Towards k-anonymous non-numerical data via Semantic Resampling

Sergio Martínez, David Sánchez, Aida Valls

Department of Computer Science and Mathematics. Universitat Rovira i Virgili
Av. Països Catalans, 26. 43007. Tarragona, Catalonia (Spain)
{sergio.martinezl, david.sanchez, aida.valls }@urv.cat

Abstract. Privacy should be carefully considered during the publication of data (e.g. database records) collected from individuals to avoid disclosing identities or revealing confidential information. Anonymisation methods aim at achieving a certain degree of privacy by performing transformations over non-anonymous data while minimising, as much as possible, the distortion (i.e. information loss) derived from these transformations. *k*-anonymity is a property typically considered when masking data, stating that each record (corresponding to an individual) is indistinguishable from at least *k-1* other records in the anonymised dataset. Many methods have been developed to anonymise data, but most of them are focused solely on numerical attributes. Non-numerical values (e.g. categorical attributes like job or country-of-birth or unbounded textual ones like user preferences) are more challenging because arithmetic operations cannot be applied. To properly manage and interpret this kind of data, it is required to have operators that are able to deal with data semantics. In this paper, we propose an anonymisation method based on a classic data re-sampling algorithm that guarantees the fulfilment of the *k*-anonymity property and is able to deal with non-numerical data from a semantic perspective. Our method has been applied to anonymise the well-known Adult Census dataset, showing that a semantic interpretation of non-numerical values better minimises the information loss of the masked data file.

Keywords: Privacy protection, anonymity, resampling, semantics.

1 Introduction

Privacy protection in published data (e.g. polls, usage logs, questionnaires, etc.) is crucial to avoid revealing individual's identities and/or linking those to confidential attributes (e.g. salary, medical conditions, etc.). Several methods have been proposed in the literature aiming at generating an anonymised version of the original data so that it can be published with low disclosure risk. These methods mask the set of records in the original dataset by introducing some distortion on the original values, making difficult the re-identification of individuals while, at the same time, trying to retain the utility of data (i.e. the fact that equal or similar conclusions can be extracted

from both the original and anonymised version of a dataset). To generate an anonymised version of a dataset, the *k*-anonymity property is commonly considered. A masked dataset is considered *k*-anonymous if each record is indistinguishable from, at least, *k*-1 other records [1]. To retain data utility, authors select the data transformation that minimizes the information loss.

Several techniques introducing distortions to mask data can be found, such as: Additive Noise, Microaggregation, Rank Swapping or Resampling [2]. These have been widely used for statistical disclosure control on numerical data [2].

In this paper we focus on Resampling [3], a method that, even though has not gained as much research attention as other algorithms like Microaggregation [4], it has demonstrated to retain a high utility of data with respect to other methods [5, 6]. In a nutshell, Resampling is based on taking *t* independent samples from the original attributes, sorting them and finally building a new masked record as the average of the ranked values of the *t* samples. Even though this method reduces the disclosure risk, it does not ensure the *k*-anonymity property on the masked dataset.

The application of this method to numerical attributes (e.g. salaries, outcomes) is straightforward by using some ordering criterion on numbers and then calculating the arithmetic average of the ranked values of the different samples. However, nowadays, thanks to the growth of the Information Society, non-numerical data (e.g. categorical attributes such as country-of-birth or even textual values like user preferences, medical visit outcomes) are commonly collected and published. The management of this kind of data for masking purposes (i.e. resampling) is challenging due to the lack of appropriate aggregation and sorting operators. A proper interpretation of non-numerical data necessarily requires a semantic interpretation of their meaning (e.g. sports like *swimming* or *sailing* are more similar than *swimming* and *trekking*). In fact, as stated by several authors [7, 8], the preservation of the semantics of non-numerical data during the masking process is crucial to maintain their utility. Because semantics is an inherently human feature that can be hardly modelled in a mathematical way, semantically-grounded masking methods have to deal with a degree of uncertainty to assess the meaning of the textual values.

This paper focuses on this kind of data, presenting a resampling method that is able to deal with non-numerical attributes from a semantic perspective, proposing appropriate sorting and averaging operators. These will exploit structured knowledge bases (i.e. taxonomies or more general ontologies) and the theory of semantic similarity [9] as the basis to semantically interpret data. As far as we know, no other semantically-grounded masking resampling methods have been proposed in the literature. Moreover, to guarantee a theoretical level of privacy, the resampling method will be designed so that it fulfils the well-known *k*-anonymity property (on the contrary to classic resampling methods [5, 10]).

The rest of the paper is organized as follows. Section 2 introduces preliminary concepts, discussing classical resampling methods and reviewing the principles that enable a semantic interpretation of data. In section 3, our semantically-grounded *k*-anonymous resampling method is presented. Section 4 is devoted to evaluate our method proposed. The final section contains the conclusions and future work.

2 Preliminaries

In this section we introduce the main elements that constitute the basis of our proposal. First we present the original resampling method. Then, we discuss how semantics modelled in knowledge bases can be exploited to interpret and compare terms.

2.1 The resampling method

Resampling was defined in [11] as a method for perturbation of output data that consists on modifying only the answers to some queries while leaving the original data unchanged. Later, in Heer [3] resampling was applied to anonymise contingency tables based on bootstrapping. A cell-oriented version resampling procedure was proposed later in [10]. However, in this case the masked data file generated does not rely on resampling the original data, but on generating binomial random perturbations that preserve some statistics (obtaining an asymptotically unbiased estimate of the original data). Resampling has been also applied to anonymise input data (i.e. original records of the published dataset). In [5] a distributed version of the resampling method is proposed for dealing with the case where the original data is partitioned among several providers. All these methods focus on numerical data.

In this paper, as done in related works [5, 10], we base our method in the Heer's resampling approach (detailed in Algorithm 1). Briefly, being *n* the number of records in the dataset, the method take *t* samples with replacement (i.e. values can be taken more than once). Each sample is sorted in increasing order. Then, the masked dataset is obtained by taking, as first value, the average of the first values of all samples, as second value, the average of the second values of all samples, and so on.

Algorithm 1. Resampling
Inputs: D (dataset), t (number of samples)
Output: D^A (a masked version of D)
1 Take t samples $S_1,\ldots,S_t$ of n records of D (with replacement)
2 Sort each sample S_i in $S_1,\ldots,S_t$
3 Make a set P_i with t elements with the records at the i-th position of the sorted samples
4 Add the average of each P_i to D^A
5 Output D^A

Comparison studies [5, 6] show that Heer's resampling achieves a high utility score with respect to other masking techniques, but with a slightly higher disclosure risk. This is related to the fact that, unlike other masking methods [4, 12, 13], the Heer's approach was designed without considering the *k*-anonymity property (formalized years later in [1]). Hence, resampled results cannot guarantee a –theoretical- level of privacy.

Our method will tackle this issue proposing a resampling method that fulfils the *k*-anonymity property while it is also able to deal with non-numerical data from a semantic perspective. Two issues that were not considered in previous works [5, 10].

2.2 Ontology-based semantic similarity

When dealing with non-numerical data, the preservation of semantics plays a crucial role to ensure the utility of the anonymised datasets [7, 8, 14, 15]. Hence, considering the semantics to compare two concepts (in the sorting or averaging processes) may improve the utility of the masked data file. To make a semantic interpretation of textual values, algorithms require some sort of structured knowledge source that represents the relations between words at a conceptual level, such as taxonomies or ontologies. Ontologies are formal and machine-readable structures of shared conceptualisations of knowledge domains expressed by means of semantic relationships. These allow mapping words to concepts, so that terms can be interpreted according to their semantic relations. As a result, semantic similarity measures can be defined [9].

In previous works on data privacy for non-numerical data [12], the distance between records was measured with the Wu & Palmer similarity measure [9] using WordNet [16] as ontology. The Wu & Palmer measure, which will also applied in in this work, evaluates the distance taking as reference the Least Common Subsumer (LCS, i.e. the most specific ancestor that generalizes the two concepts). It evaluates the number of taxonomical links from x_1 to the LCS (N1) and from x_2 to the LCS (N2), as well as the number of taxonomical links from the LCS to the root of the ontology (N_3), as shown in (Eq.1). The result ranges from 1 (for identical concepts) to 0.

$$dis_{w\&p}(x_1, x_2) = 1 - \frac{2 \times N_3}{N_1 + N_2 + 2 \times N_3} \quad (1)$$

Several ontologies are being developed in many different domains, and specific search engines may help to find the appropriate ontology for each particular dataset [17]. In addition, general-purpose ontologies are also available, such as the Wordnet thesaurus [16], which includes more than 100,000 English terms.

3 Semantic *k*-anonymous resampling for non-numerical data

This section proposes a new resampling method (named S*k*Resampling) focused on minimizing the information loss when masking non-numerical data while ensuring the fulfilment of the k-anonymity property. It is based on the Heer's resampling method [3] with the following modifications:

- *K-anonymous resampling*: the original sampling method has been modified (as detailed in section 3.1) so that masked records fulfil the k-anonymity property.
- *Semantic resampling of non-numerical data*: in order to semantically interpret non-numerical data during the resampling process, new semantically-grounded ontology-based *comparison*, *sorting* and *averaging* operators are proposed in section 3.2.

3.1 *k*-anonymous resampling

To guarantee the k-anonymity on the masked dataset we modified Heer's method (detailed in Algorithm 1) as detailed in Algorithm 2. Let *D* be the input dataset, *k* is the desired level of *k*-anonymity and *n* is the number of records in *D*.

Algorithm 2. *k*-anonymous Resampling

Inputs: D (dataset), k (level of anonymity)
Output: D^A (a new version of D that fulfils k-anonymity)

```
1  Take k samples S_1,…,S_k of n/k records of D (no replacement)
2  Take the (n mod k) remaining records as sample S_r
3  Sort each sample S_i in S_1,…,S_k
4  Make a set P_i with k elements with the records at the i-th
   position of the sorted samples
5  Generate the average of each P_i denoted as c_i.
6  For each record r_t in S_r do
7    Add r_t to the set P_i with the closest average c_i
8    Recalculate the average c_i of the modified set
9  For each P_i do
10   Replace each record in P_i by the average c_i
11   Add the new version of the record to D^A
12 Output D^A
```

The algorithm proceeds as follows. First, to guarantee the *k*-anonymity property, it is required to create *k* samples, so that *k* records can be replaced by their average in a later stage and become *k*-indistinguishable. To do so, *k* samples of *n*/*k* records are created (line 1), without replacement (i.e. each record is taken only once). It is important to avoid record repetitions to ensure that all records fulfil the k-anonymity. Then these samples are sorted (line 3) and sets P_i are created with the i^{th} records of all samples (line 4). Then, the masked dataset D^A is created by replacing, for each P_i, all records in P_i by the average c_i (lines 10-11). Because by definition P_i contains *k* records without repetition, this replacement generates a set that fulfils the *k*-anonymity.

Note that, when taking *n/k* records for sample, the remaining *n mod k* records (S_r) should be also treated. These records are added to the set with the closest average (line 7), recalculating its average (line 8) before the replacement (lines 9 and 10).

3.2 Semantic operators for non-numerical data

To consider the semantics of non-numerical data, semantically-grounded *comparison*, *sorting* and *averaging* operators for computing with words are needed.

As introduced in section 2, to semantically compare term pairs, an ontology-based measure like the Wu and Palmer distance (Eq 1) can be used. In Algorithm 2, this measure is used to discover the closest records (line 7), to sort sets of records (line 3) and to calculate averages (lines 5 and 8), as detailed below. Note that to apply this distance to multi-attributed records ($\{x_{11},\ldots,x_{1m}\}$, $\{x_{21},\ldots x_{2m}\}$, where x_{ij} is the j^{th} attribute value of the i^{th} record and m is the number of attributes), we can calculate the average of the distances between the values attribute by attribute, as follows:

$$semantic_distance\left(\{x_{11},...,x_{1m}\},\{x_{21},...,x_{2m}\}\right)=\frac{\sum_{j=1}^{m} dis_{w\&p}\left(x_{1j},x_{2j}\right)}{m} \quad (2)$$

When dealing with non-numerical attributes, the definition of a sorting operator (line 3) is not straightforward. To sort a set of elements, it is needed to have an initial point from which starting sorting according to the similarity of the rest of values to

that initial value. Numerically, this is done by taking the max (or min) value in the set. From a semantic perspective we follow a similar principle as propose Algorithm 3:

Algorithm 3. Sorting procedure

Inputs: P (dataset)
Output: P' (P sorted)
1 c = average(P)
2 f = argMax_{v_i}(semantic_distance(c,v_i)) $\forall v_i \in P$
3 Add f to P' and remove it from P
4 while (|P| > 0) do
5 r = argMin_{v_i}(semantic_distance(f,v_i)) $\forall v_i \in P$
6 Add r to P' and remove it from P
7 f = r
8 Output P'

Similarly to numbers, we begin by taking the initial point from which start sorting as the farthest record to the average of the dataset (lines 1-2). By using the semantic distance (Eq. 2) we sort the rest of the records by comparison (lines 2-5).

The sorting procedure and also the resampling method rely on the calculation of averages of non-numerical values to sort and replace values. Hence, the selection of an accurate average is crucial to minimise the information loss resulting from the masking process. Intuitively, the average of a dataset can be seen as the "centre". Based on this, we propose calculating the average as the semantic centroid of the dataset, exploiting the knowledge structure provided by a background ontology.

The usual approach to centroid construction for non-numerical data is based on the data distribution, mainly using the mode operator [4, 13]. This approach omits the semantics of data and, hence, a crucial dimension of data utility. Moreover, centroids are constrained to the values appearing in the input dataset. Otherwise, semantic centroids are constructed using the concept in the ontology that subsumes all the values in the cluster (i.e. the LCS [12]). This approach is affected by outliers and thus, it commonly results too abstract generalisations, causing a high information loss.

In our centroid construction method, a background ontology is exploited not only to assess the semantic distance between terms, but also to retrieve the centroid candidates. To calculate the centroid, we assume the classical centroid definition as the record that minimises the distance (in our case, *semantic distance*) against all the elements in a set. Formally, given a the semantic distance of Eq. 2, the centroid of a set of records $\{r_1, r_2, \ldots, r_n\}$ is defined as:

$$centroid\left(r_1, r_2, ..., r_n\right) = \arg\min_c \left(\sum_{i=1}^{n} semantic_distance\left(c, r_i\right) \right) \quad (3)$$

where c is a centroid candidate for the set of arguments.

The search space of centroid candidates c has a great importance. When selecting the centroid according to the frequency of values (i.e. mode) the number of centroid candidates is limited to the set of different values that appear in the cluster. On the contrary, using a background ontology such as WordNet, we potentially extend the search space to all concepts modelled in the ontology and, hence, we can construct the centroid from a finer grained set of candidates. For semantic coherency and efficien-

cy, we limit the search space to the hierarchical trees to which input values belong, taking as centroid candidates the set of taxonomical ancestors.

Formally, let us take $V=\{v_1,\ldots,v_n\}$ as the set of possible values for a non-numerical attribute A_i. Let us take ontology O containing and semantically modelling all v_i in V. The first step of our method consists of mapping the values in V to concepts in O, so that semantically related concepts can be extracted from O following the semantic relationships. We assume that taxonomical subsumers of a term (including itself) are valid centroid candidates. The set of candidates is then given in the *minimum subsumer hierarchy* ($H_O(V)$) that goes from the concepts corresponding to the values in V to the Least Common Subsummer of all these values. The *Least Common Subsumer* (LCS) of a set of categorical values V ($LCS_O(V)$) is the deepest taxonomical ancestor that all the terms in V have in common in the ontology O. We omitted taxonomical ancestors of the LCS because those will always be more general (i.e. more semantically distant) than the LCS and, hence, worse centroid candidates. Then, all the concepts c in $H_O(V)$ are the centroid candidates of the set V.

The centroid of a set of values V in ontology O is defined as the concept c_j in $H_O(V)$ that minimises the semantic distance $dis_{w\&p}$ with respect to all the values in V.

$$centroid_O(V) = \left\{ \arg Min\left(\sum_1^n dis_{w\&p}(c_j, v_i) \right) \right\}, \forall c_j \in H_O(V), \forall v_i in V \quad (4)$$

When more than one candidate minimises the distance, all of them would be equally representative, so any of them can be selected as the final centroid. When dealing with multi-attributed data, we assume that attributes are independent, so that, the centroid of each attribute can be calculated separately, and the final centroid is the tuple of individual ones. This assumption is usual for numerical data, where the average of attributes is calculated independently.

Hence, the centroid of a multivariate dataset MV in the ontology O is defined as:

$$centroid_O(MV) = \{ centroid_O(A_1), centroid_O(A_2), \ldots, centroid_O(A_m) \} \quad (5)$$

, where A_j is the set of values for the jth attribute in MV and m is the number of attributes in MV.

4 Evaluation

In this section, the evaluation of the proposed S*k*Resampling method is detailed. As evaluation data, we used the well-known *Adult Census* dataset [18]. Two non-numerical attributes have been taken, corresponding to "occupation" (14 distinct modalities) and "native-country" (41 modalities). This result in 30,162 records (after removing rows with missing values) distributed into 388 different responses, 83 of them being unique. Attribute values have been mapped to WordNet [16], which has been used as the knowledge base that enables the semantic interpretation of data.

Because the theoretical level of privacy is guaranteed by the fulfilment of the *k*-anonymity property, the evaluation has been focused on the quantification of the information loss. To do so, we have employed the *sum of square errors* (SSE), which has been extensively used to evaluate anonymisation methods [4, 12]. It is defined as

the sum square of distances between each original record and their corresponding masked version (Eq. 6). Hence, the higher the SSE is, the higher the information loss.

$$SSE = \sum_{i=1}^{n} (semantic_distance(x_i, x^{A}{}_i))^2 \qquad (6)$$

,where n is the number of records in the dataset, x_i is the i^{th} record and $x^{A}{}_i$ denotes the masked version of x_i. As a distance measure, we employ again the *semantic_distance* defined in Eq. 2 to quantify the loss in semantics.

To evaluate the benefits that the semantic S*k*Resampling approach brings with regards to the preservation of the data utility in comparison with non-semantic methods, we configured four settings that, while based on the same resampling algorithm (Alg. 2), they vary the criteria used to compare, sort and replace the original values:

- Setting 1: no semantics are considered when masking the dataset. On one hand, values are compared using the equality predicate (0 if identical, 1 otherwise). On the other hand, the centroid is selected as the mode (i.e., the most frequently occurring value in the set). This mimics the behavior of anonymising approaches dealing with non-numerical data in a categorical fashion [4, 13].
- Setting 2: semantics are considered when comparing values. The Wu & Palmer measure (Eq. 2) applied to WordNet as knowledge structure are used. The centroid is kept the same as in setting 1. Comparing this to setting 1, one can quantify the influence in information loss of the distance calculation method.
- Setting 3: the same as setting 2, but taking as centroid the concept in WordNet that taxonomically subsumes all elements in the set (i.e. the Least Common Subsumer, as done in [12]). This approach considers the semantics of data in the centroid calculation but neglect their distribution in the taxonomical tree.
- Setting 4 (S*k*Resampling): the method presented in section 3 is applied. It shows the contribution of our proposal on the information loss.

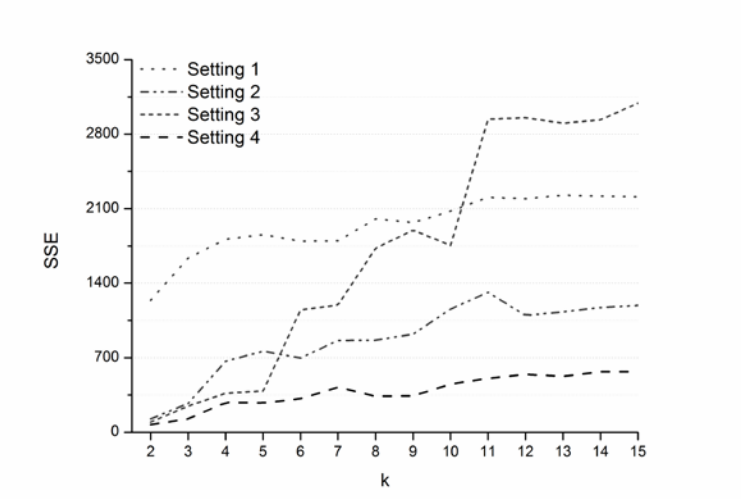


Fig. 1. Comparison of Information Loss (SSE) values for the evaluated settings.

The information loss values (i.e. *SEE*) obtained for the four settings while varying the *k*-anonymity level from 2 to 15 are shown in Fig. 1.

First of all we can see that our approach (*setting 4*) achieves the lowest information loss for all the *k*-anonymity values. Compared to other settings, we observe that a non-semantic method (*setting 1*) results in the highest information loss for most *k*-values. This shows that managing and transforming non-numerical data only by means of their distribution features (i.e. repetitions, mode) hardly preserves the meaning of masked data and hence their utility. By considering semantics in distance calculations (*setting 2*), information loss is significantly reduced showing the limitations of comparing attribute values according to the equality/inequality of their textual labels. Regarding the results obtained by *setting 3* (i.e. centroid computed as the LCS), we observe that, even though it is able to retain more information than non-sematic centroid calculation methods (i.e. mode) for low *k*-values, it produces severely worse results when a higher anonymity degree is required. The fact that the LCS is used as the centroid causes that very general abstraction are picked when a larger and more heterogeneous set of values should be generalized (i.e. higher *k*-anonymity). In comparison, our method (*setting 4*) produces more accurate centroids, by considering both the semantics and the data distribution.

5 Conclusions

In recent years, the interest in data semantics has been grown in the privacy-preserving community due to the increasing availability of non-numerical data [8, 12]. Several authors have shown the benefits that a semantic treatment of data can bring with regards to the preservation of the utility of masked data [7, 8].

In this paper, we have adapted a classic resampling method to always fulfil the *k*-anonymity property, in order to guarantee a certain level of privacy. While considering the semantics of the data by means of the definition of semantically-grounded comparison, sorting and averaging operators, we obtain an effective method for anonymising non-numerical data. The evaluation results sustain our hypothesis, being able to minimise the information loss in comparison with non-semantic approaches.

As future work, the effect of different semantic similarity measures on the quality of the masked data should be studied. A deeper analysis of the performance of the anonymised data over other types of security attacks (*e.g.* re-identification by record linkage) could be also interesting. Finally, we plan to compare our S*k*Resampling method with other anonymisation approaches that also consider data semantics up to a degree, like those based on generalizing attribute values [1], substituting values with similar ones [8] or aggregating semantically similar records [12].

Acknowledgements

This work was partly funded by the Spanish Government through the projects CONSOLIDER INGENIO 2010 CSD2007-0004 “ARES” and eAEGIS TSI2007-65406-C03-02, and by the Government of Catalonia under grant 2009 SGR 1135 and

2009 SGR-01523. Sergio Martínez Lluís is supported by a research grant of the Ministerio de Educación y Ciencia (Spain)

References

1. Samarati, P., Sweeney, L.: Protecting Privacy when Disclosing Information: k-Anonymity and Its Enforcement through Generalization and Suppression. Technical Report SRI-CSL-98-04, SRI Computer Science Laboratory (1998)
2. Domingo-Ferrer, J.: A Survey of Inference Control Methods for Privacy-Preserving Data Mining. In: Aggarwal, C.C., Yu, P.S. (eds.) Privacy-Preserving Data Mining, vol. 34, pp. 53-80. Springer US (2008)
3. Heer, G.R.: A bootstrap procedure to preserve statistical confidentiality in contingency tables. In: Int. Seminar on Statistical Confidentiality, pp. 261-271. Eurostat, (1993)
4. Domingo-Ferrer, J., Mateo-Sanz, J.M.: Practical Data-Oriented Microaggregation for Statistical Disclosure Control. IEEE Trans. on Knowl. and Data Eng. 14, 189-201 (2002)
5. Herranz, J., Nin, J., Torra, V.: Distributed Privacy-Preserving Methods for Statistical Disclosure Control Data Privacy Management and Autonomous Spontaneous Security. 5939, 33-47 (2010)
6. Karr, A.F., Kohnen, C.N., Oganian, A., Reiter, J.P., Sanil, A.P.: A Framework for Evaluating the Utility of Data Altered to Protect Confidentiality. The American Statistician 60, 224-232 (2006)
7. Torra, V.: Towards knowledge intensive data privacy. Proceedings of the 5th international Workshop on data privacy management, and 3rd international conference on Autonomous spontaneous security, pp. 1-7. Springer-Verlag, Athens, Greece (2011)
8. Martinez, S., Sanchez, D., Valls, A., Batet, M.: Privacy protection of textual attributes through a semantic-based masking method. Inf. Fusion DOI: 10.1016/j.inffus.2011.03.004, (2011)
9. Wu, Z., Palmer, M.: Verbs semantics and lexical selection. the 32nd annual meeting on Association for Computational Linguistics, pp. 133-138. Association for Computational Linguistics, Las Cruces, New Mexico (1994)
10. Domingo-Ferrer, J., Mateo-Sanz, J.M.: Resampling for statistical confidentiality in contingency tables. Computers & Mathematics with Applications 38, 13-32 (1999)
11. Jones, D.H., Adam, N.R.: Disclosure avoidance using the bootstrap and other resampling schemes. In: Proceedings of the Fifth Annual Research Conference, U.S. Bureau of the Census, pp. 446-455. (1989)
12. Abril, D., Navarro-Arribas, G., Torra, V.: Towards semantic microaggregation of categorical data for confidential documents. Proc. of the 7th int. conference on Modeling decisions for artificial intelligence, pp. 266-276. Springer-Verlag, Perpignan, France (2010)
13. Torra, V.: Microaggregation for Categorical Variables: A Median Based Approach. In: Domingo-Ferrer, J., Torra, V. (eds.) Privacy in Statistical Databases, vol. 3050, pp. 518-518. Springer Berlin / Heidelberg (2004)
14. Martinez, S., Sanchez, D., Valls, A.: Ontology-Based Anonymization of Categorical Values. In: Torra, V., Narukawa, Y., Daumas, M. (eds.) Modeling Decisions for Artificial Intelligence, vol. 6408, pp. 243-254. Springer Berlin / Heidelberg (2010)
15. Martínez, S., Sánchez, D., Valls, A., Batet, M.: The Role of Ontologies in the Anonymization of Textual Variables. Proceeding of the 2010 conference on Artificial Intelligence Research and Development: Proceedings of the 13th International Conference of the Catalan Association for Artificial Intelligence, pp. 153-162. IOS Press (2010)
16. Fellbaum, C.: WordNet: An Electronic Lexical Database (Language, Speech, and Communication). The MIT Press (1998)

17. Ding, L., Finin, T., Joshi, A., Pan, R., Cost, R.S., Peng, Y., Reddivari, P., Doshi, V., Sachs, J.: Swoogle: a search and metadata engine for the semantic web. Proceedings of the thirteenth ACM international conference on Information and knowledge management, pp. 652-659. ACM, Washington, D.C., USA (2004)
18. Hettich, S., Bay, S.D.: The UCI KDD Archive. (1999)