International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems
Vol. 0, No. 0 (1993) 000—000


(Leave $1\frac{1}{2}$ inch blank space for Publisher)

# A CRITIQUE OF THE SENSITIVITY RULES USUALLY EMPLOYED FOR STATISTICAL TABLE PROTECTION*

JOSEP DOMINGO-FERRER[1], VICENÇ TORRA[2]
[1] *Universitat Rovira i Virgili, Dept. of Computer Eng. and Maths*
*Av. Països Catalans 26, E-43007 Tarragona, Catalonia*
*jdomingo@etse.urv.es*
[2] *Institut d'Investigació en Intel·ligència Artificial*
*Campus de Bellaterra, E-08193 Bellaterra, Catalonia*
*vtorra@iiia.csic.es*

Received (received date)
Revised (revised date)

In statistical disclosure control of tabular data, sensitivity rules are commonly used to decide whether a table cell is sensitive and should therefore not be published. The most popular sensitivity rules are the dominance rule, the $p$%-rule and the $pq$-rule. The dominance rule has received critiques based on specific numerical examples and is being gradually abandoned by leading statistical agencies. In this paper, we construct general counterexamples which show that *none* of the above rules does adequately reflect disclosure risk if cell contributors or coalitions of them behave as intruders: in that case, releasing a cell declared non-sensitive can imply higher disclosure risk than releasing a cell declared sensitive. As possible solutions, we propose an alternative sensitivity rule based on the concentration of relative contributions. More generally, we suggest to complement *a priori* risk assessment based on sensitivity rules with *a posteriori* risk assessment which takes into account tables after they have been protected.

*Keywords*: Disclosure risk, Statistical disclosure control, Statistical databases, Security.

## 1. Introduction

Tabular data constitute the main type of information released by statistical agencies. Being aggregate data, one might infer that tables cannot leak information about specific respondents. As argued in [4], it turns out that table cells often 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.

Disclosure risk assessment for tables is usually performed *a priori*, that is, before tables are protected. The standard approach is to use a *sensitivity rule* to decide whether a table cell is sensitive and should be protected or even suppressed.

For magnitude tables (normally related to economic data), there are three widely accepted rules to decide whether a cell is sensitive:

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

**$(n,k)$-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 a fraction $k$ or more of the total cell value. Usually, $n$ is taken to be 1 or 2, and $k$ is a fraction higher than 0.6.

**$pq$-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 $q$ 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 $p$ percent.

**$p$%-rule** This can be seen as a special case of the $pq$-rule. In this case, no knowledge prior to table publication is assumed, *i.e.* the $pq$-rule is used with $q = 100$.

As noted by [2], the disclosure risk of contributions increases as the percent within which they can be estimated by an intruder decreases. Therefore, when releasing a cell value, the greatest risk occurs for the largest contribution to the cell. That is the reason why the risk of disclosure of the largest contribution is a sufficient measure of disclosure risk.

### 1.1. *Our contribution*

The purpose of this paper is to prove through general counterexamples that, if one or more cell contributors collude and behave as intruders, releasing a cell declared non-sensitive by the main sensitivity rules in use ($(n,k)$ dominance, $p$%-rule and $pq$-rule) can actually imply higher disclosure risk than releasing a cell declared sensitive. We also propose an alternative sensitivity rule which measures the concentration of contributions and is free from the aforementioned contradictions.

Section 2 discusses flaws related to the dominance rule and proposes a new sensitivity rule as a replacement. Section 3 extends the critique to the $p$% and the $pq$ rules; the new sensitivity rule described in Section 2 can also be used as a replacement. Some conclusions and recommendations are given in Section 4.

## 2. Analysis of the dominance rule

According to [5,6,3], 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 dominance rule to the $p$%-rule, and the German Statistisches Bundesamt did the same in 2001.

The dominance rule has received critiques for failing to adequately reflect the risk of disclosure, but these have been limited to numerical counterexamples for particular choices of $n$ and $k$. The following is a counterexample from [7] for the particular case $n = 1$:

**Example 1 (Robertson and Ethier, 2002)** *In the dominance rule, let $n = 1$ and $k = 0.6$ (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!*

We generalize below the critique in the above example for any values of $n$ and $k$. Assume a cell $X$ in a table takes a value $x$ which is formed by $N$ respondent contributions $x_1, \cdots, x_N$. Equivalently,

$$x = x_1 + x_2 + \cdots + x_N$$

The dominance rule declares $X$ to be sensitive if a few contributions ($n$ or less) add up to a substantial fraction of $x$ ($k$ or more).

In order to construct a non-sensitive cell, we need the following result.

**Lemma 1** *For any integer* $n$ *and* $k \in (0, 1]$*, there exists an integer* $N$ *and* $r \in [0, 1)$ *such that*

$$f(r) = \frac{r^n - 1}{r^N - 1} = k \tag{1}$$

**Proof:** For fixed $n$ and $N$, with $n < N$, the function

$$f(r) = \frac{r^n - 1}{r^N - 1}$$

bijectively maps the interval $[0, 1)$ onto the interval $(n/N, 1]$. The lower bound of the image interval is determined as

$$\lim_{r \to 1} f(r) = \frac{n}{N}$$

Thus, the lemma holds if we take $N$ large enough so that $n/N < k$. QED

The non-sensitive cell is now constructed as follows.

**Construction 1 (Non-sensitive cell $X_{ns}$)**

1. *Take* $r \in [0, 1)$ *and* $N$ *as defined in Lemma 1. Then it holds that*

$$k = \frac{r^n - 1}{r^N - 1} = \sum_{i=1}^{n} \frac{r^i (r - 1)}{r^{N+1} - r} \tag{2}$$

2. *Let*

$$R_i := \frac{r^i (r - 1)}{r^{N+1} - r} \tag{3}$$

3. *Consider a cell $X_{ns}$ whose $N$ relative contributions are*

$$x_i/x = \begin{cases} R_i & \text{for } i = 1, \cdots, n-1 \text{ and } i = n+2, \cdots, N \\ R_n - \varepsilon & \text{for } i = n \\ R_{n+1} + \varepsilon & \text{for } i = n+1 \end{cases}$$

*where $\varepsilon := (R_n - R_{n+1})/3$. With this choice of $\varepsilon$, one still has $x_n/x > x_{n+1}/x$.*

4. *According to Expression (2), the $n$ largest relative contributions $x_1/x, \cdots, x_n/x$ add to $k - \varepsilon$. Therefore, there is no subset of $n$ contributions adding to $k$, so $X_{ns}$ is clearly* not *sensitive according to the dominance rule.*

The sensitive cell is constructed as follows.

**Construction 2 (Sensitive cell $X_s$)**

1. *Take the same values $N$, $n$ and $k$ used in Construction 1.*

2. *Consider a cell $X_s$ whose relative contributions are*

$$x_i/x = \begin{cases} R_i & \text{for } i = 1, \cdots, n \\ (1-k)/(N-n) & \text{for } i = n+1, \cdots, N \end{cases}$$

3. *By construction, the sum of the $n$ relative contributions $x_1/x, \cdots, x_n/x$ is $k$. Thus, $X_s$ is declared sensitive by the $(n, k)$-dominance rule.*

We next show that the cell declared non-sensitive by the dominance rule can yield a closer upper bound for the largest contribution than the cell declared sensitive.

**Theorem 1** *Let $n$ and $k$ be the parameters of the dominance rule. Assume a coalition of the $n$ second largest contributors want to upper-bound the largest contribution. For any $n$ and $k$, if $N$ is taken large enough, then the coalition gets a proportionally closer upper bound for the case of $X_{ns}$ than for the case of $X_s$.*

**Proof**. For both $X_s$ and $X_{ns}$, the $n$ second largest contributors know that the largest contribution is upper-bounded by the total $x$ minus their own contributions, that is

$$x_1 \leq x - (x_2 + x_3 + \cdots + x_{n+1}) \tag{4}$$

The distance between $x_1$ and the upper bound (4) is exactly the sum of the $N-n-1$ smallest contributions. We next show that, for large enough $N$, this sum is smaller for $X_{ns}$ than for $X_s$ (since $x_1$ is the same for both cells, this is equivalent to showing that the upper bound on the largest contribution is proportionally closer for $X_{ns}$). Now, both $X_{ns}$ and $X_s$ total to $x$, so we can use in what follows relative contributions rather than absolute contributions for both cells. For $X_{ns}$, the $N-n$ smallest relative contributions add to $1 - k + \varepsilon$ by construction; therefore, the $N-n-1$ smallest relative contributions add to $1-k+\varepsilon$ minus the $n+1$-th largest relative contribution

$$(1-k) + \varepsilon - (R_{n+1} + \varepsilon) = (1-k) - \frac{kr^n(r-1)}{r^n - 1} \tag{5}$$

To obtain the last term of Expression (5), we have used that $k = (r^n - 1)/(r^N - 1)$. On the other hand, for $X_s$, the sum of the $N - n - 1$ smallest relative contributions is

$$\frac{(N-n-1)(1-k)}{N-n} = (1-k) - \frac{1-k}{N-n} \tag{6}$$

If $N \to \infty$, Expression (6) approaches $1 - k$. On the other hand, if we let $N \to \infty$, then $r$ is such that $k = 1 - r^n$ (according to Lemma 1). In this case, Expression (5) becomes

$$(1-k) - \frac{k(1-k)(r-1)}{1-k-1} = r(1-k) < (1-k)$$

Thus, $N$ can be taken large enough so that Expression (6) is larger than Expression (5), which causes the theorem to hold. QED.

The following example illustrates that $N$ does not need to be very large for Theorem 1 to hold.

**Example 2** *For* $n = 1$, $k = 0.369$ *and* $N = 3$, *we have* $r = 0.9$. *The largest relative contribution is* $x_1/x = 0.369$ *for both* $X_s$ *and* $X_{ns}$. *The tail of the* $N - n - 1 = 1$ *smallest relative contribution is* 0.298893 *for* $X_{ns}$ *and* 0.315498 *for* $X_s$.

### 2.1. *Replacing the dominance rule*

Theorem 1 highlights a major flaw in the dominance rule. Note that $n$ can be as small as 1 and, in that case, *a single cell contributor* (the second largest) can, without any help, get more precise estimates on the largest contribution for a cell declared non-sensitive than for a cell declared sensitive.

The explanation of the flaw is that the dominance rule does not capture well the concentration of contributions to a table cell (even if it defines sensitivity as concentration in a rough way). That is the reason why a cell declared non-sensitive by the dominance rule can have a smaller tail (sum of small contributions) than a cell declared sensitive, which results in better bounds on the largest contribution.

As an alternative sensitivity rule, it could seem natural to directly check whether the sum of a number of smallest contributions to the cell is below some threshold. The problem is how many smallest contributions should be considered. Note that, for $X_{ns}$ and $X_s$, the (flawed) dominance rule actually amounts to checking whether the $N - n$ smallest contributions add to $1 - k$ or less. If we consider only the $N - n - 1$ smallest contributions, then $X'_{ns}, X'_s$ could be constructed such that a result analogous to Theorem 1 can be proven for a coalition of the $n + 1$ (rather than $n$) second largest contributors. The same argument applies to any number of smallest contributions that can be chosen.

A better sensitivity rule could be based on actually measuring the concentration of contributions. Evenness of the distribution of contributions can be measured by the entropy of the relative contributions as follows:

$$H(X) = -\sum_{i=1}^{N} (x_i/x) \log_2(x_i/x) \tag{7}$$

Expression (7) is maximal when all relative contributions are the same, *i.e.* when $x_i/x = 1/N$ for all $i$. Therefore, the higher $H(X)$, the less concentration in the contributions to cell $X$ and the less sensitive should the cell be. We next compute $H(X_{ns})$ and $H(X_s)$:

- The entropy of the relative contributions of $X_{ns}$ is

$$H(X_{ns}) = -\sum_{i=1}^{n-1} R_i \log_2(R_i) - (R_n - \varepsilon)\log_2(R_n - \varepsilon)$$

$$-(R_{n+1} + \varepsilon)\log_2(R_{n+1} + \varepsilon) - \sum_{i=n+2}^{N} R_i \log_2(R_i) \tag{8}$$

- The entropy of the relative contributions of $X_s$ is

$$H(X_s) = -\sum_{i=1}^{n} R_i \log_2(R_i) - (1-k)\log_2(\frac{1-k}{N-n}) \tag{9}$$

The theorem below shows that concentration can be higher for a cell declared non-sensitive by the dominance rule than for a cell declared sensitive, that is, $H(X_{ns}) < H(X_s)$. The proof can be found in the Appendix. Thus, if the dominance rule is replaced by entropy as a sensitivity measure, $X_{ns}$ is declared as more sensitive than $X_s$ for large enough $N$, which is consistent with Theorem 1.

**Theorem 2** *Let $n$ and $k$ be the parameters of the dominance rule. Let $N$ be the number of contributors to cells $X_s$ and $X_{ns}$. For any $n$ and $k$, if $N$ is taken large enough, then $H(X_{ns}) < H(X_s)$ holds.*

The following example illustrates Theorem 2 for the same parameters used in Example 2.

**Example 3** *For $n = 1$, $k = 0.369$ and $N = 3$, we have $r = 0.9$ and*

$$H(X_{ns}) = 1.58088 < H(X_s) = 1.5809$$

*Thus, entropy gives results consistent with Example 2.*

If entropy is to be used as a sensitivity rule, it is desirable to bound its range to a fixed interval, for example $[0, 1]$. A known property of the entropy of a variable taking one of $N$ values is that $0 \leq H(X) \leq \log_2 N$. Thus, an entropy-based sensitivity rule could look like:

**Algorithm 1 (Entropy-based sensitivity(t))**

1. *Let parameter $t \in [0, 1]$ be a sensitivity threshold.*
2. *Given a cell $X$ with $N$ contributors, compute $H(X)$ using Expression (7).*
3. *If $H(X)/\log_2 N < t$ declare $X$ as sensitive; otherwise, declare $X$ as non-sensitive.*

Note that parameter $t$ in Algorithm 1 should *not* be too close to 1. If $t$ is taken close to 1, only cells with very even relative contributions are declared non-sensitive; paradoxically, this knowledge could lead to disclosure of contributions in those cells.

## 3. Analysis of the $p\%$ and the $pq$-rules

For the sake of simplicity, we will focus the discussion on the $p\%$-rule. Analysis of the $pq$-rule is analogous. The formulation of the $p\%$-rule is in practice more specific than the definition given in Section 1. In its most commonly accepted form (see [4]), the $p\%$-rule declares a cell to be sensitive if the second largest contributor can estimate the largest contribution $x_1$ to within a proportion $p$ of $x_1$ (for convenience, we will take $p$ as a proportion rather than as a percent), that is, when

$$x - x_2 - x_1 \leq px_1 \tag{10}$$

where $x_2$ is the second largest contribution and $x$ is the cell total.

Let us now construct a cell $X_{ns}$ which is non-sensitive according to the $p\%$-rule.

**Construction 3 (Non-sensitive cell $X_{ns}$)**

1. *Take $r \in [0, 1)$ such that*

$$g(r) := \frac{r^N - r^2}{r - 1} = p \tag{11}$$

*For $N \geq 3$ and $r \in [0, 1)$, $g$ is an increasing function whose image includes the whole interval $[0, 1)$. So, for any $p \in [0, 1)$ , we can find $r$ such that $g(r) = p$.*

2. *Now take the cell total to be*

$$x := \frac{r^{N+1} - r}{r - 1} \tag{12}$$

*and consider a cell $X_{ns}$ whose contributions are*

$$x_i = \begin{cases} r & \text{for } i = 1 \\ r^2 - \varepsilon & \text{for } i = 2 \\ r^3 + \varepsilon & \text{for } i = 3 \\ r^i & \text{for } i = 4, \cdots, N \end{cases}$$

*where $\varepsilon := (r^2 - r^3)/3$. With this choice of $\varepsilon$, one still has $x_2 > x_3$.*

3. *From the first condition in Expression (11),*

$$rg(r) = r^3 + r^4 + \cdots + r^N = x - r - r^2 = pr \tag{13}$$

*so that*

$$x - x_1 - x_2 = x - r - r^2 + \varepsilon > px_1 = pr$$

*Thus, the cell is clearly non-sensitive according to the $p\%$-rule.*

A sensitive cell can be constructed as follows.

**Construction 4 (Sensitive cell $X_s$)**

1. *Take the same values $p, r, N$ and $x$ used in Construction 3.*
2. *Consider a cell $X_s$ whose contributions are*

$$x_i = \begin{cases} r & \text{for } i = 1 \\ r^2 & \text{for } i = 2 \\ pr/(N-2) & \text{for } i = 3, \cdots, N \end{cases}$$

   *Note that the contributions to $X_s$ add to $x$, because*

$$\sum_{i=1}^{N} x_i = r + r^2 + pr = x \tag{14}$$

   *where the last equality is obtained using Equation (13).*

3. *From Equation (14),*

$$x - x_2 - x_1 = x - r - r^2 = pr = px_1$$

   *Thus, $X_s$ is declared sensitive by the p%-rule.*

We next show that the cell declared non-sensitive by the $p$%-rule can yield a closer upper bound for the largest contribution than the cell declared sensitive.

**Theorem 3** *Let $p$ be the parameter of the p%-rule. Assume a coalition of the two second largest contributors want to upper-bound the largest contribution. For any $p$, if $N$ is taken large enough, then the coalition gets a proportionally closer upper bound for the case of $X_{ns}$ than for the case of $X_s$.*

**Proof**. For both $X_s$ and $X_{ns}$, the two second largest contributors know that the largest contribution is upper-bounded by the total $x$ minus their own contributions, that is

$$x_1 \leq x - x_2 - x_3 \tag{15}$$

Now, the distance between $x_1$ and the upper bound (15) is exactly the sum of the $N-3$ smallest contributions. We next show that, for large enough $N$, this sum is smaller for $X_{ns}$ than for $X_s$ (since $x_1$ is the same for both cells, this is equivalent to showing that the upper bound on the largest contribution is proportionally closer for $X_{ns}$). For $X_{ns}$, the $N-3$ smallest contributions add to

$$\sum_{i=4}^{N} x_i = x - r - (r^2 - \varepsilon) - (r^3 + \varepsilon) = pr - r^3 < pr \tag{16}$$

For $X_s$, the $N-3$ smallest contributions add to

$$\sum_{i=4}^{N} x_i = pr\frac{N-3}{N-2} \tag{17}$$

As $N \to \infty$, Expression (17) approaches $pr$ and becomes thus larger than Expression (16). Thus, the theorem holds for large enough $N$. QED.

The following example illustrates that $N$ does not need to be very large for Theorem 3 to hold.

**Example 4** *For $p = 0.4375$ and $N = 5$, we have $r = 0.5$. The largest contribution is $x_1 = 0.5$ for both $X_s$ and $X_{ns}$. The tail of the $N-3$ smallest relative contributions is 0.09677 for $X_{ns}$ and 0.145833 for $X_s$.*

### 3.1. *Replacing the $p\%$ and $pq$-rules*

The weakness pointed out by Theorem 3 is due to the fact that neither the $p\%$-rule nor the $pq$-rule reflect the concentration of contributions in a proper way. Like it happened with the dominance rule, a cell $X_{ns}$ declared non-sensitive can have a smaller tail than a cell $X_s$ declared sensitive; this causes $X_{ns}$ cell to yield a better bound for the largest contribution. However, the situation is somewhat safer here than for the dominance rule: with the $p\%$-rule, no single contributor can estimate the largest contribution better for a non-sensitive cell than for a sensitive cell (at least two contributors must collude, which is more unlikely than a single-contributor intrusion).

A check whether the sum of a number of smallest contributions is below some threshold is not a good alternative sensitivity rule for the $p\%$ rule for the same arguments it was not a good replacement for the dominance rule. Note that the $p\%$-rule as discussed above actually consists of checking whether the $N-2$ smallest contributions are below $px_1$. For any number $N-c$ of smallest contributions considered, one could construct cells $X'_{ns}, X'_s$ such that a result analogous to Theorem 3 holds for a coalition of the $c$ second largest contributors.

We show below that the entropy of the relative contributions is a sensitivity measure which, for large enough $N$, declares $X_{ns}$ more sensitive than $X_s$. We first compute the entropies of both cells:

- The entropy of the relative contributions of $X_{ns}$ is

$$H(X_{ns}) = -(r/x)\log_2(r/x) - ((r^2-\varepsilon)/x)\log_2((r^2-\varepsilon)/x)$$
$$-((r^3+\varepsilon)/x)\log_2((r^3+\varepsilon)/x) - \sum_{i=4}^{N}(r^i/x)\log_2(r^i/x) \qquad (18)$$

- The entropy of the relative contributions of $X_s$ is

$$H(X_s) = -\frac{r}{x}\log_2(\frac{r}{x}) - \frac{r^2}{x}\log_2(\frac{r^2}{x}) - \frac{pr}{x}\log_2(\frac{pr}{x(N-2)}) \qquad (19)$$

The proof of the theorem below can be found in the Appendix.

**Theorem 4** *Let the proportion $p$ be the parameter of the $p\%$-rule. Let $N$ be the number of contributors to cells $X_s$ and $X_{ns}$. For any $p$, if $N$ is taken large enough, then $H(X_{ns}) < H(X_s)$ holds.*

Thus, if the $p$%-rule is replaced by entropy as a sensitivity measure, $X_{ns}$ is declared as more sensitive than $X_s$ for large enough $N$, which is consistent with Theorem 3. The following example illustrates Theorem 4 for the same parameters used in Example 4.

**Example 5** *For* $p = 0.4375$ *and* $N = 5$*, we have* $r = 0.5$ *and*

$$H(X_{ns}) = 1.82108 < H(X_s) = 1.83946$$

Thus, we propose again to use the entropy-based sensitivity rule given by Algorithm 1 to replace the $p$%-rule. To replace the $pq$-rule, prior knowledge by the intruder must be taken into account, so one might think of substituting conditional entropy for plain entropy in Algorithm 1 (entropy would be conditional to the prior knowledge by the intruder).

## 4. Conclusions and recommendations

We have constructed general counterexamples that show that none of the most widely accepted sensitivity rules for tabular protection ($(n, k)$ dominance, $p$% and —by extension— $pq$) captures disclosure risk in a proper way when coalitions of cell contributors are a realistic scenario. Further, for any value of $k$, the $(1, k)$-dominance rule is problematic even without coalitions: the second largest cell contributor can compute without help a bound on the largest contribution which might be more accurate for a cell labeled as non-sensitive than for a cell labeled as sensitive. The explanation of the above misbehaviors is that the rules in use fail to adequately reflect the concentration of contributions.

Possible and non-exclusive solutions when contributors should be regarded as potential intruders include:

- Replace sensitivity rules in use by a new entropy-based sensitivity rule (given by Algorithm 1). This new rule measures the concentration of relative contributions and has been shown to work well for the counterexamples proposed.
- Complement *a priori* risk assessment provided by sensitivity rules with *a posteriori* 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 sensitivity rules (the shuttle algorithm described in [1] can be used for that).

## References

1. L. Buzzigoli and A. Giusti, "An algorithm to calculate the lower and upper bounds of the elements of an array given its marginals", in *Proceedings of the Conference on Statistical Data Protection*. Luxemburg: Eurostat, pp. 131-147, 1999.
2. L. H. Cox, "Disclosure 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.

3. 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.
4. S. Gießing, "Nonperturbative disclosur e 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.
5. 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.
6. 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.
7. 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.

## Appendix A

**Proof (Theorem 2):** To compare Expressions (8) and (9) we must check whether

$$-(R_n-\varepsilon)\log_2(R_n-\varepsilon)-(R_{n+1}+\varepsilon)\log_2(R_{n+1}+\varepsilon)-\sum_{i=n+2}^{N}R_i\log_2(R_i)\stackrel{?}{<}$$

$$\stackrel{?}{<}-R_n\log_2(R_n)-(1-k)\log_2(\frac{1-k}{N-n})$$

The above inequality holds if the following inequality holds

$$R_n\log_2(\frac{1}{R_n-\varepsilon})+\varepsilon\log_2(\frac{R_n-\varepsilon}{R_{n+1}+\varepsilon})+\sum_{i=n+1}^{N}R_i\log_2(\frac{1}{R_i})\stackrel{?}{<}$$

$$\stackrel{?}{<}R_n\log_2(\frac{1}{R_n})+(1-k)\log_2(\frac{N-n}{1-k}) \tag{A.1}$$

Now, all terms on both sides of Inequality (A.1) are positive (the arguments of logs being all greater than one). However, as $N$ increases, the second term on the right-hand side increases more rapidly than the third term on the left-hand side (increasing $N$ on the left-hand side results only in extending the tail of a geometric series). Thus, for large enough $N$, the theorem holds. QED

**Proof (Theorem 4):** To compare Expressions (18) and (19), we can simplify computations by replacing $x$ with 1, which amounts to replacing relative contributions by absolute contributions (since both $X_s$ and $X_{ns}$ add to $x$, this does not affect the comparison). Our goal is to check whether

$$-r\log_2(r)-(r^2-\varepsilon)\log_2(r^2-\varepsilon)-(r^3+\varepsilon)\log_2(r^3+\varepsilon)-\sum_{i=4}^{N}r^i\log_2(r^i)\stackrel{?}{<}$$

$$\stackrel{?}{<} -r\log_2(r) - r^2\log_2(r^2) - pr\log_2(\frac{pr}{N-2})$$

The above

$$r^2\log_2(\frac{1}{r^2-\varepsilon}) + \varepsilon\log_2(\frac{r^2-\varepsilon}{r^3+\varepsilon})\sum_{i=3}^{N} r^i\log_2(\frac{1}{r^i}) \stackrel{?}{<}$$

$$\stackrel{?}{<} r^2\log_2(\frac{1}{r^2}) + pr\log_2(\frac{N-2}{pr}) \qquad \text{(A.2)}$$

Now, all terms on both sides of Inequality (A.2) are positive (the arguments of logs being all greater than one). However, as $N$ increases, the second term on the right-hand side increases more rapidly than the third term on the left-hand side (increasing $N$ on the left-hand side results only in extending the tail of a geometric series). Thus, for large enough $N$, the theorem holds. QED