Information-Theoretic Disclosure Risk Measures in Statistical Disclosure Control of Tabular Data*

Josep Domingo-Ferrer
Universitat Rovira i Virgili
Dept. of Comp. Eng. and Maths
Av. Països Catalans 26
E-43007 Tarragona, Spain
e-mail jdomingo@etse.urv.es

Anna Oganian
Universitat Rovira i Virgili
Dept. of Comp. Eng. and Maths
Av. Països Catalans 26
E-43007 Tarragona, Spain
e-mail aoganian@etse.urv.es

Vicenç Torra
IIIA - CSIC
Campus UAB
E-08193 Bellaterra, Spain
e-mail vtorra@iiia.csic.es

Abstract

Statistical database protection 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 shows how to use information-theoretic concepts to measure disclosure risk for tabular data. The proposed 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.

Keywords: *Statistical database protection, Tabular data protection, Information theory, Disclosure risk.*

1 Introduction

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 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 [6], 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 [10, 3] for a survey). We next 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 [5], 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.

*Work partly funded by the European Union under project ”CASC” IST-2000-25069

As noted by [3], 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.

A first approach to measuring data utility is to take generic measures such as the reciprocal of the mean squared error between the original and the released tables [3]. While this may be useful as a crude approach, a more accurate utility assessment must necessarily take into account the specific data uses the user is interested in. Thus, strictly speaking, there is no universal data utility measure.

The situation for disclosure risk is quite different. There is a number of sensitivity rules which are used to decide whether a particular table cell can be safely released. However, these rules operate on an *a priori* basis: the original data are examined *before* they are protected and the rules are used to determine whether the data can be released as they stand or should rather be protected. Note that the disclosure risk incurred if a particular protected table is released is not actually measured by sensitivity rules.

1.1 Our contribution

We will show in this paper that it is possible to use information-theoretic concepts to define a general disclosure risk measure which takes protected information into account. This measure can be termed *a posteriori*, as it measures disclosure risk *after* table protection has been used: the protected table is taken as input to compute disclosure risk and will only be released if disclosure risk is deemed low enough by the data protector. The proposed measure applies to a broad class of disclosure protection methods and is computable in practice.

Section 2 describes some sensitivity rules currently used. In Section 3, a new measure of disclosure risk based on the reciprocal of conditional entropy is proposed as an *a posteriori* alternative to sensitivity rules. Section 4 describes an application of the proposed disclosure risk measure to different table protection methods, both for simple tables and for linked tables. Section 5 is a conclusion.

2 Background on sensitivity rules for tables

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.

According to [7, 8, 4], the $n-k$ dominance rule is the most popular one for magnitude tables, followed by the $p\%$-rule and the pq-rule. Yet, it is significant to note that the U.S. Census Bureau switched in 1992 from the $n-k$ rule to the $p\%$-rule, and the German Statistisches Bundesamt did the same in 2001.

Due to their nature, the above-mentioned rules are strongly oriented toward certain methods, in particular cell suppression (a sensitivity rule is used to decide which cells should be primarily suppressed) or table redesign. Their usefulness for other metods like rounding is less clear. This is confirmed by the survey [4]: sensitivity rules are nearly always used in conjunction with cell suppression. Furthermore, being *a priori*, sensivity rules do not always correctly reflect the disclosure risk caused by the release of a particular table. The following example reported in [9] is an illustration.

Example 1 (Robertson and Ethier, 2002) *In the dominance rule, let $n=1$ and $k=60\%$. Then a cell with value 100 and contributions 59, 40, 1 is declared* not *sensitive, while a cell with value 100 and contributions 61, 20, 19 would be declared sensitive. Assume now that the second largest respondent of both cells knows the total 100 and is interested in estimating the contribution of the largest respondent. Then, for the $(59,40,1)$ cell, she removes her contribution and gets an upper bound $100-40=60$ for the largest contribution. For $(61,20,19)$ the upper bound she gets is $100-20=80$, much farther from the real largest contribution. So the cell declared non-sensitive by the rule allows better inferences than the cell declared sensitive!*

3 Conditional entropy as a general measure of disclosure risk

The discussion in Section 2 points out that, as useful as sensitivity rules can be in combination with specific table protection methods, they may fail to capture the notion of disclosure risk in a correct way. Our proposal here is to use the reciprocal of Shannon's conditional entropy to express disclosure risk in a unified manner which takes protected data into account.

Entropy-based measures were already discussed in [10] for computing information loss at the table level, but not for computing disclosure risk. However, the authors of [10] do not believe entropy is a practical information loss measure. We support their opinion with the following example.

Example 2 *Assume we use rounding with integer base b to protect a table. The entropy-based information loss measure defined in [10] 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 [3], the reciprocal of Shannon's entropy (not conditional entropy) is suggested as measure of disclosure risk at the cell level. If p_ω is the probability (as seen by the intruder) that the value of a specific cell is ω, then the disclosure risk for that cell is measured as

$$DR(X) = 1/(-\sum(p_\omega \log_2 p_\omega)) \tag{1}$$

The summation in Expression (1) extends over all possible values of cell X. What is not clear here is how to compute p_ω, that is, what distribution should be chosen. In fact, the particular distribution for an intruder depends on the knowledge held by that intruder: if the intruder is an outsider, then the only information she has is the released table; if the intruder is an insider (one of the respondents who contribute to the particular cell), then she knows her own contribution and it may be easier for her to estimate the contribution of other respondents.

The information held by an intruder does not only depend on her being outsider or insider; it clearly depends also on what information has previously been published and on how that information has been protected. The following example illustrates this.

Example 3 *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.*

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

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

where X is a variable representing an original cell and Y is a variable representing the intruder's knowledge (which is suposed 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 to 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 [10] 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

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

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

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 $m(S_y(X))$ 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.

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

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

4.1 Cell suppression

The disclosure risk computation for cell suppression is illustrated by extending an example provided in [10]. 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:

$$\begin{array}{c} H(x_1|x_1+x_2=1537, x_1+x_3=406, x_i \geq 0) \\ H(x_2|x_1+x_2=1537, x_2+x_4=2382, x_i \geq 0) \\ H(x_3|x_1+x_3=406, x_3+x_4=1251, x_i \geq 0) \\ H(x_4|x_2+x_4=2382, x_3+x_4=1251, x_i \geq 0) \end{array}$$

Expressions (4) contain constraints y_i for each suppressed cell x_i which allow $m(S_{y_i}(x_i))$ 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: $0 \leq x_1 \leq 406$, $1131 \leq x_2 \leq 1537$, $0 \leq x_3 \leq 406$ and $845 \leq x_4 \leq 1251$. By substracting the bounds we obtain $m(S_{y_i}(x_i)) = 407$ 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 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:

$$\begin{array}{rcl} x_1+x_2+x_3+x_4+x_5-x_6 &=& 0 \\ x_7+x_8+x_9+x_{10}+x_{11}-x_{12} &=& 0 \\ x_1+x_7 &=& 649 \\ x_2+x_8 &=& 406 \\ x_3+x_9 &=& 1275 \\ x_4+x_{10} &=& 2382 \\ x_5+x_{11} &=& 3467 \\ x_6+x_{12} &=& 8179 \\ x_i &\geq& 0 \text{ for } i=1,\cdots,12 \end{array} \quad (4)$$

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

4.4 Linked tables

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:

$$DR_{unif}(z_{a_i s_j r_k}|AS) = 1/H(z_{a_i s_j r_k}|z_{a_i s_j} = \sum_k z_{a_i s_j r_k}) \tag{5}$$

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

$$DR_{unif}(z_{a_i s_j r_k}|AS, AR, SR) = 1/H(z_{a_i s_j r_k}|$$
$$z_{a_i s_j} = \sum_k z_{a_i s_j r_k}, z_{a_i r_k} = \sum_j z_{a_i s_j r_k}, z_{s_j r_k} = \sum_i z_{a_i s_j r_k}) \tag{7}$$

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 take into account the number of independent constraints imposed by the released linked tables.

Another point we have to take into account 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 $995 \leq z_{a_j s_j r_k} \leq 1004$. 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 595 and 604. Thus, company A is able to estimate the turnover of company B within 1% of its value. However, the uncertainty of an outsider about the turnover of company B is almost 170% of its value: the outsider only knows that the turnover of the largest contributor is between $\epsilon > 0$ and 1004. Therefore, for an insider (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 $1/H(z_{a_i s_j r_k}|\text{released tables, insider's contribution})$.

5 Conclusion

We have given arguments in favor of using the reciprocal of Shannon's conditional entropy as an *a posteriori* disclosure risk measure for tabular data. 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. *A priori* risk assessment through sensitivy rules is indeed useful to locate table cells to be protected, but we claim that the final decision to release a table should be based on an *a posteriori* measure of disclosure risk like the one discussed in this paper, which takes the actual protected table into account.

Acknowledgments

Thanks go to Sarah Gießing for useful comments on earlier versions of this paper. Comments by three anonymous referees are also gratefully acknowledged.

References

[1] L. H. Cox, "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, pp. 167-183, 2001.

[2] D. E. Denning, *Cryptography and Data Security*. Reading, MA: Addison-Wesley, 1982.

[3] G. T. Duncan, S. E. Fienberg, R. Krishnan, R. Padman and S. F. Roehrig, "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, pp. 135-166, 2001.

[4] F. Felsö, J. Theeuwes and G. G. Wagner, "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, pp. 17-42, 2001.

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

[6] S. Gießing, "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, pp. 185-213, 2001.

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

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

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

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