SORT 27 (2) July-December 2003, 175-190

A posteriori disclosure risk measure for tabular data based on conditional entropy**

Anna Oganian and Josep Domingo-Ferrer[a*]

[a] *Universitat Rovira i Virgili, Spain*

Abstract

Statistical database protection, also known as Statistical Disclosure Control (SDC), is a part of information security which tries to prevent published statistical information (tables, individual records) from disclosing the contribution of specific respondents. This paper deals with the assessment of the disclosure risk associated to the release of tabular data. So-called sensitivity rules are currently being used to measure the disclosure risk for tables. These rules operate on an *a priori* basis: the data are examined and the rules are used to decide whether the data can be released as they stand or should rather be protected. In this paper, we propose to complement *a priori* risk assessment with *a posteriori* risk assessment in order to achieve a higher level of security, that is, we propose to take the protected information into account when measuring the disclosure risk.

The proposed *a posteriori* disclosure risk measure is compatible with a broad class of disclosure protection methods and can be extended for computing disclosure risk for a set of linked tables. In the case of linked table protection via cell suppression, the proposed measure allows detection of secondary suppression patterns which offer more protection than others.

MSC: 62P99

Keywords: statistical disclosure control, statistical databases, tabular data, security

1 Introduction

Statistical database protection is a part of information security called inference control in the classical book by Denning (1982). The most typical output offered by national statistical agencies is tabular data. Tables are central in official statistics: many survey and census data are categorical in nature, so that their representation as cross-classifications or tables is a natural reporting strategy. Tabular data being thus aggregate

** Work partly funded by the European Union under project "CASC" IST-2000-25069. The first author is supported by a grant from Generalitat de Catalunya.

* *Address for correspondence:* {aoganian, jdomingo}@etse.urv.es. Universitat Rovira i Virgili, Dept. of Computer Eng. and Maths, Av. Països Catalans 26, E-43007 Tarragona, Catalonia, Spain.

Received: April 2003

Accepted: September 2003

data, one is tempted to think they are not supposed to contain information that can reveal the contribution of particular respondents. However, as noted in Giessing (2001), in many cases table cells do contain information on a single or very few respondents, which implies a disclosure risk for the data of those respondents. In these cases, disclosure control methods must be applied to the tables prior to their release.

A number of disclosure control methods to protect tabular data have been proposed (see Willenborg and de Waal (2001), Duncan, Fienberg, Krishnan, Padman, and Roehrig (2001) for a survey). Next we list the main principles underlying those methods:

Cell suppression If a table cell is deemed sensitive, then it is suppressed from the released table (primary suppression). If marginal totals or other linked tables are also to be published, then it may be necessary to remove additional table cells (secondary suppressions) to prevent primary suppressions from being computable. Secondary suppressions should be chosen in a way such that the utility of the resulting table is maximized.

Rounding A positive integer b (rounding base) is selected and all table cells are rounded to an integer multiple of b. Controlled rounding is a variant of rounding in which table additivity is preserved (*i.e.* rounded rows and columns still sum to their rounded marginals).

Table redesign Categories used to tabulate data are recoded into different (often more general) categories, so that the resulting tabulation does not contain sensitive cells any more. A simple redesign could be to combine two rows containing sensitive cells to obtain a new row without sensitive cells.

Sampling A table is released which is based on a sample of the units on which the original table was built.

Swapping and simulation In data swapping, units are swapped so that the table resulting from the swapped data set still preserves all k-dimensional margins of the original table. A more elaborate version of swapping was proposed in Fienberg, Makov and Steele (1998), whereby the original table is replaced by a random draw from the exact distribution under the log-linear model whose minimal sufficient statistics correspond to the margins of the original table. Further extensions of this idea would lead to drawing a synthetic table from the full distribution of all possible tables with the same margins of the original table.

As noted by Duncan, Fienberg, Krishnan, Padman, and Roehrig (2001), any attempt to compare methods for tabular data protection should focus on two basic attributes:

1. *Disclosure risk:* a measure of the risk to respondent confidentiality that the data releaser (typically a statistical agency) would experience as a consequence of releasing the table.
2. *Data utility:* a measure of the value of the released table to a legitimate data user.

In this paper, we concentrate on the assessment of disclosure risk. Up to now, disclosure risk assessment for tables was usually performed *a priori*, that is, before applying any protection methods to the table. The standard approach is to use a *sensitivity rule* to decide whether a particular table cell can safely be released.

However, *a priori* measures do not actually measure the disclosure risk incurred once a particular table is released. In this paper, we propose to complement *a priori* risk assessment with a *posteriori* risk assessment, which takes protected information into account. The proposed measure applies to a broad class of disclosure protection methods and is computable in practice.

Section 2 describes existing disclosure risk measures, which are *a priori* by their nature. In Section 3, an *a posteriori* measure based on the reciprocal of conditional entropy is proposed as a complement to *a priori* measures. Section 4 describes an application of the proposed *a posteriori* measure to different table protection methods, both for simple tables and for linked tables. In the case of cell suppression methods, the proposed measure turns out to be useful to detect suppression patterns which offer more protection than others. Section 5 is a conclusion.

2 Background on *a priori* disclosure risk measures

A priori disclosure risk measures used by statistical agencies for tabular data protection are also called sensitivity rules. For magnitude tables (normally related to economic data), there are two widely accepted sensitivity rules:

$n-k$-dominance In this rule, n and k are two parameters with values to be specified. A cell is called sensitive if the sum of the contributions of n or fewer respondents represents more than a fraction k of the total cell value.

pq-rule The prior-posterior rule is another rule gaining increasing acceptance. It also has two parameters p and q. It is assumed that, prior to table publication, each respondent can estimate the contribution of each other respondent to within less than q percent. A cell is considered sensitive if, posterior to the publication of the table, someone can estimate the contribution of an individual respondent to within less than p percent. A special case is the p%-rule: in this case, no knowledge prior to table publication is assumed, *i.e.* the pq-rule is used with $q = 100$.

For tables of counts or frequencies (normally related to demographic data), a so-called **threshold rule** is used. A cell is defined to be sensitive if the number of respondents is less than a threshold k.

These sensitivity rules have received critiques for failing to adequately reflect the risk of disclosure, but these were mostly limited to numerical counterexamples for particular choices of the parameters of these rules. Recently, it was shown in Domingo-Ferrer

and Torra (2002) through general counterexamples that releasing a cell declared non-sensitive by these rules can imply higher disclosure risk than releasing a cell declared sensitive. It was proposed to use Shannon's entropy of relative contributions to a table cell as a better alternative to (n, k)-dominance, pq-rule and $p\%$-rule. Formally speaking,

$$H(X) = -\sum_{i=1}^{N} (x_i/x) \log_2 (x_i/x) \tag{1}$$

where $x = x_1 + x_2 + \cdots + x_N$ is the value of a table cell and xi are contributions to that cell.

A cell is considered sensitive by the above rule if $H(X)/\log_2 N < t$, where $t \in [0, 1]$ is a parameter; otherwise, the cell is declared non-sensitive.

3 An *a posteriori* disclosure risk measure based on conditional entropy

We have mentioned above that using only *a priori* measures may be insufficient for table protection. Now we want to illustrate this on the following examples.

Example 1 *Suppose that the person or entity who wants to guess secret information about how much a particular respondent contributed to some cell of the table is someone who also contributed to that cell. So he obviously knows his own contribution to that cell. He also may know some additional information, for example, how many respondents have contributed to that cell, who they are, etc. This internal intruder is in a better position than an outsider to estimate the contribution of his interest. This kind of information is not taken into account by a priori measures. According to these, the disclosure risk is the same for all types of intruders and that is not true.*

The information held by an intruder does not only depend on her being internal or external; it clearly depends also on what information has previously been published and on how that information has been protected.

Example 2 *Assume we have an n-dimensional table whose cells are deemed sensitive, and therefore cannot be released. Only some 2-dimensional (or $(n - i)$-dimensional) tables are released, which have been obtained as projections of the n-dimensional table. Due to their origin, the released tables are linked tables, so the uncertainty about a cell value in the n-dimensional table is conditional to the particular tables released so far.*

Therefore, we propose to complement *a priori* risk assessment provided by sensitivity rules with *a posteriori* risk assessment. The latter is performed *after* data have been protected and takes protected data into account to compute bounds on cells labeled as sensitive by a sensitivity rule.

Our proposal for *a posteriori* measure is to use the reciprocal of Shannon's conditional entropy [Shannon (1948)] to express the disclosure risk in a natural and unified manner.

Entropy-based measures were already discussed in Willenborg and de Waal (2001) for computing information loss at the table level, but not for computing disclosure risk. However, the authors of Willenborg and de Waal (2001) do not believe entropy is a practical information loss measure. We support their opinion with the following example.

Example 3 *Assume we use rounding with integer base b to protect a table. The entropy-based information loss measure defined in Willenborg and de Waal (2001) is the reciprocal of the number of original tables whose rounded version matches the published rounded table (i.e. the number of original tables "compatible" with the published one). The number of compatible tables depends on the rounding base b, but is independent on how close the published rounded values are to the original values. Thus, the entropy-based information loss measure is the same when the original table exactly corresponds to the rounded table (which happens when all cell values in the original table are multiples of b) and when all differences between corresponding cell values in the original and rounded tables are close to b/2. This does not seem to adequately reflect the utility of the published data.*

In Duncan, Fienberg, Krishnan, Padman, and Roehrig (2001), the reciprocal of Shannon's entropy (not conditional entropy) was suggested as measure of disclosure risk at the cell level. What was not clear there is how to compute the probabilities, that is, what distribution should be chosen. In fact, as we noted above, the particular distribution for an intruder depends on the knowledge of that intruder.

The above discussion suggests that the most natural *a posteriori* measure for disclosure risk is the reciprocal of conditional entropy

$$DR(X) = 1/H(X|Y = y) = 1/\left(-\sum_x p(x/y)\log_2 p(x/y)\right) \tag{2}$$

where X is a variable representing an original cell and Y is a variable representing the intruder's knowledge (which is supposed to be equal to some specific value y). The intuitive notion behind Expression (2) is that, the more uncertainty about the value of the original cell X (which depends on the constraints $Y = y$), the less disclosure risk (and conversely).

There are two practical problems in computing Expression (2):

1. Finding the set $S_y(X)$ of possible values of X given the constraints y.
2. Estimating the probabilities $p(x|y)$, *i.e.* the probability of the cell X being x given that Y is y.

As noted by Willenborg and de Waal (2001) when discussing entropy-based information loss measures, taking the uniform probability distribution over the set $S_y(X)$ can make sense for some disclosure control methods. Using the uniform distribution, Expression (2) is simplified to

$$DR_{unif}(X) = 1/\log_2 m\left(S_y(X)\right) \tag{3}$$

where $m\left(S_y(X)\right)$ is the number of possible values of the cell in $S_y(X)$.

Table 1: *A table with suppressed cells.*

Economic activity	*Size class*					*Total*
	4	5	6	7	8	
2,3	80	253	54	0	0	387
4	641	3694	2062	746	0	7143
5	592	x_1	329	x_2	1440	3898
6	57	x_3	946	x_4	2027	4281
7	78	0	890	1719	1743	4430
Total	1148	4353	4281	4847	5210	20139

Note 1 (On $m(S_y(X))$) *We assume in what follows that table cells take values in a discrete domain: either integer values or real values with a fixed number of decimal positions. This is the usual case in published statistical tables: count tables consist of integer values and magnitude tables consist of either integer values or real values with limited precision. Thus the set $S_y(X)$ of possible values is enumerable and it makes sense to speak of $m\left(S_y(X)\right)$ as the number of cell values in $S_y(X)$.*

4 Application to several table protection scenarios

We show in this Section how to compute Expression (3) for several disclosure control methods for tables; the case of linked tables will also be discussed.

4.1 Cell suppression

The disclosure risk computation for cell suppression is illustrated by extending an example provided in Willenborg and de Waal (2001). Let Table 1 be a table from which four cells x_1, x_2, x_3 and x_4 have been suppressed. Assume that the suppressed values are integer.

According to the definition given in Section 3, the disclosure risk for each suppressed cell is the reciprocal of one of the following conditional entropies:

$$\begin{aligned} &H(x_1|x_1 + x_2 = 1537,\ x_1 + x_3 = 406,\ x_i \geqslant 0)\\ &H(x_2|x_1 + x_2 = 1537,\ x_2 + x_4 = 2382,\ x_i \geqslant 0)\\ &H(x_3|x_1 + x_3 = 406,\ x_3 + x_4 = 1251,\ x_i \geqslant 0)\\ &H(x_4|x_2 + x_4 = 2382,\ x_3 + x_4 = 1251,\ x_i \geqslant 0) \end{aligned} \tag{4}$$

Table 2: *A table with two rows combined.*

Economic activity	*Size class*					*Total*
	4	5	6	7	8	
2,3	80	253	54	0	0	387
4	641	3694	2062	746	0	7143
5,6	649	406	1275	2382	3467	8179
7	78	0	890	1719	1743	4430
Total	1148	4353	4281	4847	5210	20139

Expressions (4) contain constraints yi for each suppressed cell xi which allow $m(S_{yi}(x_i))$ to be computed by solving two linear programming (LP) problems (one maximization and one minimization) and subtracting the solutions. In the case of Table 1, minimizations and maximizations bound every cell as follows: $0 \leqslant x_1 \leqslant 406$, $1131 \leqslant x_2 \leqslant 1537$, $0 \leqslant x_3 \leqslant 406$ and $845 \leqslant x_4 \leqslant 1251$. By subtracting the bounds we obtain $m(S_{yi}(x_i)) = 407$ for $i = 1, 2, 3, 4$. Using Expression (3), we can compute $DR_{unif}(x_i) = 1/\log_2 407 = 0.115$ for every cell.

4.2 Rounding

When the table is protected by rounding, the cell entropy conditional to the rounded table depends on the rounding base b. In a rounded table without marginals, if the value of a cell x_i' is $n_i b$ (*i.e.* n_i times the rounding base), then we know that the original cell x_i must lie in the interval $I_i = [(n_i - 1/2)b, (n_i + 1/2)b)$. Thus, $DR_{unif}(x_i) = 1/\log_2 m(I_i)$, where $m(I_i)$ is the number of possible cell values in I_i (keep in mind that cell values are either integer or with a fixed number of decimal positions).

4.3 Table redesign

This case is very similar to cell suppression. Imagine that the sensitive cells in Table 1 are protected by combining rows with *Economic_activity* = 5 or 6. This yields Table 2.

Let us label the six cells in the original row with *Economic_activity* = 5 as x_1 through x_6 and the six cells in *Economic_activity* = 6 as x_7 through x_{12} (x_6 is the marginal of the first row and x_{12} is the marginal of the second row).

Then the following equalities hold:

$$\begin{aligned} x_1 + x_2 + x_3 + x_4 + x_5 - x_6 &= 0 \\ x_7 + x_8 + x_9 + x_{10} + x_{11} - x_{12} &= 0 \\ x_1 + x_7 &= 649 \\ x_2 + x_8 &= 406 \\ x_3 + x_9 &= 1275 \\ x_4 + x_{10} &= 2382 \\ x_5 + x_{11} &= 3467 \\ x_6 + x_{12} &= 8179 \\ x_i &\geqslant 0 \quad \text{for} \quad i = 1, \ldots, 12 \end{aligned} \tag{5}$$

From the above, $m(S_{yi}(x_i))$ and $DR_{unif}(x_i)$ are computed in a way analogous to the case of cell suppression.

4.4 Linked tables

We will show the application of conditional entropy as *a posteriori* disclosure risk measure for linked tables with an example.

Let us consider the three-dimensional table ASR formed by cells $z_{a_i s_j r_k}$, where each cell denotes the total turnover of businesses with activity a_i and size s_j in region r_k. Assume that table ASR is not released because every cell in it is considered sensitive. Instead of ASR, some of the following tables obtained by bidimensional projection are released: $AS = \{z_{a_i s_j}\}$, which breaks down turnover by activity and business size, $AR = \{z_{a_i r_k}\}$, which breaks down turnover by activity and region, and $SR = \{z_{s_j r_k}\}$, which breaks down turnover by size and region. Assume three scenarios: 1) only AS is released; 2) AS and AR are released; 3) AS, AR and SR are released. The disclosure risk of cell $z_{a_i s_j r_k}$ in each scenario can be expressed as:

$$DR_{unif}\left(z_{a_i s_j r_k} | AS\right) = 1/H\left(z_{a_i s_j r_k} | z_{a_i s_j} = \sum_k z_{a_i s_j r_k}\right) \tag{6}$$

$$\begin{aligned} &DR_{unif}\left(z_{a_i s_j r_k} | AS, AR\right) \\ &= 1/H\left(z_{a_i s_j r_k} | z_{a_i s_j} = \sum_k z_{a_i s_j r_k}, z_{a_i r_k} = \sum_j z_{a_i s_j r_k}\right) \end{aligned} \tag{7}$$

$$DR_{unif}\left(z_{a_i s_j r_k} | AS, AR, SR\right)$$
$$= 1/H\left(z_{a_i s_j r_k} | z_{a_i s_j} = \sum_k z_{a_i s_j r_k}, z_{a_i r_k} = \sum_j z_{a_i s_j r_k}, z_{s_j r_k} = \sum_i z_{a_i s_j r_k}\right) \quad (8)$$

The released tables impose constraints on the possible cell values of the table ASR. Such constraints actually determine the simplexes $S_{AS}(z_{a_i s_j r_k})$, $S_{AS,AR}(z_{a_i s_j r_k})$ or $S_{AS,AR,SR}(z_{a_i s_j r_k})$ where $z_{a_i s_j r_k}$ should lie. By solving one LP maximization and one LP minimization for each $z_{a_i s_j r_k}$, an interval where the cell lies can be determined. Then, the cell disclosure risk is computed using Expression (3). If a cell is too closely bounded, then its disclosure risk is too high and disclosure control methods must be used.

When the disclosure control method chosen is cell suppression, it is important to notice that linked tables have the property that there are sets of linearly dependent constraints, so that one of the constraints in each such set may be suppressed without decreasing the rank of the whole constraint system. This will influence the quality of the protection offered by different suppression patterns: the best pattern is the one decreasing most the system rank, which results in more degrees of freedom, and thus more cell entropy and lower disclosure risk. We show this with in the following section.

Table 3: *Constraint matrix imposed by Table AS. Here* $i \in \{1, \ldots, 3\}$, $j \in \{1, \ldots, 4\}$, $k \in \{1, \ldots, 3\}$.

1	1	1																$z_{a_1 s_1}$
			1	1	1													$z_{a_1 s_2}$
						1	1	1										$z_{a_1 s_3}$
									1	1	1							$z_{a_1 s_4}$
																		$\vdots$
															1	1	1	$z_{a_3 s_4}$

4.5 Minimizing disclosure risk in linked table release

For the sake of concreteness, we will resume the example of three linked tables used in the previous section. We want to estimate the disclosure risk of cells in ASR depending on the released tables. Assume that $i \in \{1, \ldots, n\}$, $j \in \{1, \ldots, m\}$ and $k \in \{1, \ldots, t\}$.

4.5.1 One table released out of three linked tables

If only the table AS is released, the constraint matrix is shown in Table 3, where there is a column for each cell $z_{a_i s_j r_k}$. The matrix rank is nm, as all equations are linearly independent. Every choice for secondary suppressions causes the same rank decrease and consequently has an equivalent impact on the disclosure risk. Therefore, there is no

room for optimization (unless there are specific additional constraints specified by the data protector).

4.5.2 Two tables released out of three linked tables

If tables AS and AR are released, the constraint matrix is shown in Table 4, where there is a column for each cell $z_{a_i s_j r_k}$. This matrix consists of n submatrices of size $(t+m)\times mt$ with rank $(t+m-1)$, that is, one constraint in each submatrix is a linear combination of the remaining constraints in the submatrix. That is, using Gaussian elimination we have:

$$\sum_{k=1}^{t} z_{a_i r_k} - \sum_{j=1}^{m} z_{a_i s_j} = 0 \quad \text{for} \quad 1 \leqslant i \leqslant n \tag{9}$$

Therefore, n constraints can be suppressed and the matrix rank will not change, nor will change $H(z_{a_i s_j r_k}|AS, AR)$ nor the disclosure risk. Note that only one row per submatrix can be suppressed for disclosure risk to stay unchanged, which, in terms of tables AS and AR, means only one cell per two corresponding rows in AS and AR (*e.g.* for the submatrix related to Expression (9), the two rows are those in AS and AR with subscript a_i).

From the above discussion, we can state the following proposition:

Proposition 1 *When two out three linked tables are released, the entropy increase and the disclosure risk decrease are maximized if the suppression patterns are chosen so that the secondary suppressions are in the same columns and rows for both released tables.*

Proof. Assume that tables AS and AR are released. Now assume that $z_{a_i s_j r_k}$ in supertable ASR has a high disclosure risk which makes it necessary to increase its entropy. So, if a cell suppression is used, a natural option is to suppress the cells in the released tables which refer to $z_{a_i s_j r_k}$. These suppressions will be called primary suppressions and will be $z_{a_i s_j}$ and $z_{a_i r_k}$ in the tables AS and AR, respectively. The suppression of these two cells will decrease the rank by 1 (see the discussion above).

Secondary suppressions will be the following:

- Two cells, say $z_{a_i s_l}$ and $z_{a_f s_j}$ in the table AS in order to prevent $z_{a_i s_j}$ from being computable (in what follows, we will say —"to protect $z_{a_i s_j}$"), which decrease the rank by 1,
- The cell $z_{a_f s_l}$ in AS to protect $z_{a_f s_j}$, which decreases the rank by 1.
- Two cells in the table AR: one in the row a_i and other in the column r_k. If $z_{a_i r_v}$ is the cell in row a_i, its suppression decreases the rank by 1. When we choose the candidate for the suppression in column r_k, we should take into account that, if we choose the cell in a row other than a_f, this will not decrease the rank (because the rows where we have already suppressed cells are a_i and a_f). So, in order to

increase the entropy and to decrease disclosure risk, we have to choose row a_f, that is we have to suppress the cell $z_{a_f r_k}$.

- Finally, in order to protect $z_{a_f r_k}$, we have to suppress $z_{a_f r_v}$, which will decrease the rank by 1.

So, from the above argument, we have that choosing the same rows for secondary suppressions in both tables will decrease the rank by 1 more than if we chose different rows. Therefore, this is the strategy to maximize rank decrease, which is equivalent to to maximizing entropy increase and disclosure risk decrease. QED

Table 4: *Constraint matrix imposed by tables AS and AR. Here* $i \in \{1,\ldots,3\}, j \in \{1,\ldots,4\}, k \in \{1,\ldots,3\}$.

1	1	1																	$z_{a_1 s_1}$
1			1			1			1										$z_{a_1 r_1}$
	1			1			1			1									$z_{a_1 r_2}$
		1			1			1			1								$z_{a_1 r_3}$
			1	1	1														$z_{a_1 s_2}$
						1	1	1											$z_{a_1 s_3}$
									1	1	1								$z_{a_1 s_4}$
													⋮					⋮	
																1	1	1	$z_{a_3 s_4}$

4.5.3 All three linked tables released

Now let us consider the case where three tables AS, AR, SR are released. Again, using Gaussian elimination, we obtain several sets of constraints which are linearly dependent (exactly one constraint in each set is a linear combination of the rest). The linear combinations are as follows:

$$\sum_{k=1}^{t} z_{a_i r_k} - \sum_{j=1}^{m} z_{a_i s_j} = 0 \quad \text{for} \quad 1 \leqslant i \leqslant n \tag{10}$$

$$\sum_{k=1}^{t} z_{s_j r_k} - \sum_{j=1}^{n} z_{a_i s_j} = 0 \quad \text{for} \quad 1 \leqslant j \leqslant s-1 \tag{11}$$

$$\sum_{j=1}^{m} z_{s_j r_k} - \sum_{i=1}^{n} z_{a_i r_k} = 0 \quad \text{for} \quad 1 \leqslant k \leqslant t-1 \tag{12}$$

$$z_{s_m r_t} - \sum_{j=1}^{m-1}\sum_{k=1}^{t-1} z_{s_j r_k} + \sum_{i=1}^{n}\sum_{j=1}^{m-1} z_{a_i s_j} - \sum_{i=1}^{n} z_{a_i r_t} = 0 \tag{13}$$

Furthermore, up to $n+m+t-1$ cells can be suppressed without changing the entropy nor the disclosure risk. But this is heavily dependent on which cells are suppressed. A suppression pattern of maximal size which does not change the rank of the system may be the following: $z_{a_1 s_m}$, $z_{a_2 s_m}, \ldots, z_{a_n s_m}$, $z_{s_1 r_t}$, $z_{s_2 r_t}, \ldots, z_{s_{m-1} r_t}$, $z_{s_m r_1}$, $z_{s_m r_2}, \ldots, z_{s_m r_t}$. Now, assume that the cell $z_{a_1 s_1 r_1}$ has za1s1 = zs a a high disclosure risk which makes it necessary to increase its entropy. If cell suppression is used, a natural option is to suppress every cell in the three tables which refers to $z_{a_1 s_1 r_1}$. These suppressions will be called primary suppressions and will be $z_{a_1 s_1}$ from Table AS, $z_{a_1 r_1}$ from Table AR and $z_{s_1 r_1}$ from Table SR. Note that with these suppressions the rank of the system will decrease by 1. If $z_{a_1 s_1}$ is suppressed, the rank does not decrease because, by Expression (11), the suppressed cell is a linear combination

$$z_{a_1 s_1} = \sum_{k=1}^{t} z_{s_1 r_k} - \sum_{i \neq 1} z_{a_i s_1} \tag{14}$$

If $z_{a1_r 1}$ is suppressed next, the rank does not change either. By Expression (12), we can express the suppressed cell as a linear combination of cells which have not yet been suppressed:

$$z_{a_1 r_1} = \sum_{j=1}^{m} z_{s_j r_1} - \sum_{i \neq 1} z_{a_i r_1} \tag{15}$$

If $z_{s_1 r_1}$ is our third suppression, then the rank will decrease by 1, because that cell appears in Equations (14) and (15) and it is easy to see that there is no way to use the above equations to express that cell as a combination of the cells which have not yet been suppressed.

If the table is released with marginals, then a set of secondary suppressions is required to prevent primary suppressions from being computable. At this moment, it is important to choose the necessary strategy for secondary suppressions because the rank of the system and consequently the cell entropy will vary depending on what secondary suppressions are made. Let us analyze what happens with the secondary suppressions corresponding to each primary suppression:

1. Assume that, to protect $z_{a_1 s_1}$, we choose as secondary suppressions $z_{a_1 s_3}$, $z_{a_3 s_1}$ and $z_{a_3 s_3}$. Suppressing $z_{a_1 s_3}$ does not change the rank, because by Equation (11) we have

$$z_{a_1 s_3} = \sum_{k=1}^{t} z_{s_3 r_k} - \sum_{i \neq 1} a_{a_i s_3} \tag{16}$$

where the cells on the right-hand side have not yet been suppressed. Suppressing $z_{a_3 s_1}$ does not change the rank either, because by Equation (10):

$$z_{a_3 s_1} = \sum_{k=1}^{t} z_{a_3 r_k} - \sum_{j \neq 1} z_{a_3 s_j} \tag{17}$$

Suppression of $z_{a_3 s_3}$ causes the rank to decrease by 1, because there is no way to express the suppressed cell as combination of other cells which have not yet been suppressed: if Equation (11) is used, $z_{a_1 s_3}$ is necessary but has been suppressed already; if Equation (10) is used, then $z_{a_3 s_1}$ is necessary which has been suppressed; for a similar reason we cannot use Equation (13). Note also that, once the suppression process has started, Equation (13) is not very useful to obtain linear combinations, because it depends on nearly all cells.

2. Now assume that, to protect $z_{a_1 r_1}$, we choose as secondary suppressions $z_{a_1 r_3}, z_{a_2 r_1}$ and $z_{a_2 r_3}$. When $z_{a_1 r_3}$ is suppressed, the rank does not change, because by Equation (12)

$$z_{a_1 r_3} = \sum_{j=1}^{m} z_{s_j r_3} - \sum_{i \neq 1} z_{a_i r_3} \tag{18}$$

Suppressing $z_{a_2 r_1}$ does not change the rank either, because by Equation (10):

$$z_{a_2 r_1} = \sum_{j=1}^{m} z_{a_2 s_j} - \sum_{k \neq 1} z_{a_2 r_k} \tag{19}$$

Note that, if we now choose to suppress $z_{a_3 r_1}$, the rank would decrease by 1, because there is no way to express it as a combination of other cells not yet suppressed (if Equation (12) was used, $z_{a_1 r_1}$ would be necessary and, if Equation (10) was used, then $z_{a_3 s_1}$ would be necessary). So, *choosing for this suppression any row in Table AR other than row 3 (which was used in Table AS) does not decrease the rank and, consequently, adds hardly any protection*. Finally, using similar arguments, it is not difficult to see that suppression of $z_{a_2 r_3}$ decreases the rank by 1 (this suppression is inevitable in order to protect $z_{a_2 r_1}$ and $z_{a_1 r_3}$).

3. As to the third primary suppression, assume that, to protect $z_{s_1 r_1}$, we choose as secondary suppressions $z_{s_1 r_2}$, $z_{s_4 r_1}$ and $z_{s_4 r_2}$. Suppressing $z_{s_1 r_2}$ does not change the rank, because, by Equation (12):

$$z_{s_1 r_2} = \sum_{i=1}^{n} z_{a_i r_2} - \sum_{j \neq 1} z_{s_j r_2} \tag{20}$$

Note that, if we chose $z_{s_1 r_3}$, the rank would decrease by 1. *So, choosing for this suppression any column other than column 3 (which was used for Table AR) does not decrease the rank and adds virtually no protection*. Suppressing $z_{s_4 r_1}$ does not change the rank either, because, by Equation (11):

$$z_{s_4 r_1} = \sum_{i=1}^{n} z_{a_i s_4} - \sum_{k \neq 1} z_{s_4 r_k} \tag{21}$$

If we chose $z_{s_3 r_1}$ instead of $z_{s_4 r_1}$, then the rank would decrease by 1. *So, choosing for this suppression any row other than row 3 (which was used for Table AS when $z_{a_3 s_1}$ was suppressed) does not decrease the rank and adds no real protection*. Finally, we have to suppress $z_{s_4 r_2}$, which causes the rank to decrease by 1.

We have performed 12 suppressions altogether (primary and secondary), which decrease the rank of the system by 4. From the above, it is clear that the strategy we followed was to choose the suppression pattern which minimizes the decrease of the system rank and consequently minimizes the increase of entropy (and of protection!).

From the previous discussion, we can infer the following general result, whose proof is analogous to the proof of Proposition 1:

Proposition 2 *The entropy increase and the disclosure risk decrease in three linked tables are maximized if the suppression patterns are chosen so that the secondary suppressions are in the same columns and rows for all three tables.*

A weaker necessary condition for maximizing the decrease of disclosure risk in the case of three tables is as follows:

Proposition 3 *Proposition 3 The entropy increase and the disclosure risk decrease in three linked tables are maximized if the suppression patterns are chosen so that the secondary suppressions are in the same row for tables with the same first variable —e.g. AS, AR—, the same column for tables with the same second variable —e.g. AR, SR—, and the same row for tables which share a variable in a different position —e.g. AS, SR—).*

Using such optimal patterns we can decrease the rank of the system of 3 linked tables by 3 more units (up to an overall rank decrease of 7).

4.5.4 Disclosure risk by internal intruders

Finally, a point we have to take into account here is that disclosure risk is different for different users. Let us imagine that, when solving one LP maximization and one LP minimization for $z_{a_i s_j r_k}$, we find that $995 \leqslant z_{a_j s_j r_k} \leqslant 1004$. Then, if company A is the second largest contributor to this cell with a turnover of, say, 400, then company A knows that the largest contributor (company B) has a turnover between 401 and 604. Thus, company A is able to estimate the turnover of company B within 50% of its value. However, the uncertainty of an external intruder about the turnover of company B is roughly 200% of its value: the external intruder only knows that the turnover of the largest contributor is between $\varepsilon > 0$ and 1004. Therefore, for an internal intruder (respondent contributing to the cell), the measure of disclosure risk $1/H(z_{a_i s_j r_k}|\text{released tables})$ should be replaced by:

$$DR(X) = 1/H(z_{a_i s_j r_k}|\text{released tables, intruder's contribution}) \tag{22}$$

5 Conclusion

Due to the limitations of *a priori* disclosure risk assessment, *a posteriori* risk assessment has been proposed as a complement to *a priori* measures. That is, we propose to measure disclosure risk not only before the application of protection methods, but also after that. We have shown that the reciprocal of Shannon's conditional entropy (conditioned to the knowledge of the intruder) may be used as such a measure. While Shannon's entropy may not be suitable to evaluate the impact of disclosure control on table utility, it turns out to be extremely useful to quantify disclosure risk. As shown in Section 4, computing disclosure risk in this way can easily be done for different disclosure control methods, both with simple tables and linked tables. In the case of cell suppression methods applied to linked table protection, the proposed measure allows detection of secondary suppression patterns which offer more protection than others do. The strategy for choosing the best candidates for secondary suppressions has been outlined in the paper.

6 Acknowledgments

Thanks go to Sarah Giessing for useful comments on earlier versions of this paper.

7 References

Cox, L. H. (2001). Disclosure risk for tabular economic data, in *Confidentiality, Disclosure and Data Access*, eds. P. Doyle, J. Lane, J. Theeuwes and L. Zayatz. Amsterdam: North-Holland, 167-183.

Denning, D. E. (1982). *Cryptography and Data Security*. Reading, MA: Addison- Wesley.

Domingo-Ferrer, J. and Torra, V. (2002). A critique of the sensitivity rules usually employed for statistical table protection, in *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10, 5, 545-556.

Duncan, G. T., Fienberg, S. E., Krishnan, R., Padman, R. and Roehrig, S. F. (2001). Disclosure limitation methods and information loss for tabular data, in *Confidentiality, Disclosure and Data Access*, eds. P. Doyle, J. Lane, J. Theeuwes and L. Zayatz. Amsterdam: North-Holland, 135-166.

Felsö, F., Theeuwes, J. and Wagner, G. G. (2001). Disclosure limitation methods in use: Results of a survey, in *Confidentiality, Disclosure and Data Access*, eds. P. Doyle, J. Lane, J. Theeuwes and L. Zayatz. Amsterdam: North- Holland, 17-42.

Fienberg, S. E., Makov, U. E. and Steele, R. J. (1998). Disclosure limitation using perturbation and related methods for categorical data, *Journal of Official Statistics*, 14, 485-512.

Giessing, S. (2001). Nonperturbative disclosure control methods for tabular data, in *Confidentiality, Disclosure and Data Access*, eds. P. Doyle, J. Lane, J. Theeuwes and L. Zayatz. Amsterdam: North-Holland, 185-213.

Holvast, J. (1999). Statistical dissemination, confidentiality and disclosure, in *Proceedings of the Joint Eurostat/UNECE Work Session on Statistical Data Confidentiality*. Luxembourg: Eurostat, 191-207.

Luige, T. and Meliskova, J. (1999). Confidentiality practices in the transition countries, in *Proceedings of the Joint Eurostat/UNECE Work Session on Statistical Data Confidentiality*. Luxembourg: Eurostat, 287-319.

Robertson, D. and Ethier, R. (2002). Cell suppression: Theory and experience, in *Inference Control in Statistical Databases*, LNCS 2316, ed. J. Domingo-Ferrer. Berlin: Springer-Verlag, 9-21.

Shannon, C. E. (1948). A mathematical theory of communication, *Bell System Technical Journal*, 27, 379-423, 623-656, July and Oct. 1948.

Willenborg, L. and de Waal, T. (2001). *Statistical Disclosure Control in Practice*. New York: Springer-Verlag.

Resum

La protecció de dades estadístiques, també coneguda com a Control de Revelació Estadística (SDC), és una part de la seguretat de la informació que intenta evitar la publicació d'informació estadística (taules, registres individuals) que reveli la contribució de responents específics. Aquest article tracta de la valoració del risc de revelació associat a la difusió de dades tabulades. Les anomenades regles de sensibilitat estan sent utilitzades actualment per tal de mesurar el risc de revelació en taules. Aquestes regles operen sobre una base *a priori*: les dades són examinades i les regles s'utilitzen per decidir si les dades poden ser difoses tal com s'han elaborat o bé han de ser protegides. En aquest article, proposem complementar la mesura de *risc a priori* amb una mesura de risc *a posteriori* per tal d'aconseguir un nivell de seguretat més alt, és a dir, proposem tenir en compte la informació protegida quan es mesura el risc de revelació.

La mesura del risc de revelació *a posteriori* proposada és compatible amb una àmplia classe de mètodes de protecció de revelació i pot ser aplicada al càlcul del risc de revelació d'un grup de taules vinculades. En el cas de protecció de taules vinculades a través de la supressió de cel·les, la mesura proposada permet la detecció de patrons de supressió secundària els quals ofereixen més protecció que d'altres.

MSC: 62P99

Paraules clau: control de revelació estadística, bases de dades estadístiques, dades tabulades, seguretat