Evaluation of the disclosure risk of masking methods dealing with textual attributes

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

Received March 2011; revised July 2011

ABSTRACT. *Record linkage methods evaluate the disclosure risk of revealing confidential information in anonymized datasets that are publicly distributed. Concretely, they measure the capacity of an intruder to link records in the original dataset with those in the masked one. In the past, masking and record linkage methods have been developed focused on numerical or ordinal data. Recently, motivated by the proliferation of textual information, some authors have proposed masking methods to anonymize textual data. Textual attributes should be interpreted according to their semantics, which makes them more difficult to manage and compare than numerical data. In this paper, we propose a new record linkage method specially tailored to accurately evaluate their disclosure risk. Our method, named Semantic Record Linkage, relies on the theory of semantic similarity and uses widely available ontologies to interpret the semantics of data and propose coherent record linkages. Test performed over a real dataset show that a semantic record linkage method evaluates better the disclosure risk when compared to a non-semantic approach.*
Keywords: Privacy protection, disclosure risk, record linkage, ontologies, semantic similarity

1. **Introduction.** Statistical agencies gather data and make them available to third parties for analysis. Datasets consist on a set of records (corresponding to individuals) described by a set of attributes (corresponding to the features, such as age, job or religion). Due to the fact that microdata may contain sensible information about individuals, they must be anonymised before making them public to guarantee the privacy of the respondents. The goal of a privacy-preserving method is avoiding an intruder re-identifies the identity of an individual from the published data, associating his confidential information. A typical way to achieve some degree of anonymity is to satisfy the *k*-anonymity property [1]. This establishes that each record in a dataset must be indistinguishable with at least *k-1* other records of the same dataset. To satisfy the *k*-anonymity property, some masking methods have been designed, most of them being specific for numerical data [2]. However, in many situations, some kind of information can only be expressed by means of textual labels (*e.g.* job or user preferences regarding leisure or shopping). With the enormous growth of the

Information Society, datasets containing textual information are becoming easily available (*e.g.* user opinions, preferences, reviews and even query logs). On the contrary to numbers, the processing of textual data cannot be made by means of the arithmetic operators [3] and require semantic analysis tools to understand their meaning. Considering that words correspond to concepts with a semantic content, knowledge sources (*e.g*. structured thesaurus, ontologies, tagged corpora, etc.) are needed to interpret their semantics.

In recent years, some authors (see section 2) have proposed masking methods dealing with textual data from a semantic point of view. These works rely on predefined knowledge structures (*i.e*. taxonomies) which are analyzed to propose transformations of textual data by means of *generalization*. Terms are substituted by other more general ones that taxonomically subsume them. As a result of this data transformation process, an information loss occurs. *Information loss* is a quality measure of the reduction of the utility in the masked data, when compared to the original one [4]. In the case of textual attributes, information loss should be considered as a function of reduction of semantic content (*i.e*. the more abstract the generalizations, the higher the information loss) [1, 5-7]. Ideally, any masking method should minimize the information loss to maximize data utility [8].

Another important aspect of any masking method is the minimization of the *disclosure risk*. It measures the capacity of an intruder to obtain the information contained in the original dataset from the masked one [4]. To compute the disclosure risk, many works [4, 9, 10] consider *record linkage* (RL) methods. These try to link the records in the original dataset with those of in the masked one. Two kinds of record linkage methods are usually considered in the literature[11]. On the one hand, *distance-based record linkage* computes a distance measure between original and masked records, linking each masked one to the closest in the original dataset. For numerical data, an Euclidean distance is typically used. On the other hand, *probabilistic record linkage* bases the matching on the expectation-maximization algorithm [12] which is based on the amount of coincidences between masked and original datasets.

Classical RL methods have been defined independently of the masking method used to anonymise input data. However, some works [11, 13-15] have shown that it is possible to increase the amount of linkages by designing tailored RL methods for concrete masking schemas. In [9, 11, 13] authors show that ad-hoc designed RL methods increase the disclosure risk when assuming that input data have been anonymised by means of a micro-aggregation process. In [11] a similar work is proposed, in which an especially designed RL method increases the amount of linkages when input data is masked by with rank swapping [16]. Using especially tailored RL methods one assumes the worst possible scenario for privacy protection and, hence, better evaluates the potential disclosure risk.

Both generic and ad-hoc RL methods proposed in the literature are focused only on numerical and ordinal data. However, as stated above, several masking methods for textual attributes have been proposed in recent years. Due to textual attributes should be interpreted according to their semantics it is not straightforward to apply existing RL methods. As far as we know, no semantically-grounded RL methods have been proposed.

In this paper, we present a new distance-based RL method designed to measure the

disclosure risk of masking methods based on the generalization of textual attributes. Our method (called *Semantic Record Linkage*, SRL) relies on the theory of semantic similarity to propose linkages between original and masked datasets, discovering the most semantically similar records. This supposes an improvement over methods based solely on the number of term coincidences. Considering that the knowledge structure used by the masking method to propose generalizations remains hidden to the intruder, we propose exploiting general-purpose taxonomies/ontologies to better interpret textual values [17, 18]. Our method has been applied to evaluate the disclosure risk of a classical generalization method applied to a real dataset of textual data, and it has been compared against a non-semantic RL approach relying on counting term coincidences. Results show that a semantically-grounded RL method increases the risk of re-identification compared to existing methods and, hence, it better evaluates the potential disclosure risk of masked data.

The rest of the paper is organized as follows. Section 2 reviews masking methods dealing with textual data. Section 3 describes the semantic foundations of our method: ontologies and semantic similarity. Section 4 present and formalizes our SRL method. Section 5 tests our approach by evaluating the disclosure risk of generalization-based masking methods under different configurations, comparing it to a non-semantic approach. The final section contains the conclusions and several lines of future research.

2. **Anonymising textual attributes.** As stated in the introduction, an anonymisation method takes a dataset consisting on a set of records (*i.e.* individuals) and set of attributes (*i.e*. responses). These attributes can be classified as: *identifiers* (which unambiguously identify the individual, such as ID-card number) and *quasi-identifiers* (which may identify some of the respondents, especially if they are combined with the information provided by other attributes, such as job or place of birth). Quasi-identifiers can be divided into *confidential attributes* (which contain sensitive information, such as medical conditions) and *non-confidential attributes* (the rest). The goal of statistical disclosure control is to prevent the link of the published confidential information to unique individuals. Before publication, *identifiers* are removed from the dataset. Whereas confidential information is unknown by third parties, *non-confidential quasi-identifier attribute values* may be known and can be used to re-identify the respondent. Although, they do not link to specific respondents if they are considered separately, the problem arises if they are considered in groups (*e.g*. job + city of living + age). Consequently, before releasing the data, these attributes must be masked by means of an anonymisation algorithm, resulting in a modified dataset [19, 20]. As stated in the introduction, a classical approach is to ensure that the masked dataset is *k*-anonymous (*i.e*. any record is indistinguishable from *k-1* other ones). The value of *k* defines the desired level of privacy and influences the information loss.

In the following, we review works proposing anonymisation schemas focused on textual attributes. By analyzing and understanding their behavior, we will be able to propose a specially tailored RL method that evaluates better the potential risk of disclosure of these methods.

The most basic ones consider textual data as enumerated terms for which only Boolean

word matching operations can be performed (*i.e.* in a categorical manner). We can find methods based on data swapping (which exchange values of two different records) and noise addition (such as replacing values according to some probability distribution used by PRAM [21, 22]). Others [1, 23] perform local suppressions of certain values or select a sample of the original data while maintaining the information distribution of input data. Even though these methods achieve a certain degree of privacy, they fail to preserve the meaning of the original dataset due to their complete lack of semantic analysis. As stated in the introduction, the goal is that the conclusions drawn from the masked data would be the same or very similar to those obtained from the original dataset.

In recent years, some authors have incorporated manually tailored knowledge structures to aid the interpretation of textual data and assist the masking process. Authors represent semantic relations between the set of values of each attribute in the dataset using *Value Generalization Hierarchies* (VGHs) [1, 23-27]. VGHs are manually constructed taxonomical structures defined according to the input dataset, where attribute labels are leaves of the hierarchy and these are recursively subsumed by common generalizations. The masking process consists on substituting attribute values by more general ones obtained from the hierarchical structure associated with that attribute. This generalization process decreases the number of distinct tuples in the dataset and, therefore, increases the level of *k*-anonymity.

For each value, different generalizations are possible according to the depth of the tree. Authors proposed approaches which restrict more or less the search space of generalizations. In [1, 23], all the values of each attribute are generalized to the same level of the VGH. Iyengar [25] presented a more flexible scheme in which a value of each attribute can be generalized to a different level of the hierarchy, resulting in a larger space of possible generalizations. T. Li and N. Li [26] propose three generalization schemes. In the *Set Partitioning Scheme* (SPS), each possible partition of the attribute values represents a generalization. This provides the most flexible generalization scheme but the size of the solution space grows enormously while the benefits of a VGH are not exploited. The *Guided Set Partitioning Scheme* (GSPS) uses a VGH to restrict the partitions of the corresponding attribute. Finally, the *Guided Oriented Partition Scheme* (GOPS) adds ordering restrictions to the generalized groups of values to narrow the set of possible generalizations even more.

To retain the utility of data, the masking method should select, from all the possible combinations of generalized tuples fulfilling the *k*-anonymity, the one that minimizes the information loss. In exhaustive approaches, the search space (which depends on the generalization constraints detailed above) results in NP-hard algorithms, which can only be applied to small datasets. Due to this reason, some authors opted by a non-optimum heuristic approach [5, 8, 26, 27].

Information loss is measured in these methods according to a metric. One the one hand, we can find distributional metrics such as the Discernibility Metric (DM) [24], that evaluate the distribution of *m* records (corresponding to *m* individuals) into *c* groups of identical values. However, metrics based on data distribution do not capture how semantically similar the anonymised set is with respect to the original data. Thus, other authors [1, 23, 27] measured the information loss as a function of the level of generalization applied during the

masking process. Some authors [1, 23] quantify this loss as the number of taxonomic links needed to go from the original value to its generalization. The higher the generalization, the more abstract the masked dataset will be, resulting in a higher loss of semantic content. This measure provides more accurate assessments of the differences between the semantic content of the original and masked dataset [28].

It is important to note that, in addition to the quality metric, the design of the VGH used to assist the masking process has a direct influence in the information loss and their utility from a semantic point of view. One may construct a VGH that progressively propose fine grained generalizations for attribute labels (*e.g*. sailing -> water sport -> sport -> activity). In this case, each generalization produces a lower loss of semantics than coarser taxonomical structures (*e.g*. sailing -> activity). The disclosure risk in detailed VGHs, however, may increase because the generalizations are less abstract and can be more easily linked with the original labels. So, there is a trade-off between information loss and disclosure risk: when one decreases, the other tend to increase. Finding the equilibrium is a difficult task that should be carefully considered.

3. **Enabling semantically-grounded record linkage.** A Record Linkage method, especially when tailored for a specific masking method, can be seen as a reverse engineering process, in which an intruder tries to guess and undo the data transformations performed during the anonymisation process. In the case of masking methods dealing with textual data discussed above, two elements influence how the anonymisation is performed: the underlying knowledge structure used to propose generalizations, and the quality criteria used to decide the one that minimizes the information loss. Obviously, both elements are variables that remain hidden to the intruder. In consequence, the RL method should either guess them from input data or substitute them by other elements that are general enough to be applicable even when the masking criteria vary.

Generalization methods use ad-hoc taxonomical structures constructed according to input attribute labels to propose value generalizations. To undo generalizations, an accurate RL requires a similar knowledge base. Considering that the design and structure of the VGH depends on the way in which the anonymizer structured the knowledge, it is neither feasible nor scalable to guess the VGH. Instead, we use available knowledge structures that aim to be general enough to cover most of the concepts that may appear in a domain: ontologies. Ontologies are formal and machine-readable structures, representing a shared conceptualization of a knowledge domain, expressed by means of semantic relationships. They have been successfully applied in many areas dealing with textual resources [29] and knowledge management [30]. Ontologies present several advantages compared with VGHs. Widely used ontologies provide a taxonomical structure much larger and finer grained than VGHs, being created from the consensus of a community of knowledge experts. They represent knowledge in an objective, coherent and detailed manner. This contrasts with the ad-hoc, overspecified and coarse nature of VGHs which can be hardly assessed.

With such ontologies, attribute values (*i.e*. words) presented in input datasets can be mapped to ontological nodes (*i.e*. concepts) via simple word-concept label matching. In this

manner, the hierarchical tree to which each textual value belongs can be explored to retrieve possible generalizations and/or specializations that can assist the RL process.

From a domain independent point of view, one can use a general ontology like WordNet. WordNet [31] is a freely available lexical database that describes and organizes more than 100,000 general concepts, which are semantically structured in an ontological way. The result is a network of meaningfully related words, where the graph model can be analyzed to interpret a concept's semantics. Hypernymy is by far the most common relation, representing over 80% of all the modeled semantic links.

Once we have selected the knowledge source in which the RL will rely, it is necessary to define a criterion to match records between the masked and original datasets. Distance-based RL methods define a measure by means of which the closest records are matched. This measure should be as similar as possible to the quality metric used to anonymise data. In the case of generalization methods, one can assume that the anonymizer has selected the generalization that minimizes the information loss. From a semantic point of view, information loss is a function of the difference between the degree of generality of the original and masked values. So it can be seen as a measure of semantic alikeness. On the contrary to related works [4, 9, 10, 32] focused on numerical and ordinal data, which evaluate textual values in a categorical way, we rely on the *semantic similarity* [33] to properly compute the semantic distance between textual labels and guide the RL process. Semantic similarity estimates the taxonomical resemblance of terms based on the evidence extracted from one or several knowledge sources. In the literature, several approaches can be identified. We focus on semantic similarity measures that only rely on ontologies and, more concretely, taxonomical knowledge. Ontologies are seen as a directed graph in which taxonomic interrelations are modeled as links between concepts, and their semantic distance can be estimated by counting the number of edges separating them. Several edge-counting approaches have been developed [33-35]. They are characterized by being easily applicable and highly efficient, lacking the constraints and dependencies on external resources that other semantic similarity paradigms present [36].

The simplest way to estimate the semantic distance (*i.e.* the inverse to similarity) between two ontological nodes (c_1 and c_2) is to calculate the shortest Path Length (*i.e.* the minimum number of taxonomical links) connecting these elements, see definition (1) [33].

$$distance_{Path\_length}(c_1, c_2) = min\#\ of\ is-a\ edges connecting\ c_1\ and\ c_2 \quad (1)$$

However, this measure omits the fact that equally distant concept pairs belonging to an upper level of the taxonomy should be considered as less similar than those belonging to a lower level because they present different degrees of generality. Based on this premise, Wu and Palmer's measure [35] also takes into account the depth of the concepts (2).

$$similarity_{w\&p}(c_1, c_2) \frac{2 \times N_3}{N_1 + N_2 + 2 \times N_3} \quad (2)$$

where N_1 and N_2 are the number of is-a links from c_1 and c_2, respectively, to their Least Common Subsumer (LCS), and N_3 is the number of is-a links from the LCS to the root node.

In [36] it is proposed a new measure that aims to improve edge-counting measures by evaluating additional taxonomic knowledge modeled in ontologies. Instead on basing the assessment only on the length of the minimum path, authors evaluate, in a non-linear way, the number of non-common subsumers between the compared concepts as an indication of distance. This value is normalized by the set of subsumers of both concepts (3).

$$sim(c_1, c_2) = -\log_2 \left(1 + \frac{\left|T(c_1) \cup T(c_2)\right| - \left|T(c_1) \cap T(c_2)\right|}{\left|T(c_1) \cup T(c_2)\right|} \right) \quad (3)$$

where $T(c_i)$ is the set of taxonomic subsumers of the concept c_i, including itself.

4. **A new RL method for value generalization masking schemas.** In this section, we propose a new record linkage method for textual attributes relying on a semantic interpretation of the values. The method, named *Semantic Record Linkage* (SRL), is designed for dealing with anonymization schemas based on value generalizations, which are the most common when dealing with textual data.

4.1 **Semantic Record Linkage.** Starting from a dataset in which each record corresponds to an individual, let us consider the typical anonymisation scenario used in works like [11, 19] consisting on: i) identifier attributes (*e.g.* ID-card numbers) have been removed from the dataset, ii) if an attribute is considered confidential (*e.g.* salary) then it is not modified, and iii) the anonymisation is applied to quasi-identifier non-confidential attributes (*e.g.* job, city-of-living, personal preferences). The resulting dataset D consists on m records, each of them composed by n quasi-identifier non-confidential attributes and c confidential attributes. Let us have that D^A is the publishable and, therefore, anonymised version of D, containing m records with n anonymised quasi-identifier non-confidential attributes and the initial c confidential attributes. Let us consider that an intruder gathers information about the set of individuals in D, and builds a dataset E that contain the same n non-confidential quasi-identifiers that appear in D, together with some identifier attributes. Assuming that some (or all) of the records in E correspond to individuals that are also in D, the intruder can access confidential data (*e.g.* salary) if he is able to link a record $r_k^E \in E$ with the anonymised (and published) record $r_i^A \in D^A$, so that r_k^E and r_i^A correspond to the same individual, disclosing his identity. This can be achieved by using the common non-confidential attributes in E and D^A; that is $E \cap D^A$. The amount of correct record linkages evaluates the disclosure risk of the privacy-preserving method.

According to this scenario, the proposed *Semantic Record Linkage* method (SRL) can be applied to the set of quasi-identifier non-confidential attributes if they consist on textual values. As stated in section 3, the method relies on ontologies to assess the semantic similarity between textual values. The linkage is done calculating the maximum similarity

between the values that the intruder knows (*i.e.* the textual attributes in *E*) and the anonymised attributes published (*i.e.* the textual attributes in D^A obtained by a generalization process from the original values in *D*). The linkage method is formalized as follows.

Let us have that *D* is composed by *m* records, $r_i = (r_{i1},...,r_{in})$, D^A consist on the same number of anonymised records, $r_i^A = (r_{i1}^A,..., r_{in}^A)$, and *E*, owned by the intruder, has some records $r_k^E = (r_{k1}^E,..., r_{kn}^E)$, where r_{ij} and r_{ij}^A and r_{kj}^E are textual values.

Definition 4.1. *The set of linked records (L) with respect to each* r_k^E *is:*

$$L_{r_k^E} = \left\{ l \;\middle|\; l = \underset{\forall i=1..m, r_i^A \in D^A}{argmax} \left(record\_similarity\left(r_k^E, r_i^A\right)\right) \right\} \tag{4}$$

The intruder searches for the least distant record to r_k^E in D^A. Because the result may be non-unique (*i.e.* equally similar records), we obtain a set of linked records *L*.

The SRL method relies on the measurement of the semantic similarity between the textual values that appear in each record in order to estimate their alikeness. Considering that a generalization-based masking method tries to minimize the information loss by suggesting the closest subsumer that satisfies the *k*-anonymity (see section 2), our SRL method hypotheses that the semantically closest record in *D* for an anonymised one $r_i^A \in D^A$ should be r_i (*i.e.* the original version of r_i^A). The record similarity is then computed as follows.

Definition 4.2. *The similarity between two records* r_i *and* r_k *is defined as the arithmetic average of the semantic similarity between each of their attribute values:*

$$record\_similarity_{SRL}(r_i, r_k) = \frac{\sum_{j=1}^{n} sem\_sim_O(r_{ij}, r_{kj})}{n} \tag{5}$$

where the function sem_sim_O corresponds to any of the semantic similarity measures presented in section 3. *O* is the ontology used to calculate the similarity between r_{ij} and r_{kj}.

4.2 **Evaluation of Disclosure Risk based on Semantic Record Linkage**. The *disclosure risk* (DR) of a privacy-preserving method can be measured as the difficulty in finding correct linkages between original and masked datasets. This is done by counting the amount of correct linkages that the intruder is able to perform between *E* and D^A. DR is evaluated for the worst possible case, assuming that *E* contains all the *m* records of D^A and all the *n* non-confidential quasi-identifier attributes [9]. DR is calculated as the percentage of the average probability of linking each record r_k^E in D^A denoted as $p_{D^A}(r_k^E)$, as follows.

Definition 4.3. *The Disclosure Risk (DR) is computed as:*

$$DR = \frac{\sum_{k=1}^{m} p_{D^A}(r_k^E)}{m} \cdot 100 \tag{6}$$

where $p_{D^A}(r_k^E)$ is measured as follows.

Definition 4.4. *Being* $L \subset D^A$ *the set of records with maximum similarity with respect to each record* r_k^E*, and assuming that* r_k^A *in* D^A *and* r_k^E *correspond to the same individual, the probability of making a correct linkage is calculated as:*

$$p_{D^A}(r_k^E) = \begin{cases} 0 & if\ r_k^A \notin L \\ \dfrac{1}{|L|} & if\ r_k^A \in L \end{cases} \quad (7)$$

5. **Evaluation.** In this section we test the behavior of our SRL method in the evaluation of the disclosure risk of generalization schemas dealing with textual data, comparing it to a non-semantic approach relying on the matching of textual labels (as previous works).

5.1 **Evaluation Data.** The dataset used for evaluation consists on a set of *real* answers to polls made by the *Observatori de la Fundació d'Estudis Turístics Costa Daurada* at the *Delta de l'Ebre* National Park. Visitors were asked to respond to several questions (see an extract in Table 1). The dataset comprises 975 individual records.

TABLE 1. Extract of sample microdata used for evaluation.

Age	Gender	Visit (days)	Companion	Country	Reason	Activities
23	M	1	2	Spain	nature	fishing
26	M	3	1	Spain	landscape	sports
45	F	3	2	Belgium	sports	bicycling
56	M	1	0	France	nature	culture
26	F	5	3	France	fishing	nature
45	F	1	1	Spain	relaxation	culture
30	M	2	0	Holland	holidays	visit

The two textual attributes available in the dataset (two last columns in Table 1) have been considered as non-confidential quasi-identifiers, so they will be anonymised and used to perform the record linkage afterwards. Considering these two attributes, we obtain 211 different response combinations, 118 of which were unique (*i.e*. identifying a single person). Figure 1 shows the equivalence class structure defined by the values of these two attributes.

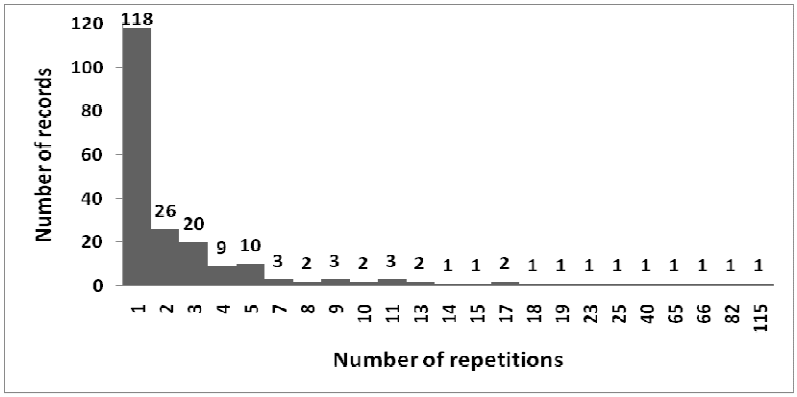

FIGURE 1. Attribute distribution according to answer repetitions

5.2 **Masking method.** To evaluate the SRL method, we have implemented a generalization algorithm that aims to depict the methods discussed in section 2. Due to the size of the data used during the evaluation, we opted by a non-exhaustive method based on a best-first search strategy (similar to [5, 8, 26, 27]). To reproduce the best scenario from the data utility point of view, we used a quality measure that quantifies the number of generalization steps performed at each transformation (like in [1, 23], as discussed in section 2). It is important to note that, on the contrary to simpler approaches [1, 23-25], the search space of the implemented algorithm is not constrained, and each value can be changed by any of the concepts that generalize it. This configures a more realistic but also challenging scenario.

Both the best-first search algorithm and the quality measure rely on a hierarchical structure that defines the possible generalizations for each value found in the dataset. In the same manner as the methods described in section 2, we have constructed ad-hoc VGHs. For this dataset, 25 distinct terms appear in the two attributes considered. In consequence, 25 leaves taxonomically connected through generalization concepts are contained in the VGH. To evaluate the influence of the VGH design, we have constructed two different VGHs. The first one (Figure 2, denoted as VGH2), incorporates up to two levels of generalization. The second on (Figure 3, named VGH3) models a finer grain classification.

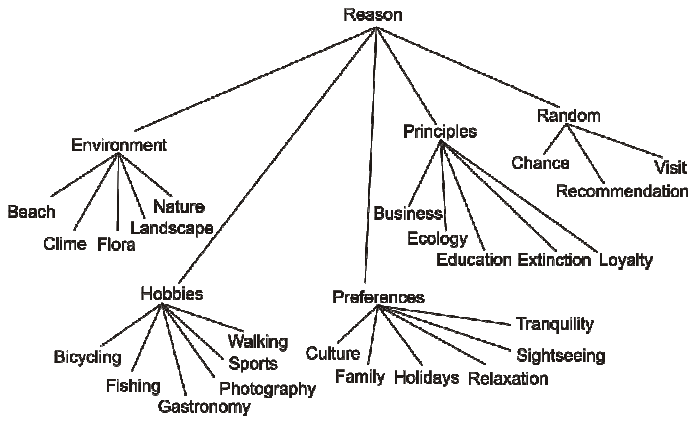

FIGURE 2. VGH2, modeling up to two levels of generalization per label.

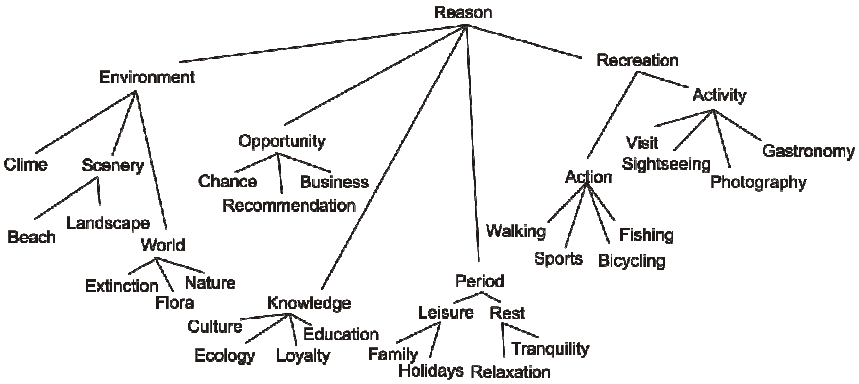

FIGURE 3. VGH3, modeling up to three levels of generalization per label.

5.3 **Evaluation of RL.** The results obtained from the anonymiszation have been evaluated by means of our SRL method, using WordNet (version 2) as ontology and the semantic similarity measures introduced in section 3 as the criteria to propose linkages. To test the adequacy of the SRL we have compared it against a non-semantic implementation of RL (named *Matching-based Record Linkage*, MRL). The MRL method represents the expected behavior of record linkage without background knowledge and dealing with textual data in a categorical fashion, like in [4, 9, 10, 32]. In this case, to build the set of linked records L (as in Eq. 4), the record similarity can only be based on the terminological matching of textual labels. It searches for records with exactly the same values in E and D^A and assigns them a maximum similarity value. Formally, the record similarity is:

$$record\_similarity_{MRL}(r_i, r_k) = \begin{cases} 1 \text{ if } r_i = r_k \\ 0 \text{ if } r_i \neq r_k \end{cases} \quad (8)$$

5.4 **Results.** The first study regards to the results obtained when using different semantic similarity measures (section 3) in comparison to a non-semantic approach (MRL). Figure 4 shows the evaluation of the disclosure risk (definition 4.3) of the generalization method for *k*-anonymity values from 2 to 20. For the SRL method, the three semantic similarity measures introduced in section 3 have been used. On the left, it is shown the percentage of correct record linkages obtained with a dataset masked using VGH2 while on the right the results when using a more detailed knowledge structure, VGH3 are given.

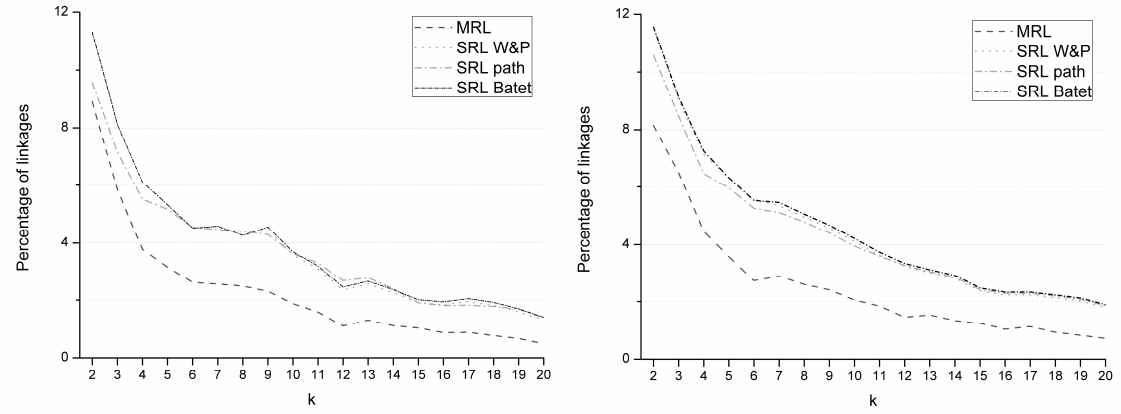

FIGURE 4. Disclosure risk evaluated by means of SRL and MRL methods, using VGH2 (on the left) and VGH3 (on the right) as the knowledge base.

Several conclusions can be extracted. First, the proposed SRL method is able to improve the amount of correct linkages proposed by the non-semantic approach (MRL). The differences are more evident for values of k from 4 to 8, because larger k-anonymity levels imply a higher loss of information. In this interval, the amount of correct linkages obtained with SRL almost double those achieved by the MRL approach. Moreover, the decreasing of the number of linkages as the k-value increases is coherent to that what it is represented in the distribution of the dataset shown in Figure 1 (most of the records have a number of

repetitions between 1 and 5). This explains the abrupt decrease in the number of linkages for same range of *k*-values. It is interesting to note that, regardless the *k*-value, the SRL method will always outperform the MRL counterpart. In fact, for *k*-values higher than the number of maximum repetitions of any record (118 in our case, as shown in Figure 1), the number of linkages obtained by the MRL method will be zero, due to all the labels in the masked dataset will be generalized. The SRL method, on the contrary, will always propose record linkages with a probability of correct linkage depending on the number of total records.

Regarding the SRL method, the differences when using each semantic similarity measure are minor, even though the approach by Wu and Palmer [35] and Batet et al. [36] provided a slightly higher amount of linkages. This is coherent to what was evaluated in [36], in which former measures improved the similarity assessment accuracy of path-based ones by a considerable margin, when compared with human ratings of term similarity. In our case, semantic similarity measures are only used to rank pairs of terms and select the most similar ones. Results in [36] showed that, even though the similarity assessment of each measure may be different, the relative order of the resulting raking is quite similar. Consequently, the selection of the semantic similarity function does not have a noticeable effect in the results.

Analyzing the differences in the disclosure risk when using VGH2 or VGH3 more correct record linkages are obtained when using the most detailed knowledge structure. Figure 5 shows the increment (in percentage) of correct linkages of the SRL method with respect to the basic MRL in both cases. We quantify among a 10-25% improvement in the amount of record linkages when using the SRL method applied to the dataset masked according to VGH3. It is worth to note that when using a more detailed knowledge base to guide the anonymisation process (VGH3), the masked values are more similar to the original ones, due to the lower level of abstraction introduced by the generalization process. In consequence, a RL method that is able to evaluate this semantic difference reveals a higher disclosure risk. On the contrary, a non-semantic RL approach obtains similar disclosure risk because, in both cases (VGH2 and 3), the original labels have been changed.

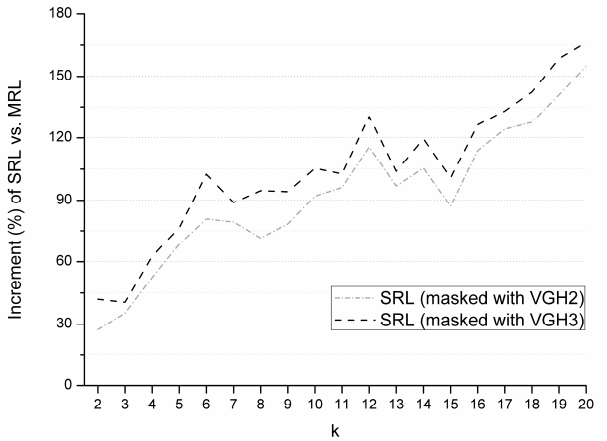

FIGURE 5. Increment in the amount of correct linkages of SRL (using Batet et al. as similarity measure) with respect to the MRL, masking data according to VGH2 and VGH3.

Finally, we can see that the difference in percentage between the SRL for both VGH2 and VGH3 (with respect to MRL) is maintained in the range 10-25% along *k* values, stating that this difference is independent of the level of privacy. The relative difference between SRL and MRL, on the contrary, increases significantly as does the *k*-value. One may conclude that, from the point of view of minimizing the risk of disclosure, one should use simpler hierarchies of concepts (with few levels of generalization), due to the higher level of abstraction of the values. However, as stated in the introduction, anonymisation methods should also maximize the utility of data, minimizing the information loss. To quantify the information loss when using knowledge structures with different levels of detail, we measured how semantically similar the masked records are with respect to the original ones. The *information loss* of D^A with respect to D has been computed as follows:

$$information\_loss_D\left(D^A\right) = \frac{\sum_{i=1}^{m} record\_similarity\left(r_i, r_i^A\right)}{m} \quad (9)$$

where *record_similarity* has been computed as defined in Eq. 5, following a similar criteria as the one used to guide the anonymisation process.

Again, WordNet has been used as ontology, enabling an objective comparison of the semantic differences when using each VGH. Figure 6 shows the evolution of information loss for each VGH, according to the *k*-anonymity level.

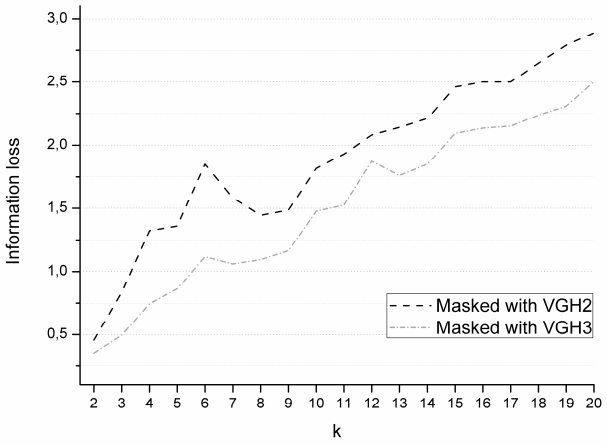

FIGURE 6. Information loss according to the type of VGH used during the anonymisation.

On one hand, we observe a higher information loss when a simpler VGH is used. Hierarchies with fewer nodes produce more abstract masked values as a result of each generalization step. When compared to a detailed ontology like WordNet, the semantic distance of the masked data is higher. In consequence, data utility will decrease because it is related to the preservation of the semantics of textual values [28] . It is also worth noting that even though using a detailed knowledge base to guide the anonymisation process is desirable from the data utility point of view, the fact that larger and finer grain generalizations are

available also increases the search space of possible value transformations. Due to the algorithmic design of generalization methods, the use of a source as large as WordNet is not feasible. The search space of possible generalizations for each value would be so high that even methods based on heuristic searches will not scale with large amounts of data [5, 8].

On the other hand, we notice a linear trend in the increasing of information loss according to the level of *k*-anonymity. Figures 5 and 6 show that the use of more detailed knowledge structures (such as VGH3) decreases information loss. Notice that the results obtained show an opposite trend with respect to the ones obtained when evaluating the disclosure risk (Figure 4). This indicates that there is a trade-off between the preservation of data utility and the disclosure risk. The differences in the curve shapes (almost linear for information loss vs. inverted log for disclosure risk) suggest that it is not convenient to protect data with high *k*-anonymity values, because the consequent loss in data utility will be comparatively higher than the decrease in the disclosure risk.

6. **Conclusions.** As stated in [11] tailored record linkage methods are convenient to reflect the *feasible* degree of data re-identification. The work presented is a step forward in this area, proposing a new record linkage method based on semantic similarity theory and using ontologies to evaluate masking methods based on generalizing textual values. Evaluation results show the convenience of using semantically-grounded RL methods compared to non-semantic algorithms. The tests have also gone a step further, evaluating the influence of the knowledge bases both in the information loss and in the disclosure risk.

The importance of textual data analysis have grown in recent years, being framed in many contexts such as the Web (*e.g.* query log analysis), digital library structuring (*e.g.* document classification) or user profile management (*e.g.* recommender systems). The work presented in this paper opens a new line of research both in the development of more suitable RL methods and in the definition of more robust anonymisation schemas focused on ensuring the privacy of textual data used for analysis.

As future research, some points can be devised. First, the use of several ontologies to assist the record linkage process could bring benefits thanks to the exploitation of additional knowledge. Second, linguistic techniques like morpho-syntactic analyses and part-of-speech tagging can be applied to extend the SRL approach not only to simple textual answers (*i.e*. words or noun phrases) but also to free textual answers consisting on complex sentences. This will permit using this kind of methods in a wider range of applications, such as private information retrieval from Web search Engines or the anonymization of free text clinical outcomes.

Acknowledgement. We would like to thank the Observatori de la Fundació d'Estudis Turístics Costa Daurada and the Delta de l'Ebre National Park for providing the data. 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 Universitat Rovira i Virgili.

REFERENCES

[1] L. Sweeney, k-anonymity: a model for protecting privacy, *International Journal Uncertainty Fuzziness Knowledge-Based Systems*, vol.10, no.5, pp.557-570, 2002 2002.

[2] J. Domingo-Ferrer, A Survey of Inference Control Methods for Privacy-Preserving Data Mining, *Privacy-Preserving Data Mining*, vol.34, Springer US, pp. 53-80, 2008.

[3] H. J. Do and J. Y. Kim, Clustering Categorical Data Based on Combinations of Attribute Values, *International Journal of Innovative Computing, Information and Control*, vol.5, no.12, pp.4393-4405, 2009.

[4] J. Domingo-Ferrer and V. Torra, Disclosure control methods and information loss for microdata, *Confidentiality, Disclosure, and Data Access: Theory and Practical Applications for Statistical Agencies*, pp. 91-110, 2001.

[5] S. Martinez, D. Sanchez, and A. Valls, Ontology-Based Anonymization of Categorical Values, *Modeling Decisions for Artificial Intelligence*, vol.6408, Springer Berlin / Heidelberg, pp. 243-254, 2010.

[6] S. Martínez, D. Sánchez, A. Valls, and M. Batet, Privacy protection of textual attributes through a semantic-based masking method, *Information Fusion*, vol.In Press, Accepted Manuscript, 2011.

[7] L. Yu, X. Liu, F. Ren, and P. Jiang, 2749–2760, *International Journal of Innovative Computing, Information and Control*, vol.5, no.12, pp.4637-4645, 2009.

[8] S. Martínez, D. Sánchez, A. Valls, and M. Batet, Privacy protection of textual attributes through a semantic-based masking method, *Information Fusion*, In Press.

[9] V. Torra and J. Domingo-Ferrer, Record Linkage methods for multidatabase data mining, *Information Fusion in Data Mining*, Springer, 2003.

[10] W. E. Winkler, Re-identification Methods for Masked Microdata, *Privacy in Statistical Databases*, vol.3050, Springer Berlin / Heidelberg, pp. 519-519, 2004.

[11] J. Nin, J. Herranz, and V. Torra, Rethinking rank swapping to decrease disclosure risk, *Data & Knowledge Engineering*, vol.64, no.1, pp.346-364, 2008 2008.

[12] G. J. Mclachlan and T. Krishnan, *The EM Algorithm and Extensions*, 1997.

[13] J. Nin, J. Herranz, and V. Torra, On the disclosure risk of multivariate microaggregation, *Data & Knowledge Engineering*, vol.67, no.3, pp.399-412, 2008.

[14] P. Medrano-Gracia, J. Pont-Tuset, J. Nin, and V. Muntés-Mulero, Ordered Data Set Vectorization for Linear Regression on Data Privacy, *Modeling Decisions for Artificial Intelligence*, vol.4617, Springer Berlin / Heidelberg, pp. 361-372, 2007.

[15] V. Torra and S. Miyamoto, Evaluating Fuzzy Clustering Algorithms for Microdata Protection, *Privacy in Statistical Databases*, vol.3050, Springer Berlin / Heidelberg, pp. 519-519, 2004.

[16] T. Dalenius and S. P. Reiss, Data-swapping: A technique for disclosure control, *Journal of Statistical Planning and Inference*, vol.6, no.1, pp.73-85, 1982.

[17] X. Q. Yang, N. Sun, T. L. Sun, X. Y. Cao, and S. J. Zheng, The Application of Latent Semantic Indexing and Ontology in Text Classification, *International Journal of Innovative Computing, Information and Control*, vol.5, no.12, pp.4491-4499, 2009.

[18] R. C. Cheng, C. T. Bau, M. Y. Tsai, and C. Y. Huang, Web Pages Cluster Based on the Relations of Mapping Keywords to Ontology Concept Hierarchy, *International Journal of Innovative Computing,*

Information and Control, vol.6, no.6, pp.2749–2760, 2010.

[19] J. Domingo-Ferrer and V. Torra, A quantitative comparison of disclosure control methods for microdata, *Confidentiality, Disclosure and Data Access*, Amsterdam: North-Holland, pp. 111–133, 2001.

[20] D.-W. Wang, C.-J. Liau, and T.-s. Hsu, An epistemic framework for privacy protection in database linking, *Data & Knowledge Engineering*, vol.61, no.1, pp.176-205, 2007 2007.

[21] L. Guo and X. Wu, Privacy Preserving Categorical Data Analysis with Unknown Distortion Parameters, *Trans. Data Privacy*, vol.2, no.3, pp.185-205, 2009.

[22] J. M. Gouweleeuw, P. Kooiman, L. C. R. J. Willenborg, and P. P. DeWolf, "Post randomization for statistal disclousure control: Theory and implementation," Voorburg: Statistics Netherlands Research paper no. 9731, 1997.

[23] P. Samarati and L. Sweeney, 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 1998.

[24] R. J. Bayardo and R. Agrawal, "Data Privacy through Optimal k-Anonymization," presented at the the 21st International Conference on Data Engineering, 2005.

[25] V. S. Iyengar, "Transforming data to satisfy privacy constraints," presented at the KDD, 2002.

[26] T. Li and N. Li, Towards optimal k-anonymization, *Data & Knowledge Engineering*, vol.65, no.1, pp.22-39, 2008 2008.

[27] Y. He and J. Naughton, Anonymization of Set-Valued Data via Top-Down, Local Generalization, *Proc. of the VLDB '09: the Thirtieth international conference on Very large data bases*, Lyon, France, 2009.

[28] S. Martinez, D. Sanchez, A. Valls, and M. Batet, The Role of Ontologies in the Anonymization of Textual Variables, *Proc. of the the 13th International Conference of the Catalan Association for Artificial Intelligence*, pp.153-162, 2010.

[29] D. Sánchez, D. Isern, and M. Millan, Content annotation for the semantic web: an automatic web-based approach, *Knowledge and Information Systems*, pp.1-26, 2010 2010.

[30] A. Valls, K. Gibert, D. Sánchez, and M. Batet, Using ontologies for structuring organizational knowledge in Home Care assistance, *International Journal of Medical Informatics*, vol.79, no.5, pp.370-387, 2010 2010.

[31] C. Fellbaum, *WordNet: An Electronic Lexical Database (Language, Speech, and Communication)*, The MIT Press, 1998.

[32] W. Yancey, W. Winkler, and R. Creecy, Disclosure Risk Assessment in Perturbative Microdata Protection, *Inference Control in Statistical Databases*, vol.2316, Springer Berlin / Heidelberg, pp. 49-60, 2002.

[33] R. Rada, H. Mili, E. Bicknell, and M. Blettner, Development and application of a metric on semantic nets, *Systems, Man and Cybernetics, IEEE Transactions on*, vol.19, no.1, pp.17-30, 1989 1989.

[34] C. Leacock and M. Chodorow, Combining local context with WordNet similarity for word sense identification, *Proc. of the WordNet: A Lexical Reference System and its Application*, 1998.

[35] Z. Wu and M. Palmer, "Verbs semantics and lexical selection," presented at the the 32nd annual meeting on Association for Computational Linguistics, Las Cruces, New Mexico, 1994.

[36] M. Batet, D. Sánchez, and A. Valls, An ontology-based measure to compute semantic similarity in biomedicine, *Journal of Biomedical Informatics*, vol.44, no.1, pp.118-125, 2011.