Anonymization Methods for Taxonomic Microdata

Josep Domingo-Ferrer[1], Krish Muralidhar[2], and Guillem Rufian-Torrell[1]

[1] Universitat Rovira i Virgili
UNESCO Chair in Data Privacy
Department of Computer Engineering and Maths
Av. Països Catalans 26, E-43007 Tarragona, Catalonia
josep.domingo@urv.cat, guillem.rufian@estudiants.urv.cat
[2] University of Kentucky
Gatton College of Business & Economics
Lexington KY 40506, USA
krish.muralidhar@uky.edu

Abstract. Often microdata sets contain attributes which are neither numerical nor ordinal, but take nominal values from a taxonomy, ontology or classification (*e.g.* diagnosis in a medical data set about patients, economic activity in an economic data set, etc.). Such data sets must be anonymized if transferred outside the data collector's premises (*e.g.* hospital or national statistical office), say, for research purposes. The literature on microdata anonymization methods is relatively limited for nominal data. Multiple imputation is a usual choice for such data, but it has computational problems when nominal attributes can take many possible different values. In this paper, we provide anonymization methods for data sets which include nominal taxonomic attributes with many possible different values.

We show how to adapt to the case of taxonomic attributes two anonymization methods, data shuffling and microaggregation, that were originally designed for numerical attributes. The above adaptation relies on a hierarchy-aware numerical mapping of nominal categories, which we call marginality. The resulting adapted methods circumvent the computational problems of multiple imputation and take the semantics of the taxonomy into account.

Keywords: Statistical disclosure control, Hierarchical attributes, Classification, Taxonomic data, Nominal data, Data anonymization.

1 Introduction

Statistical disclosure control (SDC, [2,10–12,15,25]), a.k.a. data anonymization, or privacy-preserving data mining/publishing, 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. The anonymized data should still preserve enough analytical validity for their publication to be useful to potential users.

J. Domingo-Ferrer and I. Tinnirello (Eds.): PSD 2012, LNCS 7556, pp. 90–102, 2012.

A microdata set can be defined as a file with a number of records, where each record contains a number of attributes on an individual, *e.g.* a patient. As an illustration consider the Patient Discharge Data Set 2010 that can be obtained as a public file on CD from California's Office of Statewide Health Planning and Development [19]. Among other attributes, this data set contains the six following ones: AGE_YRS (age in years), SEX (male and female), RACE (six categories), LOS (length of stay from admission to discharge in days), CHARGE (in dollars) and DIAG_P (principal diagnosis).

Attributes can be classified depending on their range and the operations that can be performed on them:

1. *Numerical*. These are attributes on which arithmetical operations can be performed. Among the six attributes listed above, AGE_YRS, LOS and CHARGE are numerical.
2. *Categorical*. These are attributes taking values over a finite set and standard arithmetical operations on them do not make sense. There are two main types of categorical attributes:
   (a) *Ordinal*. These attributes take values in an ordered range of categories; *e.g.* "Wound classification" (slight, serious, very serious, critical).
   (b) *Nominal*. These take 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 in a hierarchical taxonomy, ontology or classification. *E.g.*, among the six attributes listed above, the principal diagnosis DIAG_P is coded according to the International Classification of Diseases (ICD-9-CM, [13]). The economic activity of a company according to the NACE classification ([17]) would be another example.
      ii. *Non-hierarchical*. A non-hierarchical nominal attribute takes values in a flat hierarchy. *E.g.*, the SEX and RACE attributes are flat non-hierarchical.

Using existing data anonymization techniques with hierarchical nominal attributes while preserving the original categories (no loss of detail) is not easy. Consider the above mentioned Patient Discharge Data Set 2010. Even if we restrict to the subset of records for which DIAG_P is some form of neoplasm, the DIAG_P attribute in that subset alone takes as many as 542 different categories. For most anonymization techniques not generalizing original categories, it would be necessary to represent each possible category using a binary 0-1 attribute; in the proposed subset of neoplasm records, this would imply adding 541 binary attributes just to represent the DIAG_P attribute. For other anonymization techniques, such as PRAM [11], no binary attributes would be needed, but in the worst-case a Markov transition matrix of size 542×542 would have to be selected as a parameter to control the anonymization of DIAG_P (if the degrees of freedom of PRAM are reduced, many transition probabilities will be 0 and the matrix parameter will become certainly more manageable). Anyway, in

either case, nominal hierarchical attributes greatly increase the computational complexity of anonymization.

Consider for example multiple imputation [22], a commonly used method for data anonymization that provides high data utility and low disclosure risk for the anonymized data [21]. Multiple imputation can be implemented using the sequential regression imputation method (SRMI), described in [20] and available from those authors in the SYNTHESIZE module of the free IVEware package [14]. In order to deal with nominal attributes, SRMI creates a binary 0-1 regressor attribute B_{ij} for each nominal category j of each nominal attribute i, so that $B_{ij} = 1$ for a record if and only if attribute i takes category j for that record.

When we attempt to use SRMI on the NEOPLASM data, we are confronted to a regression model where DIAG_P is replaced with 541 binary regressors. This model is far too large for IVEware to compute with reasonable resources. The result is that SRMI is a poor choice to anonymize the DIAG_P attribute. As mentioned earlier, this is true not only for multiple imputation, but for all anonymization techniques that rely on converting the hierachical nominal data into binary 0-1 attributes.

Even if it was computationally feasible to create binary attributes, these would not capture the category hierarchy of the hierarchical nominal attribute. This is a serious shortcoming for anonymizing data sets which contain important hierarchical nominal attributes (like biomedical or economic data sets).

Other methods, like generalization or suppression, are indeed applicable to hierarchical data, but for some applications they cause an undesirable loss of information and/or detail (*e.g.* generalizing the values of the DIAG_P attribute would greatly diminish the analytical value of NEOPLASM).

1.1 Contribution and Plan of This Paper

We show how anonymization methods originally designed for numerical microdata can be adapted for use with nominal and hierarchical nominal attributes thanks to a hierarchy-aware numerical mapping. In this way, the above-mentioned problems of computational complexity and loss of detail are avoided. Specifically, we use the mapping to adapt the data shuffling and microaggregation methods.

Section 2 gives background on data shuffling and microaggregation. Section 3 introduces the hierarchy-aware numerical mapping and shows how to combine it with data shuffling and microaggregation. Empirical results are presented and discussed in Section 4. Section 5 summarizes conclusions and lines for future research.

2 Background

We next review the two anonymization methods that will be used as building blocks of the adapted methods we propose.

2.1 Data Shuffling

Data shuffling is a data masking procedure for numerical data introduced in [16]. It combines the best of perturbation and swapping. Data perturbation modifies (perturbs) the original values, which is a disadvantage, but it has the strength of providing high utility and low disclosure risk (if the perturbation is well-tuned). Data swapping by contrast does not modify the original values, but it swaps (exchanges) them between records. However, swapping cannot assure a disclosure risk as low as perturbation. By combining perturbation and swapping, data shuffling offers a versatile hybrid method that provides very high data utility and low disclosure risk.

Data shuffling can be described as follows. Let X represent the set of confidential attributes and let S represent the set of non-confidential attributes. Data shuffling models the joint distribution of $\{X, S\}$ as a multivariate normal (Gaussian) copula. In the copula approach, the joint distribution of $\{X, S\}$ is characterized by the rank order correlation matrix $\mathbf{R}$. The normalized values of the data set $\{X^*, S^*\}$ are characterized by the product moment correlation ρ, where the relationship between $\mathbf{R} = (r_{ij})$ and $\rho = (\rho_{ij})$ is $\rho_{ij} = 2\pi \sin(r_{ij})/6$ with r_{ij} and ρ_{ij} being, respectively, the rank order and product moment correlation between attributes i and j. The normalized values $\{X^*, S^*\}$ are described by a multivariate normal distribution characterized by ρ. Using this information, the perturbed normalized values Y^* are created using the conditional distribution $f(X^*|S^*)$. The generated values Y^* are independent of X^* (and X). Once the values of Y^* have been generated in this manner, the original values of X are reverse mapped to Y^* to obtain the perturbed values Y. For more details, see [16].

Data shuffling offers the following advantages:

1. The shuffled values Y have the same marginal distribution as the original values X. Hence, the results of all univariate analyses using Y provide exactly the same results as analyses using X.
2. The rank order correlation matrix of $\{Y, S\}$ is asymptotically the same as the rank order correlation matrix of $\{X, S\}$. In other words, for large data sets, data shuffling preserves linear and monotonic non-linear relationships. Hence, results of all analyses that involve linear relationship between attributes (such as regression analyses) or non-linear relationships using the shuffled data $\{Y, S\}$ will be very similar to results using the original data $\{X, S\}$. Furthermore, as the size of the data set increases, the difference in the results will also get smaller.
3. The shuffled values Y are generated independently of the original values X, based only on the values of the non-confidential attributes S. In the absence of non-confidential attributes (that is, when S is null), X and Y are statistically independent, that is, Y is synthetic. Theoretically, this provides the lowest possible level of disclosure risk. Thus, data shuffling minimizes the risk of disclosure.

2.2 Microaggregation

Microaggregation is a family of perturbative SDC methods originally defined for numerical data [1,6]. Microaggregation can be defined in terms of two steps:

Partition: The set of original records is partitioned into small 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.

Aggregation: An aggregation operator (typically the group centroid/mean) is computed for each group and is used to replace the original records. In other words, each record in a group is replaced by the group prototype.

Computational improvements of microaggregation can be found in [5, 7, 9, 18]. In [9,24], 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 (with no hierarchy) and proposed the modal value as an aggregation operator for them. Clearly, the modal value is a very coarse aggregation operator which may not even be uniquely defined, especially over a small group of values. Also, the two extensions mentioned fail to capture the semantics of hierarchical nominal data.

In [4], microaggregation-based hybrid generation for numerical data was proposed. The idea is to replace the aggregation step of microaggregation by a synthetic data generator preserving means and covariances, in such a way that the resulting anonymized data set exactly preserves the means and covariances of the original data set and approximately preserves them in random subdomains.

3 Our Method

In this section, we first describe a hierarchy-aware numerical mapping of hierarchical nominal attributes. Then we show how this approach can be used effectively for anonymization techniques designed for numerical data. Consider a nominal attribute X taking values from a hierarchical classification. Let T_X be a sample of values of X. We propose the following algorithm to compute a new measure of the marginality (non-centrality) of the values in the sample T_X.

Algorithm 1 (Marginality of nominal values)

1. *Given a sample T_X of nominal categorical 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 sub-trees 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*

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

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.

Clearly, the greater $m(x_j)$, the more marginal (*i.e.* the less central) is x_j. Marginality constitutes a hierarchy-aware numerical mapping for nominal attributes. Note that marginality also takes sample frequencies into account: if the frequency of a value in T_X increases, the marginality of that value decreases.

In addition to representing hierarchical nominal data using a numerical mapping, we need to describe the statistical characteristics of the resulting numerical attribute. Next, we provide derivations for these statistical characteristics (see [3] for proofs of correctness).

Definition 1 (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*

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

In [3] it is shown that the above marginality-based variance is equivalent to the hierarchical variance defined in [8].

Definition 2 (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.

Definition 3 (S-distance). *The S-distance between two records $\mathbf{x}_1$ and $\mathbf{x}_2$ in a data set with d attributes is*

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

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.*

In [3] the above is shown to be a mathematical distance.

Definition 4 (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 value of the above non-negative expression is higher when the marginalities of the values taken by X *and* Y *are positively correlated.*

The key feature of marginality is that it allows converting the hierarchical nominal attribute into a single numerical attribute which preserves the hierarchical structure. We have provided definitions of statistical measures that allow us to compare the original and anonymized data. Once converted in this manner, it is easy to apply selected numerical anonymization techniques without any additional computational complexity.

Note that *not* all numerical anonymization techniques can be applied after the marginality conversion. As an example, if multiple imputation or noise addition are used, the anonymized converted nominal attributes take values in a different set as the converted original nominal attributes. When this occurs, it becomes impossible to map the anonymized numerical values back to the original nominal categories. Hence, the anonymization procedure must return values for the attributes that belong to the marginality values to which original categories were mapped. We show below that shuffling fulfills this requirement and that microaggregation can be adapted to fulfill it.

3.1 Shuffling with Marginality

For nominal data, data shuffling is typically implemented by converting the nominal attribute to binary 0-1 attributes. As discussed earlier, this results in much higher computational requirements. However, if the marginality mapping is performed, data shuffling can be implemented without any additional computational requirements. More importantly, marginality captures the hierarchical nature of the data set directly. Once represented in this manner, data shuffling has the capability to maintain the same structure in the masked data as was present in the original data. Converting marginalities back to nominal categories in the shuffled data set is feasible because shuffling does not modify the original nominal categories; hence, provided that the mapping from nominal categories to marginalities is one-to-one in the original data set (which happens with great probability), the mapping is the same in the shuffled data and can be inverted.

3.2 Hybrid Microaggregation with Marginality

In microaggregation of numerical data, the partition step seeks to form groups with high within-group similarity, that is, with low within-group variance. In order to microaggregate hierarchical nominal data, we need some measure of variance for such data. The intuitive idea behind variance in a hierarchy is that a sample of nominal values belonging to categories which are all children of the same parent category in the hierarchy 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 1 above).

Using the definitions of variance, mean and distance given earlier (Definitions 1, 2 and 3), we can easily adapt to nominal and hierarchical nominal attributes any microaggregation heuristic originally designed for numerical attributes (*e.g.* MDAV, [9]). To avoid the variance reduction caused by the aggregation step of microaggregation, in this paper we will use only the partition step. Instead of replacing the records in a group by their centroid, we will replace them by a group of records generated with the following algorithm. The result are microaggregation-based hybrid data and we call the method marginality-based *MicroHybrid*. We proposed a parallel idea to generate hybrid numerical data in [4].

Algorithm 2
For each record $\mathbf{x}_j$ *in* C_i *and for each attribute* X_l *to be synthesized:*

1. *If* X_l *is hierarchical nominal, ordinal or numerical*
   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 $\delta(\cdot,\cdot)$ *is a suitable distance (S-distance if* X_l *is hierarchical nominal, ordinal if* X_l *is ordinal and Euclidean if* X_l *is numerical),* 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 group* C_i*;*
   else (X_l *is flat nominal) randomly draw a value* x'_{jl} *with replacement from the set of values of* X_l *over* C_i*;*
2. *Replace* x_{jl} *by* x'_{jl}.

4 Experimental Results

From the entire data set, we selected the subset of records for which DIAG_P was some form of neoplasm. Among these we deleted the 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). This left us with 19502 records with the six attributes listed in Section 1 above. We will refer to this data subset as the NEOPLASM data.

We applied the marginality-based shuffling and the marginality-based *MicroHybrid* to the NEOPLASM data. As observed earlier, SRMI is incapable of

Table 1. Variation $\Delta(\Theta)$ for means, variances and covariances between the original NEOPLASM data set and the data set anonymized using multiple imputation, marginality-based data shuffling and marginality-based *MicroHybrid* with $k = 600$

Statistic	Attribute	Multiple imputation	Marg.-based shuffling	Marg.-based *MicroHybrid*
Mean	AGE_YRS	0.00042	0	0.03627
	SEX	0.00373	0	0.01349
	RACE	0.00033	0	0.01532
	LOS	0.26345	0	0.57532
	CHARGE	0.44923	0	0.91774
	DIAG_P	N/A	0	0.00210
Variance	AGE_YRS	0.00785	0	0.02612
	SEX	0.00124	0	0.00465
	RACE	0.00011	0	0.01362
	LOS	0.24603	0	7.54256
	CHARGE	0.37275	0	19.81236
	DIAG_P	N/A	0	0.00197
Covariance	(AGE_YRS, LOS)	5.56559	33.69468	162.38189
	(AGE_YRS, CHARGE)	0.12300	1.60121	11.15957
	(SEX, RACE)	0.00064	0.00059	0.00514
	(SEX, DIAG_P)	N/A	0.00006	0.00205
	(RACE, DIAG_P)	N/A	0.00005	0.00914
	(LOS, CHARGE)	0.34455	0.11783	11.58553

anonymizing the DIAG_P attribute in the NEOPLASM data. However, multiple imputation is considered to be one of the techniques offering a very high level of data utility and a low level of disclosure risk. In this sense, it represents a good benchmark to assess the effectiveness of the other two techniques. Hence, in addition to the results of using adapted shuffling and adapted *MicroHybrid*, we also report results of using SRMI on the NEOPLASM data (without DIAG_P) as a benchmark.

One key aspect of the *MicroHybrid* technique is the selection of the aggregation parameter k. In order to conduct a fair comparison, we empirically determined a value of k for which the disclosure risk in *MicroHybrid* was comparable to the disclosure risk of the SRMI-generated data and the shuffled data (note that SRMI and shuffling have no parameters, so we could only adjust k). Disclosure risk was measured by using distance-based record linkage. An anonymized record was linked to the original record whose values of the original attributes were at a shortest distance (using S-distance for nominal attributes and Euclidean distance for numerical attributes). The percentage of correct matches was taken as a measure of disclosure risk. We obtained a disclosure risk of 0.015382% for SRMI (with five attributes) and of 0.005127% for shuffling. We took $k = 600$ for *MicroHybrid*, which yielded a disclosure risk of 0.03589%.

In order to assess the utility of the anonymized data sets obtained with the three methods, for several statistics Θ, we computed the variation $\Delta(\Theta) = (\theta' - \theta)/\theta$ between the value θ of the statistic over the original data set and its

Table 2. Mean variation for means, variances and covariances between one hundred 10% original and anonymized samples of the NEOPLASM data set. Anonymized samples were created using multiple imputation, marginality-based data shuffling and marginality-based *MicroHybrid* with $k = 600$.

Statistic	Attribute	Multiple imputation	Marg.-based shuffling	Marg.-based *MicroHybrid*
Mean	AGE_YRS	0.00654	0.00757	0.03644
	SEX	0.03060	0.02450	0.03044
	RACE	0.05581	0.05649	0.05370
	LOS	0.26409	0.03139	0.57156
	CHARGE	0.45352	0.03406	0.89759
	DIAG_P	N/A	0.00573	0.00745
Variance	AGE_YRS	0.04109	0.03975	0.02969
	SEX	0.01028	0.00822	0.01047
	RACE	0.04925	0.04996	0.04747
	LOS	0.26744	0.29376	7.86135
	CHARGE	0.32469	0.40743	21.53582
	DIAG_P	N/A	0	0.00296
Covariance	(AGE_YRS, LOS)	6.11472	11.01168	159.34892
	(AGE_YRS, CHARGE)	1.50547	1.94719	20.94929
	(SEX, RACE)	0.02942	0.02984	0.02772
	(SEX, DIAG_P)	N/A	0.00546	0.00674
	(RACE, DIAG_P)	N/A	0.02924	0.02821
	(LOS, CHARGE)	0.35682	0.28189	12.32575

value θ' over the anonymized data set. Specifically, we considered the statistics mean, variance and covariance. For nominal attributes, we used the definitions of mean, variance and covariance given in Section 3. While we were able to define covariance between nominal attributes, we do not have a definition of covariance between a nominal and a numerical attribute. Hence, this information is not reported. Table 1 shows the results of using the three methods on the NEOPLASM data set.

From the results in Table 1, we can see that marginality-based shuffling performs even better than the multiple imputation SRMI in most cases. Only for the covariances between (AGE_YRS, LOS) and between (AGE_YRS, CHARGE) does multiple imputation outperform data shuffling. And, as observed earlier, data shuffling also results in lower disclosure risk than multiple imputation. In summary, for the NEOPLASM data set, data shuffling offers both high utility and low disclosure risk.

The performance of *MicroHybrid* is somewhat worse than both data shuffling and multiple imputation. However, *MicroHybrid* at least offers the ability to anonymize the data, while multiple imputation does not. Hence, in the presence of hierarchical nominal data, *MicroHybrid* would be preferred over multiple imputation.

In addition to evaluating the utility over the entire data set, we evaluated the utility loss when the user restricts her analysis to a random subdomain of the

data set. Thus, for each method, we took 10% samples of the anonymized data set 100 times. We then computed the mean variation $\bar{\Delta}(\Theta)$ over the 100 samples and the corresponding samples of the original data set for the same statistics as above. The results are shown in Table 2. None of the three methods wins clearly for all statistics: shuffling no longer exactly preserves means and variances for subdomains and, for some statistics, it is slightly outperformed by one of the other two methods. However, SRMI is the clear loser to its inability to deal with DIAG_P and shuffling seems to behave somewhat better than *MicroHybrid* (the latter has quite large mean variations for four statistics).

In fact, the smaller k w.r.t. the sample size, the better *MicroHybrid* preserves statistics on random samples, because the generation of hybrid data is more constrained to being similar to the original data (in [4] we showed this for numerical data). However, we were forced to take $k = 600$ in order to match the low disclosure risk of data shuffling and multiple imputation, and $k = 600$ is *not* much smaller than the size of a 10% sample (1950 records).

5 Conclusions and Future Research

Hierarchical attributes are common in many data sets (for example biomedical or economic data), and they are often among the most important attributes (for example, "Diagnostic" or "Economic activity"). When used for secondary purposes, data containing hierarchical attributes must be anonymized, but, unfortunately, most existing anonymization techniques cannot be used. In this study, we have addressed this issue.

We have described a hierarchy-aware numerical mapping for hierarchical attributes, called marginality. With this mapping, any numerical anonymization procedure not perturbing original values (*i.e.* keeping the mapping reversible) can be employed to anonymize the data without any additional computational complexity. We have illustrated the application of this approach using data shuffling and microaggregation. The performance of these techniques was evaluated using the NEOPLASM data. Our results indicate that both techniques perform well.

Future research lines include:

- Extending the marginality-based mapping for anonymization techniques perturbing the input marginalities. One could think of an approximate reverse mapping for methods which do perturb input marginalities; that is, each numerical output marginality m could be mapped back to the hierarchical category having marginality closest to m. However, approximate reverse mapping can lead to gross distortion if there are categories very distant within the hierarchy that have similar marginalities, because they could be unduly swapped. Hence, blocking strategies or other mechanisms should be devised to avoid such undesirable effects.
- Using semantic metrics developed in artificial intelligence (*e.g.* [23]) as an alternative to assess the semantic distortion incurred by marginality-based anonymization.

Disclaimer and Acknowledgments. The first and third authors are with the UNESCO Chair in Data Privacy, but the views expressed in this paper do not necessarily reflect the position of UNESCO nor commit that organization. Most of the work on this paper by the second author was completed when he was a Visiting Professor at the Department of Computer Engineering and Maths, Universitat Rovira i Virgili. 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 project "DwB". The first author is partially supported as an ICREA Acadèmia researcher.

References

1. Defays, D., Nanopoulos, P.: Panels of enterprises and confidentiality: the small aggregates method. In: Proc. of 92 Symposium on Design and Analysis of Longitudinal Surveys, pp. 195–204. Ottawa, Statistics Canada (1993)
2. Domingo-Ferrer, J.: A survey of inference control methods for privacy-preserving data mining. In: Aggarwal, C.C., Yu, P. (eds.) Privacy-Preserving Data Mining: Models and Algorithms, pp. 53–80. Springer, New York (2008)
3. Domingo-Ferrer, J.: 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 (to appear, 2012); Preliminary version available from `http://arxiv.org/abs/1202.6009` (since February 27, 2012)
4. Domingo-Ferrer, J., González-Nicolás, Ú.: Hybrid data using microaggregation. Information Sciences 180(15), 2834–2844 (2010)
5. Domingo-Ferrer, J., Martínez-Ballesté, A., Mateo-Sanz, J.M., Sebé, F.: Efficient multivariate data-oriented microaggregation. VLDB Journal 15(4), 355–369 (2006)
6. 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)
7. Domingo-Ferrer, J., Sebé, F., Solanas, A.: A polynomial-time approximation to optimal multivariate microaggregation. Computers & Mathematics with Applications 55(4), 714–732 (2008)
8. Domingo-Ferrer, J., Solanas, A.: A measure of nominal variance for hierarchical nominal attributes. Information Sciences 178(24), 4644–4655 (2008); Erratum: Information Sciences 179(20), 3732 (2009)
9. Domingo-Ferrer, J., Torra, V.: Ordinal, continuous and heterogenerous k-anonymity through microaggregation. Data Mining and Knowledge Discovery 11(2), 195–212 (2005)
10. Duncan, G.T., Elliot, M., Salazar-González, J.J.: Statistical Confidentiality: Principles and Practice. Springer, New York (2011)
11. Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Lenz, R., Longhurst, J., Schulte-Nordholt, E., Seri, G., De Wolf, P.P.: Handbook on Statistical Disclosure Control (version 1.2). ESSNET SDC Project (2010), `http://neon.vb.cbs.nl/casc`
12. 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)

13. International Classification of Diseases, 9th Revision, Clinical Modification, 6th edn. (2008), http://icd9cm.chrisendres.com/
14. IVEware: Imputation and Variance Estimation Software, v. 0.2. University of Michigan (2010), http://www.isr.umich.edu/src/smp/ive/
15. Lenz, R.: Methoden der Geheimhaltung wirtschaftsstatistischer Einzeldaten und ihre Schutzwirkung. Statistisches Bundesamt, Wiesbaden (2010)
16. Muralidhar, K., Sarathy, R.: Data shuffling - a new masking approach for numerical data. Management Science 52(5), 658–670 (2006)
17. 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
18. Oganian, A., Karr, A.F.: Combinations of SDC Methods for Microdata Protection. In: Domingo-Ferrer, J., Franconi, L. (eds.) PSD 2006. LNCS, vol. 4302, pp. 102–113. Springer, Heidelberg (2006)
19. Patient Discharge Data, Office of Statewide Health Planning & Development-OSHPD (2010), http://www.oshpd.ca.gov/HID/Products/PatDischargeData/PublicDataSet/index.html
20. Raghunathan, T.E., Lepkowski, J.M., Van Hoewyk, J., Solenberger, P.: A multivariate technique for multiply imputing missing values using a series of regression models. Survey Methodology 27, 85–96 (2001)
21. Reiter, J.P.: Releasing multiply-imputed, synthetic public-use microdata: an illustration and empirical study. Journal of the Royal Statistical Society A 168(1), 185–205 (2005)
22. Rubin, D.B.: Discussion of statistical disclosure limitation. Journal of Official Statistics 9(2), 461–468 (1993)
23. 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)
24. Torra, V.: Microaggregation for Categorical Variables: A Median Based Approach. In: Domingo-Ferrer, J., Torra, V. (eds.) PSD 2004. LNCS, vol. 3050, pp. 162–174. Springer, Heidelberg (2004)
25. Willenborg, L., De Waal, T.: Elements of Statistical Disclosure Control. Springer, New York (2001)