Internal intrusion scenarios in inference control of tabular databases

Anna Oganian, Josep Domingo-Ferrer
Dept. of Computer Engineering and Maths
Universitat Rovira i Virgili
ETSE-Av. Països Catalans 26
E-43007 Tarragona, Catalonia
e-mail {aoganian,jdomingo}@etse.urv.es

Vicenç Torra
Institut d'Investigació en IA
(IIIA-CSIC)
Campus UAB s/n
E-08193 Bellaterra, Catalonia
e-mail vtorra@iiia.csic.es

Abstract

The main goal of inference control in statistical data bases is to prevent statistical data from disclosing confidential information about specific respondents to third parties. Thus, one of the important aspects in the protection of the statistical data is the assessment of the disclosure risk. In this paper, we propose an entropy-like disclosure risk measure general enough so as to encompass internal intruders.

Keywords: Statistical databases, Inference control, Information retrieval systems.

1 Introduction

The most typical output offered by national statistical agencies is tabular data. Tabular data are obtained from microdata files containing information on individual units (persons, enterprises or institutions) by a process called aggregation. So, being aggregated data, one is tempted to think that tables are not supposed to contain information that can reveal the contribution of particular respondents. However, as noted in [3], in many cases table cells do contain information on a single or very few respondents, which implies some disclosure risk for the data of those respondents. In these cases, statistical disclosure control (SDC) methods must be applied to the tables prior to their release. The performance of SDC methods is measured in terms of information loss, that is, loss of utility of the released data as consequence of applying the protection method, and disclosure risk, that is, risk to respondent confidentiality if the data concerning this respondent are released. In this paper, we concentrate on disclosure risk measures for tabular data protection.

Two types of disclosure risk measures are proposed in the literature: *a priori* measures, also called sensitivity rules, and *a posteriori* measures. *A priori* measures are used prior the publication of the table. Table cells are examined and these rules are used to decide whether each of the cells can safely be released as it stands or should rather be protected. An example of *a priori* measure is a (n,k)-dominance rule. 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 a fraction k or more of the total cell value. However, it was shown recently in [2] that (n,k)-dominance rule and other widely used sensitivity rules, such as p%-rule and pq-rule may not adequately reflect disclosure risk.

Another class of disclosure risk measures are *a posteriori* measures. These measures should be used after applying SDC methods. This kind of measures take the protected information into account. As an *a posteriori* measure, the reciprocal of the conditional entropy $DR(X)$ was proposed in [1]

$$\frac{1}{H(X|Y=y)} = \frac{1}{(-\sum_x p(x|y) log_2 p(x|y))} \quad (1)$$

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).

In [1], it was shown that uniform distribution over the set of possible values of X given the constraints y may be used for several SDC methods.

If the uniform probability function is used, Expression (1) can be simplified to

$$DR_{unif}(X) = 1/\log_2 m(S_y(X)) \qquad (2)$$

where $m(S_y(X))$ is the number of possible values of the cell (Note that table cells take values in a discrete domain: either integer values or real values with a fixed number of decimal positions. Thus the set of possible values is enumerable).

Now let us consider an intruder. There may be two intruder types: external intruders whose only information about the contributions comes from the released table and internal intruders having some additional information (at least their own contributions). This internal intruder may be one of the contributors, who is related to the area and is thus in a better position than someone external. So, the disclosure risk for this type of intruder is higher.

To our best knowledge, disclosure risk measures for the internal intruder scenario have not been treated in the literature. So, let us consider which of the proposed measures for disclosure risk in tables may be used to assess disclosure risk vs internal intruders. *A priori* measures by their nature do not take the side intruder's information into account. They basically measure the concentration of the contributions to the cell. On the contrary, *a posteriori measures* based on conditional Shannon entropy can be adapted to assess the risk related to internal intruders. Showing how this can be done is precisely the main point of this paper.

1.1 Contribution and plan of this paper

Section 2 presents our entropy-like disclosure risk measure that takes the intruder's side knowledge into account. In Section 3, this measure is applied to a scenario where the intruder can rank contributions. In Section 4, a scenario where the intruder knows that a contribution does *not* have a certain rank; it is also illustrated in this section that disclosure risk measures based on conditional entropy do not work well. In Section 5, a scenario is considered where the intruder can tell whether a contribution is among the m largest or not. Section 6 deals with the case where the intruder has an approximate knowledge of the groups of contributors. Section 7 deals with the case where the intruder only knows the published cell value; this case is used as a benchmark and allows a sensitivity rule to be derived from our proposed disclosure risk measure. Section 8 comments on intruder collusions. Finally, Section 9 is a conclusion.

2 Internal intruder disclosure risk measures

Assume that, after examination of the original table, some cells where considered sensitive by a sensitivity rule and they were protected, while others were labeled as nonsensitive, so they were left as they stood. After that, *a posteriori* risk assessment is applied, which takes the released information into account.

As an *a posteriori* measure, the reciprocal of Shannon's conditional entropy may be used. We want to find out how it can be applied to a scenario with internal intruders.

To that end, we need to identify the essential property which must be satisfied by a disclosure risk measure. The natural property one expects from the disclosure risk is the following:

Property 1. *Disclosure risk increases as the intruder gains more information about respondents.*

Note that, in a scenario with internal intruders, assumptions about the intruders' knowl-

edge have to stay pretty general, because the data protector who uses these assumptions to measure the risk of releasing the table does not know exactly the information held by the intruder. This is why our study will be based on various general assumptions which can be plausible in certain scenarios.

Among the assumptions that characterize a scenario is the distribution of the random variables representing the intruder's knowledge. In most cases, we will use a uniform distribution as a starting point, because it models the maximal uncertainty of the data protector about the possible values of the intruder's knowledge. Of course, if the data protector knows the specific (non-uniform) distribution of variables representing the intruder's knowledge, then this distribution should be used.

Let us first define the common hypotheses we make for all scenarios about the internal intruder's knowledge

Hypothesis 1. *The intruder knows:*

1. *The cell value x (from the released table)*
2. *Her own contribution to the cell*
3. *The identities of the remaining N contributors to the cell other than her*
4. *Some additional information about contributors to the cell depending on the intruder scenario, which will be specified later.*

Assume now that the intruder wants to estimate the contribution X_v of the specific contributor v. Let us consider the following scenarios:

1. The intruder can rank contributions.
2. The intruder knows the identity of the k^{th} largest respondent.
3. The intruder knows the identities of the m largest respondents.
4. The intruder has a fuzzy knowledge of the ranking of contributions to a sensitive cell.
5. The intruder colludes with other contributors, who reveal their contributions to her.

When entropy-based measures are used to measure disclosure risk, it must be taken into account that entropy is a measure of *average* uncertainty about the result of an experiment, *i.e.* $H(X_v)|Y) = -\sum_y P(y)H(X_v)|Y = y)$, where X_v is a random variable representing the contribution of respondent v. In a disclosure scenario, where Y represents side information about contribution X_v, the reciprocal of this average uncertainty may contradict Property 1. This contradiction will be illustrated below based on the comparison of several disclosure scenarios assuming more or less additional knowledge by the intruder.

To measure the disclosure risk for the internal intruder, we propose an additive entropy-like disclosure risk measure equal to the reciprocal of the sum of the uncertainties the intruder has about ranking the contributions and about estimating the size of the contribution given that she has found the correct ranking:

$$DR = 1/(H(r_v) + H(X_v|r_v)) \qquad (3)$$

where $H(X_v|r_v)$ is the average uncertainty about contribution X_v when its rank r_v is known by the intruder and $H(r_v)$ is the average uncertainty of the intruder about the rank. Since $H(X_v|r_v)$ represents an average uncertainty, without loss of generality we can compute it using the uniform distribution for r_v, regardless of the actual distribution of r_v assumed and used to compute $H(r_v)$ in a particular scenario. The reason is that, when the rank is known, one would expect the same $H(X_v|r_v)$ regardless of the scenario[1].

The part of Expression (3) where the rank distribution is reflected is $H(r_v)$. That is, the less information the intruder has about the rank, the higher the uncertainty. For N contributions other than the intruder's contribution, the maximum uncertainty is $\log_2 N$ when there is no knowledge at all about the possible rank value; the minimum uncertainty is $H(r_v) = 0$ when the rank is known.

[1] Strictly speaking, $H(X_v|r_v)$ is the average of $H(X_v|r_v = i)$ for all possible i. However, in order to have a single measure for every scenario rather than a set of measures for every value of the variable r_v representing side information, we compute $H(X_v|r_v)$ as an average of $H(X_v|r_v = i)$ over all i.

Since the actual distribution of r_v is not considered in $H(X_v|r_v)$, Expression (3) does not exactly correspond to the reciprocal of the joint entropy of X_v and r_v. The advantage of moving away from the reciprocal of the joint entropy is that a distribution-independent $H(X_v|r_v)$ allows Property (1) to be satisfied by our measure. This will be illustrated in the scenarios considered in the rest of this paper.

Now let us consider in turn each of the above-mentioned scenarios and compare Expressions 1 and 3 to measure disclosure risk in each scenario.

3 The intruder can rank contributions

Assume the intruder is interested in a contribution X_v with rank j where $j \in \{1 \cdots N\}$. Denote the contribution with rank i by $X_{(i)}$, that is $X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(N)}$. So, the intruder wants to estimate contribution $X_v = X_{(j)}$.

Let $x' = x - x_{intr}$, where x_{intr} is the intruder's contribution and x is the cell total.

Ranking imposes upper and lower bounds for $X_{(j)}$. In order to ensure $X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(N)}$, the ranked contribution $X_{(j)}$ should satisfy $0 < X_{(j)} < x'/N - j + 1$ for $1 \leq j \leq N-1$ and $x'/N < X_{(N)} < x'$. Denote by Num_j the cardinality of the interval in which $X_{(j)}$ lies. It is easy to see that $Num_j < Num_{j+1}$. Note that these intervals are too broad and overlap with each other, so they provide relatively little information about the contribution size.

Note also that, if the intruder can rank the contributions, she probably has some idea about the bounds for every contribution. That is, the whole interval $(0, x']$ is divided into disjoint intervals In_j, where $1 \leq j \leq N$. If the number of possible values in In_j is denoted by A_j, we have $A_j < Num_j$. Note that these bounds may be quite imprecise, with the only constraint that they should lie within the above discussed intervals imposed by ranking. We also assume a natural feature they inherited from the intervals imposed by ranking, namely that $A_j < A_{j+1}$.

General guidance on computing A_j may hardly be provided without taking into account the characteristics of the intruder. The definition of intervals A_j by the data protector is determined by the specific situation. When the rank $r_v = j$ of contribution $X_v = X_{(j)}$ is known, we have $H(r_v) = 0$ and the uncertainty about contribution X_v is $H(X_v|r_v = j) = \log_2 A_j$. Thus, the disclosure risk for every ranked and bounded contribution can be computed using using Expression (3) as:

$$DR_{r_v=j} = \frac{1}{H(X_v|r_v = j)} = \frac{1}{\log_2 A_j} \qquad (4)$$

As the data protector does not exactly know what contribution the intruder is interested in, it is sensible to model the random variable r_v indicating the contribution's rank as being uniformly distributed. So Expression (2) can be used to compute:

$$DR_{rank} = 1/H(X_v|r_v) =$$

$$= 1/(\frac{1}{N}\sum_{j=1}^{N} H(X_v|r_v = j)) =$$

$$= 1/(\sum_{j=1}^{N} \frac{1}{N} \log_2 A_j) = N/(\sum_{j=1}^{N} \log_2 A_j) \qquad (5)$$

4 The intruder knows that a contribution does *not* have a certain rank

Assume the intruder knows that the contribution X_v she is interested in has rank different from k (for a contribution with rank k, use results discussed in Section 3).

There are $N-1$ possible values for the rank r_v of X_v, all of them equally likely. Thus $H(r_v) = \log_2(N-1)$. On the other hand, $H(X_v|r_v)$ is as defined in Section 3. Thus, using Expression 3, we can estimate the disclosure risk $DR_{k-contr}$ for this scenario as:

$$DR_{k-contr} = 1/(\log_2(N-1) + \frac{\sum_i \log_2 A_i}{N}) \quad (6)$$

It is obvious that this disclosure risk is smaller than in the scenario where the ranking of the contributions. This is coherent with Property 1, because here the intruder's knowledge is more limited.

4.1 What would happen if conditional entropy was used in this scenario

As announced at the beginning of Section 2, we will next illustrate that, if disclosure risk was measured as the reciprocal of conditional entropy, Property 1 would not be satisfied.

The entropy of X_v conditional to the knowledge of the intruder is:

$$H(X_v|r_v \neq k) = \sum_{j:j\neq k} P(r_v = j) *$$

$$* \sum_{X_v:r_v=j} P(X_v|r_v = j) \log_2 \frac{1}{P(X_v|r_v = j)}$$

$$= \frac{1}{N-1} \sum_{j:j\neq k} \log_2 A_j \qquad (7)$$

We have used above the uniform distribution over possible ranks for X_v, because the intruder does not have any knowledge about the rank of $c(r_v)$ except that it is not equal to k (due to the assumption in this scenario).

In the case where the ranking of contributions is known by the intruder (Section 3), the conditional entropy of the contribution given the knowledge of the rank j is $\log_2 A_j$, where $A_j < A_{j+1}$.

In this scenario, where the intruder only knows the identity of k-th largest contributor, the conditional entropy of the contribution given that knowledge is the average of the conditional entropies for all possible $N-1$ rankings of the contribution of interest, *i.e.*

$$H(X_v|r_v \neq k) = \frac{1}{N-1} \sum_{j:j\neq k} \log_2 A_j$$

Assume now that the intruder wants to estimate the largest contribution. Since $A_1 < A_2 < \cdots A_N$,

$$H(X_v|r_v = N) = \log_2 A_N > H(X_v|r_v \neq k)$$

Thus, we reach the paradox that the entropy of the contribution X_v is larger when its rank is known than in the case when its rank is not known and the only available information is that the rank is not equal to k! The above would imply that the disclosure risk (using Expression (4)) in the scenario where the ranking is known is higher than when it is not known and the only known thing is that some contribution has rank k. This clearly contradicts Property 1.

Using in this scenario Expression (5) derived for the scenario with known ranking is no better. Indeed, for example for $k = N$ we have

$$\frac{\sum_{j=1}^{N} H(X_v|r_v = j)}{N} =$$

$$\frac{\sum_{j=1}^{N} \log_2 A_j}{N} > \frac{\sum_{j=1}^{N-1} \log_2 A_j}{N-1} = H(X_v|r_v \neq k)$$

So clearly, the above inequality does not satisfy Property 1 (there is more information on ranks on the left-hand side of the inequality than on the right-hand side).

Finally, if we generalize and consider K (the rank which the contribution cannot have) as a uniformly distributed variable, we get:

$$H(X_v|r_v \neq K) = \frac{1}{N}(\sum_k H(X_v|r_v \neq k)$$

where

$$H(X_v|r_v \neq k = \frac{1}{N-1} \sum_{i:i\neq k} \log_2 A_i$$

So, the disclosure risk $DR_{k-contr}$ computed as a conditional entropy for the second scenario is:

$$1/(\frac{1}{N}\frac{1}{N-1}(\sum_{i:i\neq 1} \log_2 A_i + \sum_{i:i\neq 2} \log_2 A_i + \cdots$$

$$+ \sum_{i:i\neq N} \log_2 A_i)) = 1/(\frac{1}{N(N-1)} \sum_{i}^{N} \log_2 A_i * (N-1)) =$$

$$= N/\sum_{i=1}^{N} \log_2 A_i = DR_{rank} \qquad (8)$$

We obtained $DR_{rank} = DR_{k-contr}$. Again, this contradicts Property 1 because the intruder has less information in the second scenario than in the first scenario.

This suite of contradictory results occur because entropy is the measure of *average* uncertainty about the random variable. Thus,

whether the intruder's knowledge is modeled as a random variable or as a set of constants, the disclosure risk computed as a reciprocal conditional entropy is not a satisfactory measure.

5 The intruder can tell whether a contribution is among the m largest or not

Assume the intruder knows the identities of the m largest contributors and also knows whether the contribution X_v she is interested in belongs to the group of the m largest or to the group of the $N-m$ smallest.

As in the previous scenario, it is easy to show that the entropy of the contribution conditional to the knowledge of group membership (a Boolean, since there are only two groups) may be less than the entropy of the contribution conditional to the known rank of this contribution (an integer).

Fortunately, if we use Expression (3) we do not have this contradiction. To use that expression, we first compute

$$H(r_v) = \frac{1}{2(N-1)} \sum_{m=1}^{N-1} (\log_2 m + \log_2 (N-m))$$

$$= \frac{\sum_{m=1}^{N-1} \log_2 m}{N-1}$$

Note that $H(r_v)$ is conditioned to the variable *group* (which may be first or second) and to the variable m which represents the cardinality of the second group. We assume that both variables are uniformly distributed.

So now, the disclosure risk in this scenario is:

$$DR_{m-group} = 1/(H(r_v) + H(X_v|r_v))$$

$$\frac{\sum_{i=1}^{N-1} \log_2 i}{N-1} + H(X_v|r_v) \qquad (9)$$

where $H(X_v|r_v)$ is defined as the reciprocal of Expression (5). It can be seen that this risk measure does not contradict Property 1.

6 The intruder has approximate knowledge of the groups of respondents

Assume the intruder can approximately split the set of contributors into groups G_i where $1 \leq i \leq q$ depending on the the size of their contributions. That is, the intruder approximately knows whose contributions are small, whose contributions are medium and so on, but she does not know the values of these contributions.

We assume that the intruder has got no knowledge about the relative position of the contributions within the groups.

We denote the contributions from G_i by X_{G_i}. If we want to distinguish among them, we will use the notation $X_{G_i,t}$ to denote the contribution of a contributor t from group G_i. The cardinality of G_i will be expressed as $|G_i|$. Approximate knowledge means that the intruder is not 100% sure that the contribution of a certain respondent belongs to G_i. She only assumes that it is the most likely case, but she does not totally rule out the possibility that this contribution may fall in the nearest group to G_i, that is G_{i-1} or G_{i+1}. In particular, it may fall in one of the nearest positions to the group G_i, not far from the frontier of G_i (not far, in the sense that only few positions in comparison with the number of positions in G_i may be possible candidates for it). We assume that only neighboring groups may overlap with each other.

We will call the positions of some group where the elements of other group may fall the *overlap zones* and the positions of the group where only elements of this group may fall the *proper zone*. So, every group G_i (except G_1 and G_q) has two overlap zones:

- the zone where elements from the left neighboring group G_{i-1} may fall, denoted: $(i, i-1)$
- the zone where elements from the right neighboring group G_{i+1} may fall, denoted: $(i, i+1)$

The overlap zone of the first group is $(1, 2)$ and the overlap zone of the last group is $(q, q-1)$. The proper zone of group G_i we denote by (i, i).

Here we assume that the cardinality of overlap zones is the same for every group and denote it by n (a natural assumption when all groups are equally fuzzy to the intruder). We will denote the positions in each zone by j_{ml}, where ml is the zone. If some contribution from G_i is not in the group G_i, the possibility that it lands in a position in a neighboring group $G_{i\pm1}$ decreases the farther is the position from the frontier of the group G_i and becomes zero after a certain position in $G_{i\pm1}$. A natural assumption we make here is that the possibility that a position j in an overlap zone is occupied by an element from another group, denoted by p_j^{ov}, is the same for all groups and only depends on its distance j from the frontier: the farther the contribution falls from the frontier of its group, the less the possibility. In terms of probability theory, for every group G_i and G_t, p_j^{ov} is:

$$P(r_{G_i,1} = j_{ml} \vee r_{G_i,2} = j_{ml} \vee \cdots$$

$$\vee r_{G_i,|G_i|} = j_{ml}) = P(r_{G_t,1} = j_{m'l'} \vee$$

$$r_{G_t,2} = j_{m'l'} \vee \cdots r_{G_t,|G_t|} = j_{m'l'}) = p_j^{ov} \quad (10)$$

where $(ml) \in \{(i+1,i),(i-1,i)\}$, $(m'l') \in \{(t+1,t),(t-1,t)\}$. The disclosure risk is

$$DR_{approx} = 1/(H(r_v) + H(X_v|r_v)$$

where $H(X_v|r_v)$ is calculated like in the previous scenarios and

$$H(r_v) = \frac{\sum_{i=1}^{q} H(r_v|group_v = i)}{q}$$

and

$$H(rank|group = i) =$$

$$= - \sum_{j_{ml}:(ml)\in\{(i,i),(i,i\pm1),(i\pm1,i)\}} p(j_{ml}) \log_2 p(j_{ml}) \quad (11)$$

where j_{ml} is the position such that (ml) is one of the zones where the contribution may fall and $\{P(j_{ml})\}$ is the distribution over possible positions j_{ml} which satisfy the conditions of this scenario described above.

Considering the distribution model as a variable will complicate the case too much, but some guidance on how to choose an acceptable model may be given. A possible strategy is as follows.

Let us consider some group G_i. There are three types of zone where elements from G_i may fall: a proper zone (i,i), an overlap zone $(i,i\pm1)$ in G_i where elements from the neighboring group may fall and an overlap zone $(i\pm1,i)$ in the neighboring group where elements from G_i may fall.

First let us consider (i,i). Its positions may be occupied only by the elements of G_i, so:

$$P(r_{G_i,1} = j_{ii} \vee r_{G_i,2} = j_{ii} \vee \cdots r_{G_i,|G_i|} = j_{ii}) = 1$$

As the events in the above probability are mutually exclusive and the intruder has no knowledge about the rank of elements within the group, we have:

$$P(r_{G_i,1} = j_{ii}) = P(r_{G_i,2} = j_{ii}) = \cdots$$

$$\cdots P(r_{G_i,|G_i|} = j_{ii}) = \frac{1}{|G_i|} \quad (12)$$

Regarding the zone $(i,i\pm1)$, the situation is the following:

$$P(r_{G_i,1} = j_{i,i\pm1} \vee r_{G_i,2} = j_{i,i\pm1} \vee \cdots$$

$$r_{G_i,|G_i|} = j_{i,i\pm1} \vee r_{G_{i\pm1},1} = j_{i,i\pm1} \vee$$

$$\cdots r_{G_{i\pm1},|G_{i+1}|} = j_{i,i\pm1}) = 1$$

this yields

$$|G_i|P(r_{G_i} = j_{i,i\pm1}) = 1 - |G_{i\pm1}|P(r_{G\pm1} = j_{i,i\pm1})$$

As $|G_{i\pm1}|P(r_{G_{i\pm1}} = j_{i,i\pm1}) = p_j^{ov}$, we have:

$$P(r_{G_i} = j_{i,i\pm1}) = \frac{1 - p_j^{ov}}{|G_i|} \quad (13)$$

For the zone $i\pm1,i$ using Expression (10) we obtain:

$$P(c(r_{G_i}) \to j_{i\pm1,i}) = \frac{p_j^{ov}}{|G_i|} \quad (14)$$

The only thing that should be specified are the probabilities p_j^{ov}. Whether these probabilities should be taken large or small depends on how fuzzy the frontier of the groups may be seen by the intruder.

7 The intruder knows only the published cell value

We consider here the scenario which corresponds to the maximal uncertainty about the contribution X_v. Note that we still assume that the intruder knows the contributors to the cell. This scenario is a benchmark to compare different scenarios and devise some

threshold algorithm for our disclosure risk measure vs internal intruders. So, according to Expression (3), the disclosure risk in this scenario is

$$DR_{min} = 1/(\log_2 N + \frac{\sum_{i=1}^{N} \log_2 A_i}{N}) \qquad (15)$$

Note that this disclosure risk is minimal in comparison with other scenarios where some additional intruder's knowledge is assumed. Just compare with the expressions determined above for DR_{rank}, $DR_{k-contr}$, $DR_{m-group}$ and DR_{approx}.

Thus, the sensitivity rule for the scenarios considered in this paper based on the proposed measure could look like:

Algorithm 1 (t-sensitivity rule for the internal intruder).

1. *Let parameter $t \in [0,1]$ be a sensitivity threshold.*
2. *Given a cell X, compute $DR_{scenario}$ using the expression corresponding to the particular scenario and DR_{min} using Expression (15).*
3. *If $DR_{min}/DR_{scenario} < t$, declare X as sensitive; otherwise, declare X as non-sensitive.*

8 Collusion

If n contributors collude with the intruder, then we can adapt the measures described above for the different scenarios by just taking into account that:

- Instead of $N-1$, the number of cell contributors other than the colluders will be $N-n$.
- The sum of the contributions to the cell from non-colluding contributors is $x' - \sum_{colluder} x_{colluder}$.

With these two cautions, A_j can be redefined and the measures can be computed.

9 Conclusion

We have considered several scenarios for the internal intruder. Although the number of possible scenarios is virtually infinite, we think that similar considerations may be used for measuring the disclosure risk in a considerably larger number of scenarios. The proposed scenarios can be used as a base to derive measures for more complex situations in which more additional knowledge is available to the intruder. This knowledge will influence the entropy $H(r_v)$ of the position of the contribution and the related bounds, while the entropy $H(X_v|r_v)$ of the contribution given the rank will be the same.

Acknowledgements

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

References

[1] J. Domingo-Ferrer, A. Oganian and V.Torra, "Information-theoretic disclosure risk measures in statistical disclosure control of tabular data", in *Proceedings of the 14th Intl. Conf. on Scientific and Statistical Database Management (SSDBM'2002)*, (J. Kennedy, ed.), Los Alamitos CA: IEEE Computer Society, pp. 227-231, 2002.

[2] J. Domingo-Ferrer and V.Torra, "A critique of sensitivity rules usually employed for statistical table protection", in *Intl. J. of Unc., Fuzz. and Knowledge-Based Systems*, Vol. 10, No. 5, pp. 545-556, 2002.

[3] 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.

[4] L. Willenborg and T. de Waal, *Elements of Statistical Disclosure Control*. New York: Springer-Verlag, 2001.