Semantic-Based Aggregation for Statistical Disclosure Control

Aïda Valls,[1,*] Vicenç Torra,[2,†] Josep Domingo[1,‡]
[1]*Department of Computer Engineering and Maths—ETSE, Universitat Rovira i Virgili, Av Països Catalans 26, 43007 Tarragona, Catalonia, Spain*
[2]*Institut d'Investigació en Intel.ligència Artificial—CSIC, Campus UAB s/n, 08193 Bellaterra, Catalonia, Spain*

In this paper we show how clustering can be used to aggregate different versions of the same data set in order to discover confidential information. Having these tools helps to not publish data that could be reidentified, which is known as Statistical Disclosure Control. In particular, the paper is focused on the case of dealing with categorical values. © 2003 Wiley Periodicals, Inc.

1. INTRODUCTION

In recent years, the so-called information explosion has caused the development of new techniques for data analysis and information management. One class of techniques where this improvement can be found is the one related with information fusion and knowledge integration. As the number of available information sources and the amounts of information increase, the need for these techniques also increases. Now, applications of these techniques are as diverse as scientific fields. One of the particular applications of information fusion techniques is Statistical Disclosure Control (SDC).[1]

The mission of National Statistical Offices (NSOs) is to collect information from respondents and to their posterior publication. In fact, data dissemination is a requirement for NSOs as is the main justification for the resources spent and of their existence. However, data dissemination usually is a sensitive task because of reidentification risk. NSOs should process data before publication so that published data ensure that particular individuals or organizations cannot be reidentified, i.e., no sensible data are published in a way that can be reidentified with a particular respondent (see Ref. 2 for a review of reidentification methods). Thus, data have to be protected (this is the so-called disclosure control problem) to avoid possible

*Author to whom all correspondence should be addressed: e-mail: avalls@etse.urv.es.
†e-mail: vtorra@iiia.csic.es.
‡e-mail: jdomingo@etse.urv.es.

INTERNATIONAL JOURNAL OF INTELLIGENT SYSTEMS, VOL. 18, 939–951 (2003)
 Published online in Wiley InterScience (www.interscience.wiley.com). • DOI 10.1002/int.10129

reidentification. Failure of protection can cause major problems because of legal marks and because respondents would refuse new collaborations with the NSOs.

At present, several SDC methods (or masking methods) have been developed to protect data. These methods can be classified (see Refs. 3 and 4) according to the two main kinds of data that NSOs release (different data require different masking methods). On the one hand, masking methods for tabular data exist, i.e., tables containing aggregated data as the cell values. On the other hand, methods for microdata sets exist, i.e., sets of records, each containing information about an individual entity such as a person, household, business, etc. Additional classification can be defined for microdata methods: differences exist according to the type of information in the microdata file. In particular, classification depends on the types of variables used to describe the records in the file. Some methods are only applicable to numerical microdata (e.g., microaggregation, additive noise, and sampling), others to categorical microdata (e.g., Post-RAndomization Method (PRAM), local suppression, and sampling), and a third group to both types (rank swapping, global recoding, and top and bottom coding).

When different microdata methods are applied to the same original file, different masked files are generated. Selection of the appropriate method depends on several aspects: the protection required to avoid reidentification, the information loss advisable to avoid making the file useless, and the intended data use. In some cases, multiple protected versions of the same confidential data set are released, each one protected to minimize information loss for a particular use. In this case, an additional threat for reidentification risk appears because of the formation of coalitions of users. This is so because data fusion techniques can integrate the information contained in n different distorted versions of the data set, thus compromising statistical confidentiality. Note that the better the reconstruction, the larger the disclosure risk. This suggests that data fusion tools can be applied to multiple masked data files to evaluate to what extent the original data file can be reconstructed.

In this work, we focus on the problem of fusing categorical data and on evaluating the reconstruction achieved. We propose the use of the system Radamès based on clustering techniques and suggest the use of the so-called information loss measures to evaluate this reconstruction.

The structure of this work is as follows. Section 2 reviews the Radamès system, the system for fusing categorical data. Section 3 describes the method for evaluating the reconstruction of the original data set and, therefore, gives some risk evaluation; and Section 4 details the application of the method to a particular disclosure problem. Section 5 includes the conclusions and future work.

2. THE RADAMÈS SYSTEM

This section is devoted to the description of the Radamès system. This system was built to aggregate information of several individuals described using different variables for its application to multicriteria decision making. Therefore, as usual, data are represented as a two-dimensional table in which one dimension corresponds to the set of objects (or elements, individuals, persons, or business) and the

other is the set of variables (or attributes). The table contains a value for each pair (object and variable). From this information, the system builds a new variable that aggregates the information from all the original variables.

As information is represented using different types of variables, Radamès considers several data formalisms, i.e., different types of variables: boolean, numerical (the system distinguishes between ratio scale and interval scale), ordinal scale (a set of linguistic labels or categories totally ordered), and nominal (a set of linguistic labels in which order is not relevant). Different aggregation methods apply to different types of data. In this work we focus on the aggregation of categorical data; therefore, the description of the system is limited to this type of information. For numerical information the system supplies a set of aggregation functions such as weighted means, quasi-geometric means, ordered weighted averaging (OWA) operators,[5] and weighted OWA (WOWA) operators.[6] The system also permits the construction of the aggregated variable when the original variables are not all in the same type scale. In this case, we apply the same technique used for categorical data.

2.1. Fusion of Categorical Variables: An Outline

The main characteristics of the method is that it does not satisfy Arrow's condition of irrelevant alternatives. This is, the aggregated value for each object does not only depend on the values of the variables for that object but also on the values for the other objects. Usually, Arrow's condition is claimed to be a requirement for technical reasons[7] (it simplifies the computations because each object can be operated in a separate way). However, when data are categorical, the actual value of a given object usually is not important *per se* but in relation to the values that are assigned to other objects. For example, when nominal scales are considered, assignment of values to objects defines a partition of the set of objects. Then, it seems natural that objects that are kept together in different partitions also are kept together in the aggregated variable. In other words, the similarities of the objects according to each variable are relevant to find global relationships and calculate the aggregated value. Thus, in general, it seems convenient to use all the information about all the objects when calculating the aggregation of one of them. From our point of view[8] these relationships can be obtained through clustering methods. Following this idea, a clustering method[9] was defined to deal with categorical variables (in both ordinal and nominal scales). When several variables in ordinal scales are considered, an additional step is added to clustering: a ranking process. This is so because for ordinal scales, the aggregated variable Radamès builds also is defined on an ordinal scale. As clustering methods typically lead to a set of clusters in the product space of all the variables, these classes have to be ordered to be able to define an ordinal scale.

The definition of a distance to compare two objects is one of the key points of the clustering process. For categorical variables we use a new distance, based on a sematics induced from negation functions. We describe this semantics and the clustering process in the following section.

For the application of the system to the SDC problem, we assume that each masking method corresponds to one variable and that the aggregated variable obtained by our system corresponds to an approximation of the original data.

2.2. Negation Function-Based Semantics for Categories in Ordinal Scales

Defining a semantics for linguistic labels is a difficult task because it is not always possible for an expert to give detailed information of each label (either an interval or a fuzzy set). To avoid this detailed information, we have followed a different approach considering a negation function for each set of ordered linguistic labels. This approach has the advantage that the meaning of a negation can be understood as the opposite of a label in the vocabulary (i.e., the antonym in the sense of Ref. 10). In fact, the use of negations functions to deal with categorical data is usual in intelligent systems (see, e.g., Refs. 11, 12, and 13).

Let $L = \{x_0, \ldots, x_n\}$ be the set of ordered linguistic labels with $x_0 < \cdots < x_n$. Then, classical negation functions (see the aforementioned references) are functions from L to L that satisfy the following conditions:

(N1) If $x < x'$ then $N(x) > N(x')$ for all x, x' in L
(N2) $N(N(x)) = x$ for all x in L

These conditions completely determine the negation function as the following theorem shows:

PROPOSITION 1.[14] *For each set of ordered linguistic labels $L = \{x_0, \ldots, x_n\}$, there exists only one negation function that satisfies conditions N1 and N2. This negation function is defined by*

$$N(x_i) = x_{n-i} \quad \forall x_i \in L$$

This proposition makes explicit that classical negation functions correspond to situations where each label in the pair $\langle x_i, x_{n-i}\rangle$ is equally informative. If we consider the semantics of a set of labels as a mapping from the set of labels into (disjoint) subsets of the unit interval, then the interval attached to x_i will be equal to the one x_{n-i}.

Therefore, each time conditions N1 and N2 are required for a negation function, equal informativeness for all pairs $\langle x_i, x_{n-i}\rangle_{i \in \{1, \ldots, n/2\}}$ is assumed, and semantic constraints apply. However, equal informativeness is not always appropriate. This is the case, e.g., if we consider the set of linguistic labels {negative, zero, small, medium, large}. It is clear that the "extension" of negative (in whatever scale is expressed) is larger than small or medium. However, although equal informativeness is not acceptable, it also is not always possible for the expert to define an interval or a fuzzy set for each label because that would require a degree of accuracy that the expert can not always supply. To present an alternative to equal informativeness but keeping the required information to the minimum, ref. 15 introduced a new class of negation functions. Using these negations, an expert can provide more information about the semantics of each label in a more natural

way. The usefulness of these negation functions for expressing additional knowledge was proven in Ref. 16. In this work, classical and negation-based semantics were compared in a knowledge acquisition framework.

We give the definition of these negation functions. This definition shows that the negation of a label is a set. Conditions C1 and C2 for this alternative negation are, respectively, generalizations of the conditions N1 and N2 given previously. In fact, C2 is a generalization of the condition N3) if $x = N(x')$ then $x' = N(x)$ that is equivalent to N2. C0 is a technical condition.

DEFINITION 1.[15] *A function N from L to $\wp(L)$ is a negation function if it satisfies*

(C0) N is not empty and convex [i.e., $N(x)$ is a nonempty interval of linguistic labels in L]
(C1) If $x < x'$ then $N(x) \geqslant N(x')$ for all x, x' in L
(C2) If $x \in N(x')$ then $x' \in N(x)$

Note that because $N(x)$ is a set, $N(x) \geqslant N(x')$ means that all the linguistic labels in $N(x)$ are larger or equal to all the labels in $N(x')$, i.e., $\min N(x) \geqslant \max N(x')$.

Because for this definition the negation of a label is a set, it is no longer required (and not even desirable) to be equal in length for the pairs $\langle x_i, x_{n-i}\rangle$. Equal informativeness is now, at most, between x_i and the set $N(x_i)$.

Based on this assumption, a semantics (a mapping into subintervals in [0, 1]) can be inferred from the negation. We consider that $P = \{I(x_0), \ldots, I(x_n)\}$ corresponds to the semantics of all labels in L. To build such mapping, it is assumed that the sets cover the unit interval and that the intersection of any two sets is empty or punctual (if they are contiguous), i.e., $\cup_{p\in P}\, p = [0, 1]$ and $I(x_i) \cap I(x_j) = \varnothing$ or equal to $\{x\}$ for some $x \in I(x_i)$. However, not all partitions in the unit interval are adequate: negation functions introduce constraints on appropriate sets P. In particular, relationships between labels expressed by means of the negation functions also should be true in the intervals in P when the negation in the unit interval is considered. In particular, the following elements are considered for consistency: (i) the negation of all the elements of the interval $I(x_i)$ belong to the intervals attached to the negation of x_i; (ii) if $N(x_i) = \{x_{i0}, \ldots, x_{in}\}$, then neither the label x_{i0} nor the label x_{in} are *superfluous* in relation to the negation function. This latter condition means that there exists at least one element of the interval attached to x_i such that its negation belongs to $I(x_0)$ [respectively, to $I(x_{in})$]. Given a negation function, there are several consistent semantics. One of them is the following:

DEFINITION 2.[15] *Let N be a negation function from L to $\wp(L)$, according to definition* 1; *we define P_N as the set $P_N = \{I(x_0), \ldots, I(x_n)\}$, where*

$$I(x_i) = [m_{in}, m_{ix}] = \left[\frac{\sum_{x<x_i} |N(x)|}{\sum_{x\in L} |N(x)|}, \frac{\sum_{x \leqslant x_i} |N(x)|}{\sum_{x \in L} |N(x)|}\right]$$

where $|\cdot|$ stands for the cardinality operator.

In Ref. 15, the consistency of P_N in relation to the negation function $N(x) = 1 - x$ is proven mathematically. It also can be proven that classical negation functions imply, using this approach, a *classical* semantics: all the labels have intervals with the same length (i.e., same precision).

The distance function for linguistic labels used in Radamès uses this semantics: each label is mapped into the central point of the corresponding interval. Clustering obtains a set of classes that are interpreted as the ones induced by the aggregated value. Then, classes are ordered (taking into account the orderings of the initial variables) and, thus, the ordinal scale for the aggregated variable is defined. For the clustering process, Radamès uses the general classifier system Sedàs. This is a general classifier because it implements several sequential, agglomerative, hierarchic, nonoverlapping clustering (SAHN) methods. See Ref. 9 for details.

3. EVALUATING THE RECONSTRUCTION

In SDC literature there exist several methods to evaluate information loss. These methods are applied to evaluate to what extent a masked data file is different from the original one. Different methods focus on different aspects and properties of the files.

Because the procedure to integrate the information of several files intend to reconstruct the original matrix, it is appropriate to use information loss measures to compare the original data and the reconstructed one. Differences between both files should be minimal.

For categorical data, three types of information loss measures are distinguished (see Ref. 3 for details):

(1) Comparison of categorical values. A distance is computed between the original variable and the masked one. This distance is based on the distance between the original values and the masked ones. This requires the definition of a distance over the categories.
(2) Comparison of contingency tables. Given a subset of variables, contingency tables are built for both the original file and the masked file. The number of differences between the positions in both tables give an information loss measure.
(3) Entropy-based measures. The masking process is modeled as the noise added to the original file when transmitted through a noisy channel. The information loss measure uses conditional probabilities (the probability of a value in the original file given the value in the masked file).

In the case of evaluating the reconstruction of a data file when multiple releases are considered, the last alternative is not applicable because the needed probabilities are not known. Therefore, we can use only the first two approaches.

4. EXPERIMENTATION

The approach described here has been applied to evaluate to what extent several releases of masked files corresponding to an original one permit the reconstruction of the original file.

Table I. Negation function for the variable *DEGREE.*

N(coldest) = {hot}	N(mixed) = {cool, cold}
N(cold) = {hot, mixed}	N(hot) = {cold, coldest}
N(cool) = {mixed}	N(mild) = {cool, cold, coldest}

To do this evaluation, we have considered 1000 records of the *American Housing Survey 1993* (data publicly available from the U.S. Bureau of the Census through the *Data Extraction System*[17] and the variable *DEGREE* (it corresponds to long-term average heating degree days) defined on an ordinal scale. We have applied to these records a set of masking methods obtaining a set of masked variables. Risk assessment has been computed in the case that all masked variables are released and assuming that all users are in coalition. The example here is limited to 20 records. Table III includes in the first column the original values being masked.

4.1. The Variable and Its Semantics

The range of variable *DEGREE* is {coldest, cold, cool, mild, mixed, hot}. In this case, distance and aggregation have been defined using negation functions. We have used the one given in Table I. Note that the term mild overlaps with hot and mixed because its negation is defined as {cool, cold, coldest}. This term is not considered when building the intervals but their interval is defined as the union of the intervals inferred for hot and mixed. Using Definition 2, the negation in Table I induces the intervals given in Table II (a graphical interpretation is given in Figure 1).

4.2. Masking Variables

The variable has been masked using the most common masking methods for categorical data. In particular, we have considered the following perturbative and nonperturbative methods: top coding, bottom coding, global recoding, rank swapping, and PRAM. Because these methods depend on some parameters, particular parameterizations have been selected. In the case of PRAM, several parameterizations have been considered. We now briefly review the masking methods (see Ref. 3 for a detailed review of masking methods) and the selected parameterizations chosen. Parameterizations follow the extensive study in Ref. 18 comparing the performance of different masking methods with respect to information loss and reidentification risk in SDC.

Table II. Induced intervals for the variable *DEGREE.*

I(coldest) = [7/8, 8/8]	I(mixed) = [2/8, 4/8]
I(cold) = [5/8, 7/8]	I(hot) = [0/8, 2/8]
I(cool) = [4/8, 5/8]	I(mild) = [0/8, 4/8]

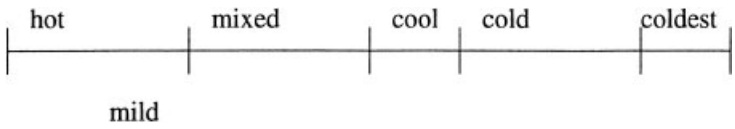

Figure 1. Intervals induced for the variable *DEGREE.*

Top Coding (abbreviated *Tp* in Table III, where *p* is the particular parameterization chosen). This method, appliable only to variables in ordinal scales, consists on the recoding of the highest *p* values of the variable into a new category. We have used the symbol "&" to denote the new category in Table III. A recoding of four categories has been considered (T4 in Table III). Top coding is applied to avoid the reidentification of the largest values because they are frequently easy to reidentify.

Bottom Coding (abbreviated *Bp*, where *p* is as mentioned previously). This masking method is similar to the previous case but now the lowest *p* categories are recoded into a new one. As before, we have selected $p = 4$ and the new category is codified by "&." As in the case of top coding, this masking method is applied to avoid the reidentification of the smallest values when the availability of this information allows the reidentification of the individuals.

Global recoding (abbreviated *Gp*, where *p* is the parameter). Global recoding consists of the recodification of some categories by some other ones. Selection

Table III. Records used.

Name	o.v.	B4	T4	G4	R10	P8	P9	P4	a.v.
a	3	&	&	3	3	3	3	3	3
b	3	&	&	3	2	3	3	3	3
c	3	&	&	3	3	3	3	3	3
d	3	&	&	3	3	3	3	3	3
e	4	&	&	4	4	4	4	4	4
f	4	&	&	4	4	1	1	1	1
g	4	&	&	4	3	4	4	4	4
h	4	&	&	4	4	4	4	4	4
i	4	&	&	4	4	4	4	4	4
j	1	&	1	*n*	1	1	1	1	1
k	3	&	&	3	4	3	3	3	3
l	3	&	&	3	2	3	3	3	3
m	3	&	&	3	3	3	3	3	3
n	2	&	2	2	2	2	2	2	2
o	3	&	&	3	3	3	3	3	3
p	2	&	2	2	2	2	2	2	2
q	3	&	&	3	3	3	3	3	3
r	5	5	&	*n*	5	3	3	4	5
s	2	&	2	2	2	2	2	2	2
t	2	&	2	2	2	3	4	2	2

First column corresponds to a name for the record; second column is the original value (o.v.); columns 3–9 are masked variables; column 10 is the aggregated value (a.v.).

of categories is done on the basis of increasing the number of individuals that match a particular category. For example (see Ref. 3), if there is a record with "marital status = widow/er" and "age = 17 years," global recoding could be applied to "marital status" to create a broader category "widow/er or divorced," so that the probability of the foregoing record being unique would diminish. In our experimentation, the following parameterization has been considered: recode the p lowest frequency categories into a single one. We have used $p = 4$.

Post-RAndomization Method or PRAM (abbreviated Pp, where p is as mentioned previously). This is a perturbative probabilistic method in which the value of a given individual is changed according to a prescribed probability mechanism (a Markov matrix). This method reduces the number of matching for all categories (reduction depends on the Markov matrix). The selected Markov matrix is based on the approach described in Ref. 19. This approach is as follows: Let $\mathbf{T}_V = (\mathbf{T}_V(1), \ldots, \mathbf{T}_V(K))^t$ be the vector of frequencies of the K categories of variable V in the original file (assume without loss of generality that $\mathbf{T}_V(k) \geqslant \mathbf{T}_V(K) > 0$ for $k < K$) and let θ be such that $0 < \theta < 1$. Then, the PRAM matrix for variable V is defined as

$$p_{kl} = \begin{cases} 1 - \theta \mathbf{T}_V(K)/\mathbf{T}_V(k) & \text{if } l = k \\ \theta \mathbf{T}_V(K)/((K-1)\mathbf{T}_V(k)) & \text{if } l \neq k \end{cases}$$

In our example we have considered different parameterizations $p = 4, 8, 9$ and $p := 10\theta$.

Rank swapping (abbreviated Rp, where p is a parameter). This method is better explained from their operational point of view. First, values of variable V_i are ranked in ascending order; then, each ranked value of V_i is swapped with another ranked value randomly chosen within a restricted range (e.g., the rank of two swapped values can not differ by $>p\%$ of the total number of records). We have used $p = 10\%$.

The results of the masking methods are displayed in Table III. For each method, original names for categories are used when possible. For the sake of conciseness, the following recoding is given in Table III: value 1 stands for coldest, 2 stands for cold, 3 stands for cool, 4 stands for mild, 5 stands for mixed, and 6 stands for hot.

4.3. Reconstructing the Original Matrix

The reconstruction of the 20 records was achieved applying the clustering approach as described in Section 2 and using the semantics for linguistic labels described previously. As in this example, we have considered seven different masking methods, each record has seven different variables and, therefore, is a point in the seven-dimensional space. In the same way, the clusters obtained are regions in this space. In Figure 2 we show the distances between pairs of records.

	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t
a	0	0.083	0	0	0.283	0.263	0.253	0.283	0.283	0.249	0.126	0.083	0	0.185	0	0.185	0	0.176	0.185	0.191
b	0.083	0	0.083	0.083	0.295	0.276	0.266	0.295	0.295	0.231	0.151	0	0.083	0.166	0.083	0.166	0.083	0.199	0.166	0.172
c	0	0.083	0	0	0.283	0.263	0.253	0.283	0.283	0.249	0.126	0.083	0	0.185	0	0.185	0	0.176	0.185	0.191
d	0	0.083	0	0	0.283	0.263	0.253	0.283	0.283	0.249	0.126	0.083	0	0.185	0	0.185	0	0.176	0.185	0.191
e	0.283	0.295	0.283	0.283	0	0.292	0.126	0	0	0.377	0.253	0.295	0.283	0.338	0.283	0.338	0.283	0.222	0.338	0.291
f	0.263	0.276	0.263	0.263	0.292	0	0.318	0.292	0.292	0.188	0.231	0.276	0.263	0.25	0.263	0.25	0.263	0.275	0.25	0.303
g	0.253	0.266	0.253	0.253	0.126	0.318	0	0.126	0.126	0.349	0.283	0.266	0.253	0.314	0.253	0.314	0.253	0.226	0.314	0.262
h	0.283	0.295	0.283	0.283	0	0.292	0.126	0	0	0.377	0.253	0.295	0.283	0.338	0.283	0.338	0.283	0.222	0.338	0.291
i	0.283	0.295	0.283	0.283	0	0.292	0.126	0	0	0.377	0.253	0.295	0.283	0.338	0.283	0.338	0.283	0.222	0.338	0.291
j	0.249	0.231	0.249	0.249	0.377	0.188	0.349	0.377	0.377	0	0.286	0.231	0.249	0.166	0.249	0.166	0.249	0.305	0.166	0.24
k	0.126	0.151	0.126	0.126	0.253	0.231	0.283	0.253	0.253	0.286	0	0.151	0.126	0.224	0.126	0.224	0.126	0.171	0.224	0.229
l	0.083	0	0.083	0.083	0.295	0.276	0.266	0.295	0.295	0.231	0.151	0	0.083	0.166	0.083	0.166	0.083	0.199	0.166	0.172
m	0	0.083	0	0	0.283	0.263	0.253	0.283	0.283	0.249	0.126	0.083	0	0.185	0	0.185	0	0.176	0.185	0.191
n	0.185	0.166	0.185	0.185	0.338	0.25	0.314	0.338	0.338	0.166	0.224	0.166	0.185	0	0.185	0	0.185	0.255	0	0.157
o	0	0.083	0	0	0.283	0.263	0.253	0.283	0.283	0.249	0.126	0.083	0	0.185	0	0.185	0	0.176	0.185	0.191
p	0.185	0.166	0.185	0.185	0.338	0.25	0.314	0.338	0.338	0.166	0.224	0.166	0.185	0	0.185	0	0.185	0.255	0	0.157
q	0	0.083	0	0	0.283	0.263	0.253	0.283	0.283	0.249	0.126	0.083	0	0.185	0	0.185	0	0.176	0.185	0.191
r	0.176	0.199	0.176	0.176	0.222	0.275	0.226	0.222	0.222	0.305	0.171	0.199	0.176	0.255	0.176	0.255	0.176	0	0.255	0.261
s	0.185	0.166	0.185	0.185	0.338	0.25	0.314	0.338	0.338	0.166	0.224	0.166	0.185	0	0.185	0	0.185	0.255	0	0.157
t	0.191	0.172	0.191	0.191	0.291	0.303	0.262	0.291	0.291	0.24	0.229	0.172	0.191	0.157	0.191	0.157	0.191	0.261	0.157	0

Figure 2. Distances for the original records (a representative is given for all records with distance zero).

A dendrogram of the classification is given in Figure 3 (in Figure 3, only one representative is given for all records with distance zero).

With the complete dendrogram, the goal is to find a set of about five clusters so that each cluster corresponds to one of the categories in the aggregated set. To achieve this, two aspects have been considered: (i) records with all the variables with the same value should correspond to different clusters and (ii) clusters should be defined according to the dendrogram. According to Condition i, the records *a*, *n*, and *e* should belong to different clusters. Forcing that *a* and *n* belong to different clusters, the record *r* is left aside because otherwise it does not satisfy Condition ii (see the dendrogram in Figure 3). To sum up, Condition i is the cut A in Figure 3 and Condition ii is translated into the cut B in the same figure (this defines five clusters). It has to be noted that Condition i was not originally required in Radamès, but here it is mandatory to have meaningful results.

Once the clustering process is applied, we need to assign a linguistic label to each class. These categories have to correspond to the ones in the original set of labels. Selection of the category is based on the most used one in the cluster (and giving to each cluster a different value). The last column of Table III shows the selected category for each record.

4.4. Evaluating the Reconstruction

To evaluate the reconstruction, two alternatives have been considered in Section 3: comparison of categorical values and comparison of contingency tables.

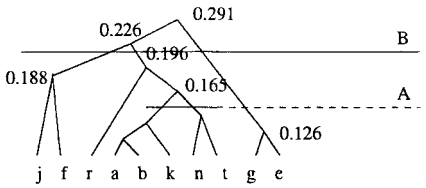

Figure 3. Dendrogram for the clustering of the records in Table III.

In this case, comparing the original values and the aggregated values, we see that only one of the values was not reconstructed correctly. The comparison of the contingency tables will report only the divergence in one position.

A more detailed analysis of the results shows that the only value not correctly reconstructed is the one for record $f = (\&\&44111)$ which is fused as 1 while it should be attached to the value 4. This was caused by presence of several values of 1 in the record and the fact that the distance between this record and the ones with values (&&44444) is larger than the one between the same record and $(\&1n1111)$. In both cases, the nonnumerical values have been considered as unknown values & and because of the fact that unknown values are not considered, the actual calculated distances were $d((44111), (44444))$, and $d((4111), (1111))$.

The analysis of the data also shows that record $r = (5\&n5334)$ has been fused as the value 5. It is interesting to note that this value is equal to the original one. This value is obtained even though most of the values are different from the value 5. This is caused by the approach of applying clustering to the data. In this case, this record is different enough to all the other records to become a different cluster.

5. CONCLUSIONS AND FUTURE WORK

In this work, we have presented a method to aggregate data from multiple sources and applied it to the statistical disclosure risk. We have shown that multiple protected releases of the same data can provoke disclosure when recipients of the data make coalitions.

The approach introduced here is suitable for categorical data and builds the aggregated value through the application of clustering. In Section 4 we have given an example of the application of the system to a set of 20 records.

We have shown that the reconstruction method introduced here permits the reconstruction of the original data from seven different releases obtained through the application of seven different masking methods.

The system described for data fusion does not satisfy the condition of irrelevant alternatives. This means that the aggregated value for a given record depends on the other records. In the example presented in Section 4, record r could lead to another result if records similar to it would be similar also to record a. For example, the presence of a record equal to (53334) would cause r to be related to this one and this one would be related to a. Therefore, a and r would be put together in the same cluster. However, in this case, record r is labeled correctly although there are only two values of 5 in it.

An advantage of the approach is that the method can not only be applied when the set of categories for all masked variables is equal, but also when each method applies some recoding that makes it difficult to know the original set of categories. For example, this would be the case if $G4$ codifies $1, 2, \ldots, 5$ by $\alpha, \beta, \ldots, \gamma$ (thus changing the range of the variable). In such a situation, the final values obtained by our method would be the same because it does not depend on the values themselves but on the clustering process. Note that this would not be the case with other

aggregation methods (as the voting technique) and is a clear advantage of the approach.

As a future work, we consider the need to include knowledge in the aggregation process. At this point we have not included any knowledge about the masking methods and how they distort the data. Including this knowledge would improve the capabilities of the system.

Also, further analysis of the method and extensive testing with larger sets is needed. Additionally, the approach considered here would be of interest when combined with reidentification procedures. Combination of data with similar characteristics would improve the reidentification scores.

Finally, we consider the use of techniques for determining the set of adequate categories to be used to describe the aggregated variable. This is not required in this case because all masked variables use the same set of categories but it would be appropriate in the foregoing example if $G4$ uses α, $\beta, \ldots,$ γ instead of 1, 2, . . . , 5 after the application of the masking method. To work in this direction, we plan to follow the approach outlined in Ref. 20.

Acknowledgments

We acknowledge partial support of the European Community under the contract "CASC" IST-2000-25069 and of the CICYT under the project "STREAMOBILE" (TIC2001-0633-C03-01/02).

References

1. Doyle P, Lane JI, Theeuwes JJM, Zayatz LM. Confidentiality, disclosure, and data access: Theory and practical applications for statistical agencies. New York: Elsevier; 2001.
2. Torra V, Domingo-Ferrer J. Record linkage methods for multidatabase mining. In: Torra V, editor. Information fusion in data mining. New York: Springer; 2003.
3. Domingo-Ferrer J, Torra V. Disclosure control methods and information loss for microdata. In: Doyle P, Lane JI, Theeuwes JJM, Zayatz LM, editors. Confidentiality, disclosure, and data access: Theory and practical applications for statistical agencies. New York: Elsevier; 2001. pp 91–110.
4. Willenborg L, De Waal T. Elements of statistical disclosure control. New York: Springer-Verlag; 2001.
5. Yager RR. On ordered weighted averaging aggregation operators in multi-criteria decision making. IEEE Trans Syst Man Cybern 1998;18:183–190.
6. Torra V. The weighted OWA operator. Int J Intell Syst 1997;12:153–166.
7. Dubois D, Koning J-L. Social choice axioms for fuzzy set aggregation. Fuzzy Sets Syst 1991;43:257–274.
8. Torra V, Cortés U. Towards an automatic consensus generator tool: EGAC. IEEE Trans Syst Man Cybern 1995;25(5):888–894.
9. Valls A, Riaño D, Torra V. Sedàs: A semantic based general classifier system. Math Soft Comput 1997;4:267–279.
10. de Soto AR, Trillas E. On antonym and negate in fuzzy logic, Intl J Intell Syst 1999; 14(3):295–303.
11. Bonissone PP, Decker KS. Selecting uncertainty granularity: An experiment in trading-off precision and complexity. In: Kanal LH, Lemmer JF, editors. Uncertainty in artificial intelligence. Amsterdam: North-Holland; 1986. pp 217–247.

12. Herrera F, Herrera-Viedma E, Verdegay JL. A sequential selection process in group decision making with a linguistic assessment approach. Int J Inf Sci 1995;80:223–239.
13. Herrera F, Herrera-Viedma E, Verdegay JL. A model of consensus in group decision making under linguistic assessments. Fuzzy Sets Syst 1996;78:73–87.
14. Agustí J, Esteve F, Garcia P, Godo L, Sierra C. Combining multiple-valued logics in modular expert systems. In: Proc 7th Conf on Uncertainty in AI, Los Angeles, CA, July, 1991. pp 17–29.
15. Torra V. Negation functions based semantics for ordered linguistic labels. Int J Intell Syst 1996;11:975–988.
16. Valls A, Torra V. On the semantics of qualitative attributes in knowledge elicitation. Int J Intell Syst 1999;14(2):195–209.
17. http://www.census.gov/DES/www/welcome.html.
18. Domingo-Ferrer J, Torra V. A quantitative comparison of disclosure control methods for microdata. In: Doyle P, Lane JI, Theeuwes JJM, Zayatz LM, editors. Confidentiality, disclosure, and data access: Theory and practical applications for statistical agencies. New York: Elsevier; 2001. pp 111–133.
19. Kooiman P, Willenborg L, Gouweleeuw J. PRAM: A method for disclosure limitation of microdata. Research Report, Voorburg NL: Statistics Netherlands; 1998.
20. Valls A, Torra V. Explaining the consensus of opinions with the vocabulary of the experts. In: Proc of the Int Conf IPMU 2000, Madrid, Spain, 2000. pp 746–753.