Marginality: A Numerical Mapping for Enhanced Exploitation of Taxonomic Attributes

Josep Domingo-Ferrer

Universitat Rovira i Virgili
Dept. of Computer Engineering and Mathematics
UNESCO Chair in Data Privacy
Av. Països Catalans 26
E-43007 Tarragona, Catalonia
josep.domingo@urv.cat

Abstract. Hierarchical attributes appear in taxonomic or ontology-based data (*e.g.* NACE economic activities, ICD-classified diseases, animal/plant species, etc.). Such taxonomic data are often exploited as if they were flat nominal data without hierarchy, which implies losing substantial information and analytical power. We introduce marginality, a numerical mapping for taxonomic data that allows using on those data many of the algorithms and analytical techniques designed for numerical data. We show how to compute descriptive statistics like the mean, the variance and the covariance on marginality-mapped data. Also, we define a mathematical distance between records including hierarchical attributes that is based on marginality-based variances. Such a distance paves the way to re-using on taxonomic data clustering and anonymization techniques designed for numerical data.

Keywords: Hierarchical attributes, Classification, Taxonomic data, Ontologies, Descriptive statistics, Numerical mapping, Anonymization.

1 Introduction

Taxonomic attributes are common in economic, medical or biological data sets and, more generally, in ontology-based data sets. For example, data about companies often include an attribute "Economic activity" which takes values in a standard classification, like NACE [12] or ISIC [9]; data about employees include their position within the company's hierarchy; data about patients include an attribute "Diagnosis" which takes values in some classification of diseases, like ICD9 [8]; data about plants or animals include the name of the plant or animal in the Linnaean taxonomy [11,14], etc.

Statistical analyses tend to treat taxonomic data as if they came from flat nominal attributes without hierarchy, thereby disregarding their hierarchical semantics and losing useful information. Such a wasteful approach can be explained

V. Torra et al. (Eds.): MDAI 2012, LNAI 7647, pp. 367–381, 2012.

by the lack of analytical techniques and algorithms specifically designed for taxonomic data. Indeed, numerical data are the type of data for which a greatest choice of techniques exists; categorical ordinal data are often mapped to integers and treated like numerical data; nominal data, whether drawn from a flat or hierarchical taxonomy, are most of the time treated as flat.

The situation described in the previous paragraph repeats itself for statistical disclosure control (SDC, [6,7,17,5,10]), a.k.a. data anonymization and sometimes as privacy-preserving data mining. SDC aims at making possible the publication of statistical data in such a way that the individual responses of specific users cannot be inferred from the published data and background knowledge available to intruders. If the data set being published consists of records corresponding to individuals, usual SDC methods operate by masking original data (via perturbation or detail reduction), by generating synthetic (simulated) data preserving some statistical features of the original data or by producing hybrid data obtained as a combination of original and synthetic data. The choice of SDC methods is greatest for numerical data.

The attributes in a data set can be classified depending on their range and the operations that can be performed on them:

1. *Numerical.* An attribute is considered numerical if arithmetical operations can be performed on it. Examples are income and age.
2. *Categorical.* An attribute is considered categorical when it takes values over a finite set and standard arithmetical operations on it do not make sense. Two main types of categorical attributes can be distinguished:
   (a) *Ordinal.* An ordinal attribute takes values in an ordered range of categories. Thus, the $\leq$, max and min operators can still be used on this kind of data. The instruction level and the political preferences (left-right) are examples of ordinal attributes.
   (b) *Nominal.* A nominal attribute takes values in an unordered range of categories. The only possible operator is comparison for equality. Nominal attributes can further be divided into two types:
      i. *Hierarchical.* A hierarchical nominal attribute takes values from a hierarchical classification. For example, plants are classified using Linnaeus's taxonomy, the type of a disease is also selected from a hierarchical taxonomy, and the type of an attribute can be selected from the hierarchical classification we propose in this section.
      ii. *Non-hierarchical.* A non-hierarchical nominal attribute takes values from a flat taxonomy. Examples of such attributes could be the preferred soccer team, the address of an individual, the civil status (married, single, divorced, widow/er), the eye color, etc.

This paper focuses on finding a numerical mapping for taxonomic data. Such a mapping can be used to obtain richer descriptive statistics, inspired on those for numerical data. It also makes it possible to use on taxonomic data techniques designed for numerical data (*e.g.* clustering, SDC).

Assuming a hierarchy is less restrictive than it would appear, because very often a non-hierarchical attribute can be turned into a hierarchical one if its flat

hierarchy can be developed into a multilevel hierarchy. For instance, the preferred soccer and the address of an individual have been mentioned as non-hierarchical attributes; however, a hierarchy of soccer teams by continent and country could be conceived, and addresses can be hierarchically clustered by neighborhood, city, state, country, etc. Furthermore, well-known approaches to anonymization, like k-anonymity [15], assume that any attribute can be generalized, *i.e.* that an attribute hierarchy can be defined and values at lower levels of the hierarchy can be replaced by values at higher levels.

1.1 Contribution and Plan of This Paper

We propose to associate a number to each categorical value of a hierarchical nominal attribute, namely a form of centrality of that category within the attribute's taxonomy. We show how this allows computation of centroids, variances and covariances of hierarchical nominal data.

Section 2 gives background on the variance of hierarchical nominal attributes. Section 3 defines a tree centrality measure called marginality and presents the numerical mapping. Section 4 exploits the numerical mapping to compute means, variances and covariances of hierarchical nominal data. Section 5 contains a discussion and conclusions.

2 Background

We next recall the variance measure for hierarchical nominal attributes introduced in [4]. To the best of our knowledge, this is the first measure which captures the variability of a sample of values of a hierarchical nominal attribute by taking into account the semantics of the hierarchy. The intuitive idea is that a set of nominal values belonging to categories which are all children of the same parent category in the hierarchy has smaller variance that a set with children from different parent categories.

Algorithm 1 (Nominal variance in [4])

1. *Let the hierarchy of categories of a nominal attribute X be such that b is the maximum number of children that a parent category can have in the hierarchy.*
2. *Given a sample T_X of nominal categories drawn from X, place them in the tree representing the hierarchy of X. Prune the subtrees whose nodes have no associated sample values. If there are repeated sample values, there will be several nominal values associated to one or more nodes (categories) in the pruned tree.*
3. *Label as follows the edges remaining in the tree from the root node to each of its children:*
   - *If b is odd, consider the following succession of labels $l_0 = (b-1)/2$, $l_1 = (b-1)/2-1$, $l_2 = (b-1)/2+1$, $l_3 = (b-1)/2-2$, $l_4 = (b-1)/2+2$, $\cdots$, $l_{b-2} = 0$, $l_{b-1} = b-1$.*

- *If b is even, consider the following succession of labels $l_0 = (b-2)/2$, $l_1 = (b-2)/2+1$, $l_2 = (b-2)/2-1$, $l_3 = (b-2)/2+2$, $l_4 = (b-2)/2-2$, $\cdots$, $l_{b-2} = 0$, $l_{b-1} = b-1$.*
- *Label the edge leading to the child with most categories associated to its descendant subtree as l_0, the edge leading to the child with the second highest number of categories associated to its descendant subtree as l_1, the one leading to the child with the third highest number of categories associated to its descendant subtree as l_2 and, in general, the edge leading to the child with the i-th highest number of categories associated to its descendant subtree as l_{i-1}. Since there are at most b children, the set of labels $\{l_0, \cdots, l_{b-1}\}$ should suffice. Thus an* edge label *can be viewed as a b-ary digit (to the base b).*

4. *Recursively repeat Step 3 taking instead of the root node each of the root's child nodes.*
5. *Assign to values associated to each node in the hierarchy a* node label *consisting of a b-ary number constructed from the edge labels, more specifically as the concatenation of the b-ary digits labeling the edges along the path from the root to the node: the label of the edge starting from the root is the most significant one and the edge label closest to the specific node is the least significant one.*
6. *Let L be the maximal length of the leaf b-ary labels. Append as many l_0 digits as needed in the least significant positions to the shorter labels so that all of them eventually consist of L digits.*
7. *Let $T_X(0)$ be the set of b-ary digits in the least significant positions of the node labels (the "units" positions); let $T_X(1)$ be the set of b-ary digits in the second least significant positions of the node labels (the "tens" positions), and so on, until $T_X(L-1)$ which is the set of digits in the most significant positions of the node labels.*
8. *Compute the variance of the sample as*

$$Var_H(T_X) = Var(T_X(0)) + b^2 \cdot Var(T_X(1)) + \cdots$$

$$+b^{2(L-1)} \cdot Var(T_X(L-1)) \quad (1)$$

where $Var(\cdot)$ is the usual numerical variance.

In Section 4.2 below we will show that an equivalent measure can be obtained in a simpler and more manageable way.

3 A Numerical Mapping for Nominal Hierarchical Data

Consider a nominal attribute X taking values from a hierarchical classification. Let T_X be a sample of values of X. Each value $x \in T_X$ can be associated two numerical values:

- The sample frequency of x;
- Some centrality measure of x within the hierarchy of X.

While the frequency depends on the particular sample, centrality measures depend both on the attribute hierarchy and the sample. Known tree centralities attempt to determine the "middle" of a tree [13]. We are rather interested in finding how far from the middle is each node of the tree, that is, how marginal it is. We next propose an algorithm to compute a new measure of the marginality of the values in the sample T_X.

Algorithm 2 (Marginality of hierarchical values)

1. *Given a sample T_X of hierarchical nominal values drawn from X, place them in the tree representing the hierarchy of X. There is a one-to-one mapping between the set of tree nodes and the set of categories where X takes values. Prune the subtrees whose nodes have no associated sample values. If there are repeated sample values, there will be several nominal values associated to one or more nodes (categories) in the pruned tree.*
2. *Let L be the depth of the pruned tree. Associate weight 2^{L-1} to edges linking the root of the hierarchy to its immediate descendants (depth 1), weight 2^{L-2} to edges linking the depth 1 descendants to their own descendants (depth 2), and so on, up to weight $2^0 = 1$ to the edges linking descendants at depth $L-1$ with those at depth L. In general, weight 2^{L-i} is assigned to edges linking nodes at depth $i-1$ with those at depth i, for $i = 1$ to L.*
3. *For each nominal value x_j in the sample, its marginality $m(x_j)$ is defined and computed as*

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

where $d(x_j, x_l)$ is the sum of the edge weights along the path from the tree node corresponding to x_j and the tree node corresponding to x_l.

Note 1 (On distances and marginality). The above construction of marginality can be generalized by allowing other distance functions to be used in Expression (2), not necessarily based on edge weights. For example, in [3] it is suggested to use the semantic distance proposed in [16], in which the distance between two categories in a taxonomy is a function of the number of non-common ancestors divided by the total number of ancestors of the category pair.

Clearly, the greater $m(x_j)$, the more marginal (*i.e.* the less central) is x_j. We give next a toy running example to illustrate the computation of marginality.

Example 1. Assume a hierarchical attribute *"Diagnosis"*, for which a sample is available whose nominal values can be hierarchically classified as shown in Figure 1. The hierarchy is a pruned one, so that only leaves with some value in the sample are depicted. The sample has one element for each diagnostic category, except for "Epilepsy" and "Nose cold", for each of which there are two elements. Figure 1 also shows the weights assigned by Algorithm 2 to each edge in the hierarchy tree.

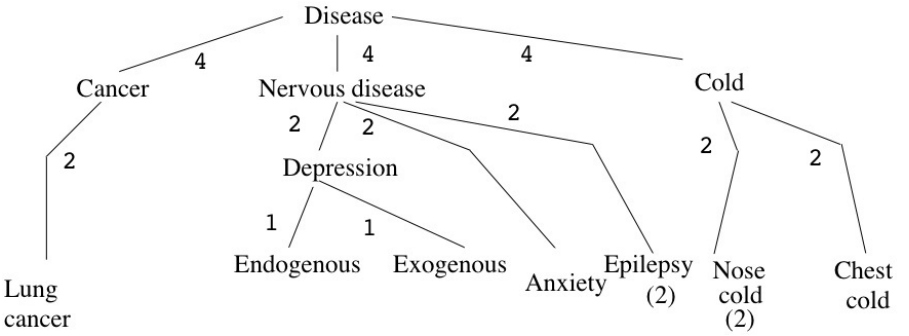

Fig. 1. Example pruned hierarchy of a sample of a *"Diagnosis"* attribute

Label the elements in the sample as follows: x_1 (lung cancer), x_2 (endogenous depression), x_3 (exogenous depression), x_4 (anxiety), x_5 (first epilepsy element), x_6 (second epilepsy element), x_7 (first nose cold element), x_8 (second nose cold element) and x_9 (chest cold). The distance matrix between elements is given below, where component (j, l) represents the sum $d(x_j, x_l)$ of edge weights along the path between x_j and x_l (only the upper diagonal matrix is represented):

$$\begin{pmatrix} 0 & 13 & 13 & 12 & 12 & 12 & 12 & 12 & 12 \\ & 0 & 2 & 5 & 5 & 5 & 13 & 13 & 13 \\ & & 0 & 5 & 5 & 5 & 13 & 13 & 13 \\ & & & 0 & 4 & 4 & 12 & 12 & 12 \\ & & & & 0 & 0 & 12 & 12 & 12 \\ & & & & & 0 & 12 & 12 & 12 \\ & & & & & & 0 & 0 & 4 \\ & & & & & & & 0 & 4 \\ & & & & & & & & 0 \end{pmatrix}$$

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 hierarchy in Figure 1. On the other hand, x_5 and x_6 are the least marginal elements, due to both the central position of epilepsy in the hierarchy and the fact that there are two epilepsy elements. In fact, the higher frequency of epilepsy is what makes the marginality of x_5 and x_6 lower than the marginality of x_4 (anxiety); otherwise, epilepsy and anxiety have equally central positions in the hierarchy. This illustrates that marginality is a function of both the hierarchy of categories and their frequency in the sample.

Some properties are next stated which illustrate the rationale of the distance and the weights used to compute marginalities.

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

x_j	$m(x_j)$
x_1	$0+13+13+12+12+12+12+12+12=98$
x_2	$13+0+2+5+5+5+13+13+13=69$
x_3	$13+2+0+5+5+5+13+13+13=69$
x_4	$12+5+5+0+4+4+12+12+12=66$
x_5	$12+5+5+4+0+0+12+12+12=62$
x_6	$12+5+5+4+0+0+12+12+12=62$
x_7	$12+13+13+12+12+12+0+0+4=78$
x_8	$12+13+13+12+12+12+0+0+4=78$
x_9	$12+13+13+12+12+12+4+4+0=82$

Lemma 1. *$d(\cdot,\cdot)$ is a distance in the mathematical sense.*

Being the length of a path, it is immediate to check that $d(\cdot,\cdot)$ satisfies reflexivity, symmetry and subadditivity. The rationale of the above exponential weight scheme is to give more weight to differences at higher levels of the hierarchy; specifically, the following property is satisfied.

Lemma 2. *The distance between any non-root node n_j and its immediate ancestor is greater than the distance between n_j and any of its descendants.*

Proof: Let L be the depth of the overall tree and L_j be the depth of n_j. The distance between n_j and its immediate ancestor is 2^{L-L_j}. The distance between n_j and its most distant descendant is

$$1+2+\cdots+2^{L-L_j-1}=2^{L-L_j}-1$$

□

Lemma 3. *The distance between any two different nodes at the same depth is greater than the longest distance within the subtree rooted at each node.*

Proof: Let L be the depth of the overall tree and L_j be the depth of the two nodes. The distance between two different nodes is shortest when they have the same parent and it is

$$2\cdot 2^{L-L_j}=2^{L-L_j+1}.$$

The longest distance within any of the two subtrees rooted at the two nodes at depth L_j is the length of the path between two leaves at depth L, which is

$$2\cdot(1+2+\cdots+2^{L-L_j-1})=2(2^{L-L_j}-1)=2^{L-L_j+1}-2$$

□

4 Statistical Analysis of Numerically Mapped 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 magnitude can be used in statistical analysis.

4.1 Mean

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

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

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

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

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

if one wants a numerical mean value.

Example 2. It can be seen from Table 1 that, for the sample of Example 1, the marginality-based mean is "Epilepsy" (which is the least marginal value) and the numerical marginality-based mean is 62.

4.2 Variance

In Section 2 above, we recalled a measure of variance of a hierarchical nominal attribute proposed in [4] which takes the semantics of the hierarchy into account. Interestingly, it turns out that the average marginality of a sample is an equivalent way to capture the same notion of variance.

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

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

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

$$\frac{98 + 69 + 69 + 66 + 62 + 62 + 78 + 78 + 82}{9} = 73.78$$

The following lemma is proven in the Appendix.

Lemma 4. *The $Var_M(\cdot)$ measure and the $Var_H(\cdot)$ specified by Algorithm 1 in Section 2 are equivalent.*

4.3 Covariance Matrix

It is not difficult to generalize the sample variance introduced in Definition 2 to define the sample covariance of two nominal attributes.

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*

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

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 $T_{jl} = (T_j, T_l)$.

4.4 Variance-Based Distance

Based on variances (whether plain numerical or marginality-based), we can define the following distance for records with numerical, hierarchical or flat nominal attributes.

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*

$$\delta(\mathbf{x}_1, \mathbf{x}_2) = \sqrt{\frac{(S^2)^1_{12}}{(S^2)^1} + \cdots + \frac{(S^2)^d_{12}}{(S^2)^d}} \tag{3}$$

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.

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 on multivariate records consisting of hierarchical attributes based on the hierarchical variance computed as per Definition 2 is a distance in the mathematical sense.*

Theorem 2. *The S-distance on multivariate records consisting of ordinal or numerical attributes based on the usual numerical variance 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 hierarchical variance is used for hierarchical and flat nominal attributes and the usual numerical variance is used for ordinal and numerical attributes, is a distance in the mathematical sense.*

The above distance can be used for a variety of purposes, including clustering. Specifically, it allows microaggregating hierarchical data [1,2] in view of anonymization.

5 Discussion and Conclusions

We have presented a centrality-based mapping of hierarchical nominal data to numbers. We have shown how such a numerical mapping allows computing means, variances and covariances of nominal attributes, and distances between records containing any kind of attributes.

Such enhanced flexibility of manipulation can be used to adapt methods intended for numerical data to the treament of hierarchical attributes. If reverse mapping to nominal categories is required at the end of the treatment, two situations arise:

- *Each numerical output of the method exactly equals one of the input marginalities. E.g.* this happens for SDC methods that involve swapping input values that are within a certain distance of each other. In this case, each numerical output m is mapped back to the nominal category having marginality m.
- *Numerical outputs do not correspond to marginalities. E.g.* such is the case if numerical outputs are the result of applying a regression model on the input marginalities. In this case, a reasonable option is to map each numerical output m back to the category having marginality closest to m.

Reverse mapping may be problematic if there are categories which are semantically very different and have similar marginalities or the same marginality. For example, if the nose colds are suppressed from the sample depicted in Figure 1, then chest cold and lung cancer would have exactly the same marginality. A way to prevent semantic confusion in reverse mapping is to use blocking, that is, to split the hierarchy tree into several subtrees based on semantic criteria and treat each subtree separately: *e.g* divide the sample of Example 1 into a subsample of cancers, a subsample of nervous diseases and a subsample of colds, and treat subsamples separately to avoid big confusions during reverse mapping (we are assuming that confusing two categories within the same subtree is tolerable).

Future research will involve developing real-life applications of marginality, for example data anonymization of hierarchical attributes using SDC methods intended for numerical data (like multiple imputation or microaggregation).

Appendix

Proof (Lemma 4): We will show that, given two samples $T_X = \{x_1, \cdots, x_n\}$ and $T'_X = \{x'_1, \cdots, x'_n\}$ of a nominal attribute X, both with the same cardinality n, it holds that $Var_M(T_X) < Var_M(T'_X)$ if and only if $Var_H(T_X) < Var_H(T'_X)$.

Assume that $Var_M(T_X) < Var_M(T'_X)$. Since both samples have the same cardinality, this is equivalent to

$$\sum_{j=1}^{n} m(x_j) < \sum_{j=1}^{n} m(x'_j)$$

By developing the marginalities, we obtain

$$\sum_{j=1}^{n} \sum_{x_l \in T_X - \{x_j\}} d(x_j, x_l) < \sum_{j=1}^{n} \sum_{x'_l \in T'_X - \{x'_j\}} d(x'_j, x'_l)$$

Since distances are sums of powers of 2, from 1 to 2^{L-1}, we can write the above inequality as

$$d_0 + 2d_1 + \cdots + 2^{L-1} d_{L-1} < d'_0 + 2d'_1 + \cdots + 2^{L-1} d'_{L-1} \tag{4}$$

By viewing $d_{L-1} \cdots d_1 d_0$ and $d'_{L-1} \cdots d'_1 d'_0$ as binary numbers, it is easy to see that Inequality (4) implies that some i must exist such that $d_i < d'_i$ and $d_{\hat{i}} \leq d'_{\hat{i}}$ for $i < \hat{i} \leq L-1$. This implies that there are less high-level edge differences associated to the values of T_X than to the values of T'_X. Hence, in terms of $Var_H(\cdot)$, we have that $Var(T_X(i)) < Var(T'_X(i))$ and $Var(T_X(\hat{i})) \leq Var(T'_X(\hat{i})$ for $i < \hat{i} \leq L-1$. This yields $Var_H(T_X) < Var_H(T'_X)$.

If we now assume $Var_H(T_X) < Var_H(T'_X)$, we can prove $Var_M(T_X) < Var_M(T'_X)$ by reversing the above argument. □.

Lemma 5. *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*

$$\sqrt{A+B} \leq \sqrt{A'+B'} + \sqrt{A''+B''} \tag{5}$$

Proof (Lemma 5): 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 (5)

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

$$= A' + A'' + B' + B'' + 2(\sqrt{A'A''} + \sqrt{B'B''}) \tag{6}$$

Squaring the right-hand side of Expression (5), we get

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

$$= A' + B' + A'' + B'' + 2\sqrt{(A'+B')(A''+B'')} \tag{7}$$

Since Expressions (6) and (7) both contain the terms $A' + B' + A'' + B''$, we can neglect them. Proving Inequality (5) is equivalent to proving

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

Suppose the opposite, that is,

$$\sqrt{A'A''} + \sqrt{B'B''} > \sqrt{(A' + B')(A'' + B'')} \tag{8}$$

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 (8) is false and the lemma follows. □

Proof (Theorem 1): 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 Expression (3), we obtain

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

$$\sqrt{\frac{(S^2)^1_{12}}{(S^2)^1} + \cdots + \frac{(S^2)^d_{12}}{(S^2)^d}} + \sqrt{\frac{(S^2)^1_{23}}{(S^2)^1} + \cdots + \frac{(S^2)^d_{23}}{(S^2)^d}} \tag{9}$$

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 (9) with $d = 1$, we can ignore the variance in the denominators (it is the same on both sides) and we just need to check

$$\sqrt{S^2_{13}} \stackrel{?}{\leq} \sqrt{S^2_{12}} + \sqrt{S^2_{23}} \tag{10}$$

We have

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

$$= \frac{d(x_1, x_3)}{2} + \frac{d(x_3, x_1)}{2} = d(x_1, x_3) \tag{11}$$

Similarly $S^2_{12} = d(x_1, x_2)$ and $S^2_{23} = d(x_2, x_3)$. Therefore, Expression (10) is equivalent to subaddivitity for $d(\cdot, \cdot)$ and the latter holds by Lemma 1. 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

$$\sqrt{A + B} \stackrel{?}{\leq} \sqrt{A' + B'} + \sqrt{A'' + B''} \tag{12}$$

which holds by Lemma 5 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 $\sqrt{B} \leq \sqrt{B'} + \sqrt{B''}$. □

Proof (Theorem 2): Non-negativity, reflexivity and symmetry are proven in a way analogous as in Theorem 1. As to subaddivity, we just need to prove the case $d = 1$, that is, the inequality analogous to Expression (10) for numerical variances. The proof for general d is the same as in Theorem 1. 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 (10) obviously holds with equality in the case of numerical variances because

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

□

Acknowledgments and Disclaimer. Thanks go to Klara Stokes for help with Lemma 5. 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" and CONSOLIDER INGENIO 2010 CSD2007-00004 "ARES", and by the European Comission under FP7 projects "DwB" and "Inter-Trust". The author is partially supported as an ICREA Acadèmia researcher by the Government of Catalonia. The author is with the UNESCO Chair in Data Privacy, but he is solely responsible for the views expressed in this paper, which do not necessarily reflect the position of UNESCO nor commit that organization.

References

1. Domingo-Ferrer, J., Mateo-Sanz, J.M.: Practical data-oriented microaggregation for statistical disclosure control. IEEE Transactions on Knowledge and Data Engineering 14(1), 189–201 (2002)
2. Domingo-Ferrer, J., Torra, V.: Ordinal, continuous and heterogeneous k-anonymity through microaggregation. Data Mining and Knowledge Discovery 11(2), 195–212 (2005)
3. Domingo-Ferrer, J., Sánchez, D., Rufian-Torrell, G.: Anonymization of clinical data based on semantic marginality (manuscript, 2012)
4. Domingo-Ferrer, J., Solanas, A.: A measure of nominal variance for hierarchical nominal attributes. Information Sciences 178(24), 4644–4655 (2008); Erratum in Information Sciences 179(20), 3732 (2009)
5. Duncan, G.T., Elliot, M., Salazar-González, J.-J.: Statistical Confidentiality: Principles and Practice. Springer, New York (2011)
6. Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Lenz, R., Longhurst, J., Schulte-Nordholt, E., Seri, G., DeWolf, P.-P.: Handbook on Statistical Disclosure Control (version 1.2). ESSNET SDC Project (2010), `http://neon.vb.cbs.nl/casc`
7. Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Schulte Nordholt, E., Spicer, K., De Wolf, P.P.: Statistical Disclosure Control. Wiley, New York (2012)
8. ICD9 - International Classification of Diseases, 9th Revision, Clinical Modification, 6th edn., October 1 (2008), `http://icd9cm.chrisendres.com/`
9. ISIC Rev. 4 - International Standard Industrial Classification of All Economic Activities, United Nations Statistics Division, `http://unstats.un.org/unsd/cr/registry/regcst.asp?Cl=27&prn=yes`
10. Lenz, R.: Methoden der Geheimhaltung wirtschaftsstatistischer Einzeldaten und ihre Schutzwirkung. Statistik und Wissenschaft, vol. 18. Statistisches Bundesamt, Wiesbaden (2010)
11. McNeill, J., et al. (eds.): International Code of Botanical Nomenclature (Vienna Code). International Association for Plant Taxonomy (2006), `http://ibot.sav.sk/icbn/main.htm`
12. NACE Rev. 2 - Statistical Classification of Economic Activities in the European Community, Rev. 2. Eurostat, European Commission (2008), `http://epp.eurostat.ec.europa.eu/cache/ITY_OFFPUB/KS-RA-07-015/EN/KS-RA-07-015-EN.PDF`

13. Reid, K.B.: Centrality measures in trees. In: Kaul, H., Mulder, H.M. (eds.) Advances in Interdisciplinary Applied Discrete Mathematics, pp. 167–197. World Scientific eBook (2010)
14. Ride, W.D.L., et al. (eds.): International Code of Zoological Nomenclature, 4th edn., January 1. International Union of Biological Sciences (2000), http://www.nhm.ac.uk/hosted-sites/iczn/code/
15. Samarati, P.: Protecting respondents' identities in microdata release. IEEE Transactions on Knowledge and Data Engineering 13(6), 1010–1027 (2001)
16. Sánchez, D., Batet, M., Isern, D., Valls, A.: Ontology-based semantic similarity: a new feature-based approach. Expert Systems with Applications 39(9), 7718–7728 (2012)
17. Willenborg, L., DeWaal, T.: Elements of Statistical Disclosure Control. Springer, New York (2001)