 CSD2007-00004 "ARES", and by the European Commission under FP7 projects "DwB" and "Inter-Trust". The second author is partially supported as an ICREA Acadèmia researcher by the Government of Catalonia. The authors are with the UNESCO Chair in Data Privacy, but they are solely responsible for the views expressed in this paper, which neither necessarily reflect the position of UNESCO nor commit that organization.

References

[1] D. Abril, G. Navarro-Arribas, V. Torra. Towards semantic microaggregation of categorical data for confidential documents, In: Proc. of the 7th International Conference on Modeling Decisions for Artificial Intelligence-MDAI 2010, LNCS 6408, Springer, 2010, pp. 266-276.

[2] M. Batet, A. Valls, K. Gibert. A distance function to assess the similarity of words using ontologies, In: XV Congreso Español sobre Tecnologías y Lógica Fuzzy, Huelva, Spain, 2010, pp. 561-566.

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

[4] M. Batet, D. Sánchez, A. Valls, K. Gibert. Semantic similarity estimation from multiple ontologies, Applied Intelligence 38(1) (2012) 29-44.

[5] R.J. Bayardo, R. Agrawal. Data privacy through optimal k-anonymization, In: Proc. of the 21st International Conference on Data Engineering-ICDE 2005, IEEE Computer Society, 2005, pp. 217-228.

[6] D. Defays, P. Nanopoulos. Panels of enterprises and confidentiality: the small aggregates method, In: Proc. of 92 Symposium on Design and Analysis of Longitudinal Surveys, Ottawa, Statistics Canada, 1993, pp. 195-204.

[7] J. Domingo-Ferrer. Marginality: a numerical mapping for enhanced exploitation of taxonomic attributes, In: Proc. of the 9th International Conference on Modeling Attributes for Artificial Intelligence-MDAI 2012, LNCS, Springer, 2012, pp. 367-381.

[8] J. Domingo-Ferrer, U. González-Nicolás. Hybrid data using microaggregation, Information Sciences 180(15) (2010) 2834-2844.

[9] 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.

[10] J. Domingo-Ferrer, F. Sebé, A. Solanas. A polynomial-time approximation to optimal multivariate microaggregation, Computers & Mathematics with Applications 55(4) (2008) 714-732.

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

[12] G.T. Duncan, M. Elliot, J.J. Salazar-González. Statistical Confidentiality: Principles and Practice, Springer, 2011.

[13] C. Dwork. Differential privacy, In Proc. of the 33rd International Colloquium on Automata, Languages and Programming-ICALP 2006, Part II, LNCS 4052, Springer, 2006, pp. 1-12.

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

[15] J.M. Gouweleeuw, P. Kooiman, L.C.R.J. Willenborg, P.P. De Wolf. Post randomisation for statistical disclosure control: theory and implementation, Research paper no. 9731, Voorburg: Statistics Netherlands, 1997.

[16] N. Guarino. Formal ontology in information systems, In: N. Guarino (Ed.), 1st International Conference on Formal Ontology in Information Systems, Trento, Italy, 1998, pp. 3-15.

[17] Y. He, J. Naughton. Anonymization of set-valued data via top-down, local generalization, In: Proc. of the 30th International Conference on Very Large Data Bases, Lyon, France, 2009, pp. 934-945.

[18] A. Hundepool, J. Domingo-Ferrer, L. Franconi, S. Giessing, R. Lenz, J. Longhurst, E. Schulte-Nordholt, G. Seri, P.P. De Wolf. Handbook on Statistical Disclosure Control (version 1.2), ESSNET SDC Project, 2010. http://neon.vb.cbs.nl/casc

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

[20] International Classification of Diseases, 9th Revision, Clinical Modification, Sixth Edition, 2008. http://icd9cm.chrisendres.com/

[21] T. Li, N. Li. Towards optimal k-anonymization, Knowledge and Data Engineering 65 (2008) 22-39.

[22] 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 (2004) 179-192.

[23] 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.

[24] Medical Subject Headings, U.S. National Library of Medicine, 2012. http://www.nlm.nih.gov/mesh

[25] K. Muralidhar, R. Sarathy. Data shuffling - a new masking approach for numerical data, Management Science 52(5) (2006) 658-670.

[26] L. Ohno-Machado, P.S.P. Silveira, S. Vinterbo. Protecting patient privacy by quantifiable control of disclosures in disseminated databases, International Journal of Medical Informatics 73 (2004) 599-606.

[27] Patient Discharge Data, Office of Statewide Health Planning & Development-OSHPD, 2010. http://www.oshpd.ca.gov/HID/Products/PatDischargeData/PublicDataSet/index.html

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

[29] K.B. Reid. Centrality measures in trees, In: Advances in Interdisciplinary Applied Discrete Mathematics, World Scientific eBook, 2010, pp. 167-197.

[30] D.B. Rubin. Discussion of statistical disclosure limitation, Journal of Official Statistics 9(2) (1993) 461-468.

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

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

[33] SNOMED-Systematized Nomenclature of Medicine, U.S. National Library of Medicine, 2012. `http://www.nlm.nih.gov/research/umls/Snomed/snomed/main.html`

[34] L. Sweeney. k-Anonymity: a model for protecting privacy, International Journal on Uncertainty, Fuzziness and Knowledge-based Systems 10 (2002) 557-570.

[35] V. Torra. Microaggregation for categorical variables: a median based approach, In: Privacy in Statistical Databases-PSD 2004, LNCS 3050, Springer, 2004, pp. 162-174.

[36] L. Willenborg, T. de Waal. Elements of Statistical Disclosure Control, Springer, 2001.

[37] Z. Wu, M. Palmer. Verbs semantics and lexical selection, In: Proc. of the 32nd Annual Meeting on Association for Computational Linguistics, Las Cruces, New Mexico, 1994, pp. 133-138.