A Unified Mathematical Framework for different Statistical Disclosure Limitation Methods for Tabular Data *

Juan-José Salazar-González
DEIOC, Facultad de Matemáticas, Universidad de La Laguna,
38271 La Laguna, Tenerife, Spain.
Fax: + 34 922 318170; e-mail: jjsalaza@ull.es

June 2002

Abstract

This paper concerns with Statistical Disclosure Control methods to minimize the information loss while keeping small the disclosure risk from different data snoopers. This issue is of primary importance in practice for statistical agencies when publishing data. A common definition of protection is introduced and used in four different methodologies. In particular, two Integer Linear Programming models are described for the well-known *Cell Suppression* and *Controlled Rounding* techniques. Also two relaxed techniques are presented through two associated Linear Programming models, called *Interval Publication* and *Cell Perturbation*, respectively. A final discussion shows how to combine the four methods and how to implement a cutting-plane approach for the exact and heuristic resolution of the combinatorial problems in practice. All the presented methodologies inherently guarantee protection levels on all cells and against a set of different intruders (possibly respondents), thus the standard post-phase to test the protection requirements (and typically named *Disclosure Auditing*) is unnecessary.

Keywords: Statistical Disclosure Control; Cell Suppression; Rounding; Integer Linear Programming

1 Introduction

Statistical agencies are often required by law or policy to protect the confidentiality of the information that they collect from persons, businesses, or other units. The *microdata* is the collection

*Juan José Salazar González is Associate Professor of Mathematical Programming, "Departamento de Estadística, Investigación Operativa y Computación", University of La Laguna, Av. Astrofísico Francisco Sánchez s/n, 38271 Tenerife, Spain (E-mail: jjsalaza@ull.es). This work was partially supported by the European Union IST-2000-25069, entitled "Computational Aspects of Statistical Confidentiality" (CASC) and coordinated by Anco Hundepool (Central Bureau of Statistics, The Netherlands). The author thanks David Brown (Office for National Statistics, England), Larry Cox (National Center for Health Statistics, U.S.A.), Ramesh Dandekar (Department of Energy, U.S.A.), Sarah Giessing (Statistisches Bundesamt, Germany), Dale Robertson (Statistics Canada, Canada) and Laura Zayatz (Census of Bureau, U.S.A.) for fruitful comments on Statistical Disclosure Control.

of all the individual responses, and a *statistical table* is the aggregation of one variable according to other variables and including marginal sums. Before releasing statistical tables (or microdata files), these agencies use a variety of statistical methods to protect their data and to ensure that the risk of disclosure is controlled and very small. In essence, statistical agencies protect the confidentiality of the data that they collect by restricting the amount of information in tabular data products (or microdata) that they release. Therefore, a common characteristic of all the methodologies is that they reduce the information to limit disclosure risk, but with the aim of minimizing the loss of information. There are methodologies to protect microdata and others to protect statistical tables. This paper concerns only with methodologies to protect statistical tables directly, i.e. modifying the table itself and not the original microdata. The importance of protecting tabular data has been clearly stated by governments awarding contracts to conduct research and issue reports on Disclosure Limitation Methods for Tabular Data Protection. For example, the *National Institute of Statistical Sciences* is supporting the U.S. project entitled "Digital Government", and *EUROSTAT* is coordinating the E.U. project entitled "Computational Aspects on Statistical Confidentiality", both addressing the protection of tabular data (among other topics).

The most popular methodologies for protecting an statistical table are variants of the well-known Cell Suppression and Controlled Rounding methods. Nevertheless, the different methodologies are usually applied by practitioners without sharing similar hypothesis, thus making very hard a comparative even on the same data. Even more, in practice, some implementations cannot inherently guarantee the protection requirements and a heavy computational effort must be applied to check the proposed release. This checking is called *Disclosure Auditing* phase. See, e.g., Willenborg and de Waal [28] for a wider introduction to the Statistical Data Protection. This paper presents a common framework (including concepts, models and algorithms) for several methodologies that implicitly guarantee the required protections on different cells and against different attackers. Section 1 introduces the main concepts of the Statistical Disclosure Control problem in a general context. Section 2 considers the well-known *Cell Suppression Methodology*, and Section 3 points out a relaxed version here refereed as *Interval Publication methodology*. Section 4 deals with the *Controlled Rounding Methodology* and Section 6 proposes a relaxed version named *Cell Perturbation Method*. For each version two mathematical models are described emphasizing the common definitions and features. The paper ends with some conclusions leading to an all-in-one methodology.

This work has been presented in several seminars on Disclosure Limitation Methods (Plymouth, April 2002; Ottawa, May 2002; Washington, June 2002).

	A	B	C	Total
Activity I	20	50	10	80
Activity II	8	19	**22**	49
Activity III	17	32	12	61
Total	45	101	44	190

Figure 1: Investment of enterprises by activity and region.

2 General Situation

A statistical agency is typically given with a set of n values $a = [a_i : i \in I]$, where $I := \{1, \dots, n\}$. Vector a is known as "nominal table" and satisfies a set of m equations $\sum_{i \in I} m_{ij} y_i = b_j$ for $j \in J$, where $J := \{1, \dots, m\}$. For convenience of notation the linear system will be denoted by $My = b$, thus $Ma = b$ holds. Each solution y of $My = b$ is called *congruent table*. Matrix M (with n columns representing the cells and m rows representing the equations) has typically elements m_{ij} in $\{-1, 0, +1\}$ with one -1 per row associated to the marginal-cell variable, while vector b is typically the zero vector. Table in Figure 1 is a 2-dimensional table consisting on $n := 16$ cells and $m := 8$ equations (one from each row and from each column in the table). When the table is a 2-dimensional table, then M is the edge-node matrix of a bipartite graph (see Cox [8]), thus a congruent table can be represented as a flow circulation in a network and some tools from Graph Theory can be applied. This is not the case when M is associated to a more complex table.

On statistical tables there could be some sensitive data, i.e., information that cannot be disclosured since they show confidential information on particular respondents. The sensitive cells in a tabular data are typically determined by common-sense *rules*. A typical rule is the so-called *dominance rule* (see, e.g., [28]), described as follows. We are given with the microdata from which the table is computed, and with two input numbers α and β (e.g. $\alpha := 80$ and $\beta := 3$). Whenever the biggest β respondents from the microdata to value in cell p of the table produce more than α percentage of the total value a_p, then cell p is classified as *sensitive*. We denote the subset of sensitive cells by P. In the example represented in Figure 1, cell in Activity II and Region C is assumed to be a sensitive cell to be protected because (say) it is publicly known that there is only one respondent in Region C dedicated to Activity II.

In a general situation, all the sensitive cells in a table must be protected against a set of *attackers*. The attackers are the intruder or data snoopers that will analyze the final product data and will try to disclosure confidential information. The aim of the Disclosure Limitation Methods is to reduce the risk that they succeed. The set of attackers will be denoted by K. Each attacker knows the set of linear system $My = b$ plus extra information that bound each cell value.

For example, the simplest attacker is the so-called *external intruder* knowing only that unknown cell values are (say) nonnegative. Other more accurate attackers known tighter bounds on the cell values, and they are called *internal attackers*. For example, an internal attacker could be a respondent that had contributed to cell i with (say) 10 units; then he/she knows that $y_i \geq 10$, while the external attacker only knows $y_i \geq 0$. If the internal attacker also knows that he/she is the only contributor to cell i with value 10, then $10 \leq y_i \leq 10$ when attacking the output data. In general, attacker k is associated with two bounds lb_i^k and ub_i^k such that $a_i \in [lb_i^k \dots ub_i^k]$ for each cell $i \in I$. Literature on statistical disclosure control (see, e.g., Willenborg and de Waal [28]) typically addresses the situation where $|K| = 1$, thus protecting the table against the external intruder with the only knowledge of the linear system and some external bounds; nevertheless this is a simplification of the real problem in Disclosure Limitation and statistical offices are interested in protecting tables against several intruders (see, e.g., Jewett [18]).

To protect sensitive cell p containing value a_i in a input table, the statistical office is interested in publishing an output containing several congruent tables, including the original nominal table but also others such that no attacker can disclosure private information. The output of a Disclosure Limitation Method is generally named *pattern*, and it can assume a particular structure depending on the methodology considered. The sections of this paper deal with several methodologies, and hence illustrate different patterns. In all cases they share the following common definition of "protection" defined as follows.

The congruent tables associated to a pattern must differ so as each attacker analyzing the pattern will not compute the original value of a sensitive cell within a narrow approximation. For each potential intruder, the idea is to define a protection range for p and to demand that protection be such that any value in the range is potentially the correct cell value. To be more precise, by observing the published pattern, attacker k will compute an interval $[\underline{y}_p^k \dots \overline{y}_p^k]$ of possible values for each sensitive cell p. The pattern will be considered *valid* to protect cell p against attacker k if the computed interval is "wide enough". To set up the definition of "wide enough" in a precise way, the statistical office gives three input parameters for each attacker k and each sensitive cell p with nominal value a_p:

- Upper Protection Level: it is a number UPL_p^k representing the minimum value for $\overline{y}_p^k - a_p$;
- Lower Protection Level: it is a number LPL_p^k representing the minimum value for $a_p - \underline{y}_p^k$;
- Sliding Protection Level: it is a number SPL_p^k representing the minimum value for $\overline{y}_p^k - \underline{y}_p^k$.

The values of this parameters can be also defined by using common-sense rules. For example, simple values for the protection levels are percentages of the nominal value of the cell (e.g., 20%, 15% and 40%, respectively). In more sophisticated situations where intruder k is an original respondent (i.e., an internal attacker), the protection levels could be chosen to be proportional to his/her contributions s_p^k to the nominal value of the cell a_p and/or to the complement $a_p - s_p^k$ (see, e.g, Cox [5], Robertson [25], Sande [26]). Of course, an elementary assumption is that

$$lb_p^k \leq a_p - LPL_p^k \leq a_p \leq a_p + UPL_p^k \leq ub_p^k$$

and

$$ub_p^k - lb_p^k \geq SPL_p^k,$$

for each attacker k and each sensitive cell p. For notational convenience, let us also define absolute protection levels and relative nominal bounds:

$$lpl_p^k := a_p - LPL_p^k,$$

$$upl_p^k := a_p + UPL_p^k,$$

$$LB_i^k := a_i - lb_i^k,$$

$$UB_i^k := ub_i^k - a_i.$$

In the example represented in Figure 1 the statistical office could be interested in protecting the sensitive cell (Activity II, Region C) against one attacker with a lower protection level of 10 units, an upper protection level of 12 units, and a sliding protection level of 0 units.

Given a pattern, the mathematical problems of computing values $\underline{y}_p^k$ and $\overline{y}_p^k$ are known as *attacker problems* for cell p and attacker k. The overall problem of solving the attacker problems for all cells is named as *Disclosure Auditing Problem*. The attacker problems associated with cell p and attacker k can be formulated as two Linear Programming (LP) models on an array of variables $y = [y_i : i \in I]$ representing a table. Indeed, an attacker problem is

$$\underline{y}_p^k := \min y_i$$

subject to

$$\begin{array}{ll} My = b & \\ lb_i^k \leq y_i \leq ub_i^k & \qquad \text{for all } i \in I \end{array}$$

plus a set of additional constraints that make y feasible according to the published pattern. The precise additional constraints depend on the structure of the pattern, and therefore on the considered methodology. The other attacker problem is obtained by replacing the objective function with $\overline{y}_p^k := \max y_i$. Each section of this paper shows the precise attacker problems for each methodology.

	A	B	C	Total
Activity I	20	50	10	80
Activity II	*	19	*	49
Activity III	*	32	*	61
Total	45	101	44	190

Figure 2: Cell Suppression pattern.

Finally, among all possible valid patterns, the statistical office is interested in finding one with minimum information loss. The *information loss* of a pattern is intended to be a measure of the number of congruent tables in the pattern. Indeed, a valid pattern must always allow the nominal table to be a congruent table feasible with it, but it must contains also other different congruent tables so to keep the risk of disclosure controlled. For example, when the pattern contains only the original table (because there is no sensitive data to be protected) then the loss of information is clearly zero. The precise definition of loss of information depends on the structure of the pattern, and hence on the methodology to be considered.

In practice most of the available software are based on techniques to find "good" patterns with no inherently guarantee on the protection level requirements, i.e. not necessarily valid (see, e.g., [11]). Therefore, it is necessary to check the proposed pattern before it is made public by solving the Disclosure Auditing Problem, and to try a different technique when the result is negative. It is well-known (see, e.g., Doyle, Lane, Theeuwes and Zayatz [11]) that auditing a pattern could consume many computing resources. In the next sections we introduce precise methodologies to find a valid pattern (if any exists) with minimum (or near-minimum) information loss, hence the Disclosure Auditing Problem is not required.

3 Cell Suppression Methodology

Cell suppression is one of the most popular techniques for protecting sensitive information in statistical tables. The standard Cell Suppression technique is based on the idea of protecting the sensitive information by hiding (*suppressing*) the values of some cells with a symbol (e.g. $*$). Obviously, the sensitive cells must be suppressed and they are named *primary suppressions*, but also other cells must be suppressed and they are named *secondary suppressions*.

A pattern in Cell Suppression is then defined by a subset of cells SUP to be unpublished. Obviously, $P \subseteq SUP$. Then, the feasible region for the attacker problems associated to attacker k

is defined by

$$\begin{aligned} My &= b & & \\ y_i &= a_i & & \text{if } i \notin SUP \\ lb_i^k \leq y_i &\leq ub_i^k & & \text{if } i \in SUP. \end{aligned}$$

Figure 2 illustrates a pattern for the instance in Figure 1, where SUP is the set of cells containing an asterisk. Assuming that there is one attacker who knows that each missing value is a non-negative number (i.e., $lb_i^k = 0$ and $ub_i^k = \infty$), then the minimum value $\underline{y}_{II,C}$ for the sensitive cell in row II and column C can be computed by solving an LP model in which the values $y_{i,j}$ for the suppressed cells in row i and column j are treated as unknowns, namely

$$\underline{y}_{II,C} := \min y_{II,C}$$

subject to

$$\begin{array}{ccccc} y_{II,A} & & +y_{II,C} & & = 30 \\ & y_{III,A} & & +y_{III,C} & = 29 \\ y_{II,A} & +y_{III,A} & & & = 25 \\ & & y_{II,C} & +y_{III,C} & = 34 \end{array}$$

$$y_{II,A} \geq 0 \ , \ y_{III,A} \geq 0 \ , \ y_{II,C} \geq 0 \ , \ y_{III,C} \geq 0.$$

Notice that the right-hand-side values are known to the attacker, as they can be obtained as the difference between the marginal and the published values in a row/column.

The maximum value $\overline{y}_{II,C}$ for the sensitive cell can be computed in a perfectly analogous way, by solving the LP model maximizing $y_{II,C}$ subject to the same constraints as before. Notice that each solution of this common set of constraints is a congruent table according with the published suppression pattern in Figure 2 and with the extra knowledge of the external bounds (non-negativity on this example).

In the example, $\underline{y}_{II,C} = 5$ and $\overline{y}_{II,C} = 30$, i.e., the sensitive information is "protected" within the *protection interval* $[5 \ldots 30]$. If this interval is considered sufficiently wide by the statistical office according to the protection level requirements, then the pattern in Figure 2 is valid; otherwise, different complementary suppressions are needed.

For the definition of the loss of information of a Cell Suppression pattern, it is given an estimation w_i of the loss of information when cell i (with nominal value a_i) is not published. Typical definitions of w_i (see, e.g., [28]) are

- $w_i := a_i$,
- $w_i := 1$,

- $w_i := \log(a_i)$,
- $w_i :=$ the number of responses in the microdata contributing to value a_i of cell i,
- $w_i :=$ a (linear) combination of the above criteria.

Then the loss of information of a pattern determined by SUP is defined as sum of w_i for all $i \in SUP$.

The problem of finding a valid pattern with minimum loss of information is a very difficult combinatorial problem, known as *Cell Suppression Problem* (CSP, for short). The task is so complex that there are in literature mainly heuristic approaches (i.e., procedures providing approximated —probably overprotected— suppression patterns) for special situations. For example, a relevant situation occurs when there is an entity which contributes to several cells, leading to the so-called *common respondent problem*. Possible simplifications valid for this situation consist on replacing all the different attackers by one stronger intruder with "protection capacities" (see, e.g., Jewett [18] or Sande [27] for details), or on aggregating some sensitive cells into new "union" cells with stronger protection level requirements (see, e.g., Robertson [25]). But even the relaxation that consider one intruder (an external attacker) is an strongly $\mathcal{NP}$-hard problem (see, e.g., Kelly et al. [22]), meaning that it is very unlikely the existence of an algorithm for the exact solution of CSP which guarantees an efficient (i.e., polynomial-time) performance for all possible input instances. Previous work on the classical CSP from the literature mainly concentrate on 2-dimensional tables with marginals and protection against a single attacker. Heuristic solution procedures have been proposed by several authors, including Cox [4, 8], Sande [26], Kelly et al. [22], and Carvalho et al. [2]. Kelly [19] proposed a mixed ILP formulation involving a huge number of variables and constraints (for instance, the formulation involves more than 20,000,000 variables and 30,000,000 constraints for a two-dimensional table with 100 rows, 100 columns and 5% sensitive entries). Geurts [17] refined this model, and reported computational experiences on small-size instances, the largest instance solved to optimality being a table with 20 rows, 6 columns and 17 sensitive cells. Heuristics for 3-dimensional tables have been proposed in Robertson [24], Sande [26], and Dellaert and Luijten [9]. Fischetti and Salazar [15] proposed a new method capable of solving to proven optimality, on a personal computer, 2-dimensional tables with about 250,000 cells and 10,000 sensitive entries. An extension of this methodology capable of solving to proven optimality real-world 3- and 4-dimensional tables is presented in Fischetti and Salazar [15], but always protecting against one intruder (the external attacker). The following section extends the methodologies in [15] to deal with multiple intruders.

3.1 Mathematical Model

Let us consider a binary variable x_i associated to each cell $i \in I$, assuming value 1 if such cell must be suppressed in the released pattern, or 0 otherwise. Notice that attacker k will minimize and maximize unknown values on the set of consistent tables in the pattern, defined by:

$$\begin{aligned} My &= b \\ lb_i^k \le y_i &\le ub_i^k \qquad \text{for all } i \in I : x_i = 1 \\ y_i &= a_i \qquad \text{for all } i \in I : x_i = 0, \end{aligned}$$

equivalently represented as the solution set of the following linear system:

$$\left. \begin{aligned} My &= b \\ a_i - LB_i^k x_i \le y_i &\le a_i + UB_i^k x_i \qquad \text{for all } i \in I. \end{aligned} \right\} \tag{1}$$

Therefore, the CSP optimization problem is to find a value for each x_i such that the total loss of the information in the released pattern is minimized, i.e.:

$$\min \sum_{i \in I} w_i x_i \tag{2}$$

subject to, for each sensitive cell $p \in P$ and for each attacker $k \in K$,

- the upper protection requirement must be satisfied, i.e.:

$$\max \{ y_p : (1) \text{ holds } \} \ge upl_p^k \tag{3}$$

- the lower protection requirement must be satisfied, i.e.:

$$\min \{ y_p : (1) \text{ holds } \} \le lpl_p^k \tag{4}$$

- the sliding protection requirement must be satisfied, i.e.:

$$\max \{ y_p : (1) \text{ holds } \} - \min \{ y_p : (1) \text{ holds } \} \ge SPL_p^k \tag{5}$$

Finally, each variable must assume value 0 or 1, i.e.:

$$x_i \in \{0, 1\} \qquad \text{for all } i \in I. \tag{6}$$

Mathematical model (2)–(6) contains all the requirements of the statistical office (according with the definition given in Section 1), and therefore a solution $[x_i^* : i \in I]$ defines an optimal valid Cell Suppression pattern. The inconvenient is that it is not an easy model to be solved, since it does not belong to the standard (Mixed) Integer Linear Programming (ILP). In fact, the existence of optimization problems as part of the constraints of a main optimization problem classifies the model in the so-called "Bilevel Mathematical Programming", which today is not provided with efficient algorithms to solve the model (2)–(6) even on instances of small size. Observe that the inconvenience of model (2)–(6) is not the number of variables (which is at most the number of cells, both for the master optimization problem and for each subproblem in the second level), but the fact there are nested optimization problems in two levels. The better way to avoid the direct resolution is to look for a transformation into a classical ILP model, as the following section shows.

3.2 A first model

A first idea arises by observing that the optimization problem in condition (3) can be replaced by the existence of a congruent table $[f_i^{kp} : i \in I]$ such that it is feasible (i.e., it satisfies (1)) and it guarantees the upper protection level requirement, i.e.:

$$f_p^{kp} \geq upl_p^k.$$

In the same way, the optimization problem in condition (4) can be replaced by the existence of a congruent table $[g_i^{kp} : i \in I]$ such that it is also feasible (i.e., it satisfies (1)) and it guarantees the lower protection level requirement, i.e.:

$$g_p^{kp} \leq lpl_p^k.$$

Finally, the two optimization problems in condition (5) can be replaced by the above congruent tables if they guarantee the sliding protection level, i.e.:

$$f_p^{kp} - g_p^{kp} \geq SPL_p^k.$$

Figure 3 shows a first attempt to have an ILP model, where x_i, f_i^{kp}, g_i^{kp} are the variables.

Clearly, this new model is an ILP model, and therefore —in theory— there are efficient approaches to solve it. Nevertheless, the number of new variables (f_i^{kp} and g_i^{kp}) is really huge even on small tables. For example, the model associated with a table with 100×100 cells, with 1% sensitive, and 100 attackers would have millions of variables. Therefore, it is necessary another approach to transform model (2)–(6) into an ILP model without adding so many additional variables.

3.3 A second model

An alternative approach which does not add any additional variable follows the idea described in Fischetti and Salazar [15] for the Cell Suppression problem against one attacker.

$$\min \sum_{i\in I} w_i x_i$$

subject to:

$$x_i \in \{0,1\} \qquad \text{for all } i \in I$$

and, for all $p \in P$ and all $k \in K$:

$$\textstyle\sum_{i\in I} m_{ij} f_i^{kp} = b_j \qquad \text{for all } j \in J$$

$$a_i - LB_i^k x_i \le f_i^{kp} \le a_i + UB_i^k x_i \qquad \text{for all } i \in I$$

$$\textstyle\sum_{i\in I} m_{ij} g_i^{kp} = b_j \qquad \text{for all } j \in J$$

$$a_i - LB_i^k x_i \le g_i^{kp} \le a_i + UB_i^k x_i \qquad \text{for all } i \in I$$

$$f_p^{kp} \ge upl_p^k$$

$$g_p^{kp} \le lpl_p^k$$

$$f_p^{kp} - g_p^{kp} \ge SPL_p^k.$$

Figure 3: First ILP model for Cell Suppression.

Imposing the upper protection level requirements

Based on the Farkas' Lemma, it is possible to replace the second level subproblems of model (2)–(6) by linear constraints on the x_i variables. Indeed, assuming that values y_i in a congruent table are continuous numbers, the two LP models in conditions (3)–(5) can be rewritten in their dual format. More precisely, by Dual Theory in Linear Programming (see, e.g., Wolsey [29]):

$$\max \{ y_p : (1) \text{ holds } \}$$

is equivalent to

$$\min \sum_{j\in J} \gamma_j b_j + \sum_{i\in I} [\alpha_i (a_i + UB_i^k x_i) - \beta_i (a_i - LB_i^k x_i)]$$

subject to

$$\left.\begin{array}{rl} \alpha_p - \beta_p + \sum_{j\in J} m_{pj}\gamma_j = 1 & \\ \alpha_i - \beta_i + \sum_{j\in J} m_{ij}\gamma_j = 0 & \text{for all } i \in I \setminus \{p\} \\ \alpha_i \ge 0 & \text{for all } i \in I \\ \beta_i \ge 0 & \text{for all } i \in I \\ \gamma_j \text{ unrestricted in sign} & \text{for all } j \in J. \end{array}\right\} \qquad (7)$$

Because of (7) and $[a_i : i \in I]$ is a consistent table, we have

$$\sum_{j \in J} \gamma_j b_j + \sum_{i \in I} (\alpha_i a_i - \beta_i a_i) = \sum_{i \in I} \sum_{j \in J} \gamma_j m_{ij} a_i + \sum_{i \in I} (\alpha_i - \beta_i) a_i = a_p.$$

Hence the above LP model can be rewritten as

$$a_p + \min \sum_{i \in I} (\alpha_i UB_i^k + \beta_i LB_i^k) x_i$$

subject to $\alpha_i, \beta_i, \gamma_j$ satisfying (7).

From this observation, condition (3) can be now written as:

$$\sum_{i \in I} (\alpha_i UB_i^k + \beta_i LB_i^k) x_i \geq UPL_p^k \qquad \text{for all } \alpha_i, \beta_i, \gamma_j \text{ satisfying (7).}$$

In other words, the last system defines a family of linear constraints, in the x-variables only, representing the condition (3) which concerns with the upper protection level requirement for sensitive cell p and attacker k.

Notice that this family contains in principle an infinite number of constraints, each associated with a different point $[\alpha_i : i \in I; \beta_i : i \in I; \gamma_j : j \in J]$ of the polyhedron defined by (7). However, it is well known that only the extreme points (and rays) of such polyhedron can lead to undominated constraints, i.e., a finite number of such constraints is sufficient to impose the upper protection level requirement for a given sensitive cell p and a given attacker k. Again, there is in infinite number of points in (7), but only the one corresponding to extreme points of the polyhedron (7), and it is well-known that this is a finite number (see, e.g., Wolsey [29]).

Imposing the lower protection level requirements

In a similar way, the optimization problem in (4) is:

$$-\max\left\{-y_p : (1) \text{ holds }\right\},$$

which, by Duality Theory, is equivalent to

$$-\min \sum_{j \in J} \gamma_j b_j + \sum_{i \in I} [\alpha_i (a_i + UB_i^k x_i) - \beta_i (a_i - LB_i^k x_i)]$$

subject to

$$\left.\begin{array}{rl} \alpha_p - \beta_p + \sum_{j \in J} m_{pj} \gamma_j = -1 & \\ \alpha_i - \beta_i + \sum_{j \in J} m_{ij} \gamma_j = 0 & \text{for all } i \in I \setminus \{p\} \\ \alpha_i \geq 0 & \text{for all } i \in I \\ \beta_i \geq 0 & \text{for all } i \in I \\ \gamma_j \text{ unrestricted in sign} & \text{for all } j \in J. \end{array}\right\} \qquad (8)$$

Because of (8) and $[a_i : i \in I]$ is a consistent table, we have

$$\sum_{j\in J}\gamma_j b_j + \sum_{i\in I}(\alpha_i a_i - \beta_i a_i) = \sum_{i\in I}\sum_{j\in J}\gamma_j m_{ij} a_i + \sum_{i\in I}(\alpha_i - \beta_i) a_i = a_p.$$

Hence the above linear program can be rewritten as

$$-a_p - \min \sum_{i\in I}(\alpha_i UB_i^k + \beta_i LB_i^k) x_i$$

subject to $\alpha_i, \beta_i, \gamma_j$ satisfying (8).

From this observation, condition (4) can be now written as:

$$\sum_{i\in I}(\alpha_i UB_i^k + \beta_i LB_i^k) x_i \geq LPL_p^k \qquad \text{for all } \alpha_i, \beta_i, \gamma_j \text{ satisfying (8).}$$

In other words, the last system defines a family of linear constraints, in the x-variables only, representing the condition (4) which concerns with the lower protection level requirement for sensitive cell p and attacker k.

Imposing the sliding protection level requirements

As to the sliding protection level for sensitive cell p and attacker k, the requirement is that

$$SPL_p^k \leq \max\{y_p : (1) \text{ hold }\} + \max\{-y_p : (1) \text{ hold }\}.$$

Again, by LP duality, this condition is equivalent to

$$\begin{aligned} SPL_p^k \quad \leq \quad & \min\{\sum_{j\in J}\gamma_j b_j + \sum_{i\in I}[\alpha_i(a_i + UB_i^k x_i) - \beta_i(a_i - LB_i^k x_i)] : (7) \text{ holds }\} + \\ & \min\{\sum_{j\in J}\gamma_j b_j + \sum_{i\in I}[\alpha_i(a_i + UB_i^k x_i) - \beta_i(a_i - LB_i^k x_i)] : (8) \text{ holds }\}. \end{aligned}$$

Therefore, the feasibility condition can now be formulated by requiring

$$SPL_p^k \leq \sum_{j\in J}(\gamma_j + \gamma_j')b_j + \sum_{i\in I}[(\alpha_i + \alpha_i')(a_i + UB_i^k x_i) - (\beta_i + \beta_i')(a_i - LB_i^k x_i)]$$
$$\text{for all } \alpha, \beta, \gamma \text{ satisfying (7) and for all } \alpha', \beta', \gamma' \text{ satisfying (8),}$$

or, equivalently,

$$\sum_{i\in I}[(\alpha_i + \alpha_i')UB_i^k + (\beta_i + \beta_i')LB_i^k] x_i \geq SPL_p^k$$
$$\text{for all } \alpha, \beta, \gamma \text{ satisfying (7) and for all } \alpha', \beta', \gamma' \text{ satisfying (8).}$$

$$\min \sum_{i \in I} w_i x_i$$

subject to:

$$x_i \in \{0, 1\} \qquad \text{for all } i \in I$$

and, for all $p \in P$ and all $k \in K$:

$$\sum_{i \in I} [\alpha_i UB_i^k + \beta_i LB_i^k] x_i \geq UPL_p^k$$
$$\text{for all } \alpha, \beta, \gamma \text{ satisfying (7)}$$

$$\sum_{i \in I} [\alpha_i' UB_i^k + \beta_i' LB_i^k] x_i \geq LPL_p^k$$
$$\text{for all } \alpha', \beta', \gamma' \text{ satisfying (8)}$$

$$\sum_{i \in I} [(\alpha_i + \alpha_i') UB_i^k + (\beta_i + \beta_i') LB_i^k] x_i \geq SPL_p^k$$
$$\text{for all } \alpha, \beta, \gamma \text{ satisfying (7) and}$$
$$\text{for all } \alpha', \beta', \gamma' \text{ satisfying (8).}$$

Figure 4: Second ILP model for Cell Suppression.

Overall model

In conclusion, Figure 4 summarizes an alternative model to (2)–(6) with only the 0-1 variables. The inequalities in the model are called *capacity constraints* in analogy with similar constraints introduced in Fischetti and Salazar [15] for enforcing a sufficient "capacity" of certain cuts in the network representation of problem on 2-dimensional tables with marginals. Intuitively, the capacity constraints force to suppress (i.e., to set $x_i = 1$) a sufficient number of cells whose positions within the table and contributions to the overall protection are specified by the dual variables (α, β, γ) of the attacker subproblems.

Trying to solve ILP models, standards approaches in Mathematical Programming are based on *branch-and-bound* schemes, where the *bound* is based on solving the model without the integrality conditions on the variables through an LP solver. This relaxed model is called called *LP relaxation*, and it is typically strengthened by introducing other valid inequalities. The overall procedure is

	A	B	C	Total
Activity I	$[18\dots24]$	50	$[6\dots12]$	80
Activity II	$[4\dots10]$	19	$[20\dots26]$	49
Activity III	17	32	12	61
Total	45	101	44	190

Figure 5: Interval Publication pattern.

known as *branch-and-cut*. See, e.g., Wolsey [29] for details.

The solution of the LP relaxation of the first model (in Figure 3) can in principle be obtained in polynomial time, as it involves a number of variable and constraints which is (huge but) polynomially bounded in the input size. In practice, however, the model cannot be handled effectively, even for very small instances.

At first glance, the LP relaxation of the second model in (Figure 4) seems even more difficult to solve due to the exponential number of capacity constraints needed. In fact, in the second model the reduction on the number of variables was obtained through the introduction of a really huge number of inequalities, which cannot be handled explicitly by any LP solver. However, for a given CSP instance just a few capacity constraints are typically needed in the LP relaxation to force the protection level requirements. In fact, experience reported in Fischetti and Salazar [15] shows that the number of capacity constraints that need to be incorporated explicitly in the model seldom exceeds 4–5 times the number of sensitive cells, a quite reasonable figure for practical solution approaches. Although we cannot decide in advance which capacity constraints are the relevant ones, we can apply a dynamic constraint-generation technique that identifies them "on the fly", during the solution of a relaxed LP problem.

4 Interval Publication

In the *Interval Publication Methodology* a pattern is a set of intervals, one $[y_i^- \dots y_i^+]$ for each cell i. For each cell i and each value $y_i \in [y_i^- \dots y_i^+]$ there is a congruent tables y' where $y'_i = y_i$ and $y'_l \in [y_l^- \dots y_l^+]$ for all $l \in I$. For the instance in Figure 1 a feasible pattern could be the considered in Figure 5. Then, the feasible region for the attacker problems associated to attacker k is defined by

$$\begin{array}{rl} My = b & \\ y_i^+ \le y_i \le y_i^- & \text{for all } i \in I; \\ lb_i^k \le y_i \le ub_i^k & \text{for all } i \in I. \end{array}$$

The input parameters w_i^+ and w_i^- can be defined by common-sense rules similar to the ones mentioned for the Cell Suppression technique.

The loss of information for publishing $[y_i^- \ldots y_i^+]$ instead of a_i can be measure as a proportion of $a_i - y_i^-$ and of $y_i^+ - a_i$. To this end two input parameters w_i^- and w_i^+ are given for each cell i, and the "loss of information" of a pattern is defined as

$$\sum_{i \in I} w_i^+(y_i^+ - a_i) + w_i^-(a_i - y_i^-).$$

The Interval Publication Methodology (introduced by Fischetti and Salazar [16] against one attacker with the name *Partial Cell Suppression Methodology*) is quite related to the *Cell Suppression Methodology* described in Section 2, but they differ in important features. Indeed, from a Cell Suppression pattern each attacker, after solving the disclosure auditing problem, will replace the missing values by intervals of possible values. Therefore, from an attacker point of view, patterns from both methods could have the same form. Nevertheless, the classical Cell Suppression methodology is a "yes-or-not" approach, where a cell must be published or not, while in the Interval Publication the value of each unsafe cell is replaced by an interval that can be more or less wide. That is why the Interval Protection method was originally named *Partial* Cell Suppression. This extra freedom in Interval Publication has the advantage of providing patterns containing an smaller number of congruent tables than the ones from classical Cell Suppression, and hence increasing the data utility of the pattern to the user. In other words, the set of congruent tables associated to a valid Cell Suppression pattern coincides with the set of congruent tables associated to a valid Interval Publication pattern, but the reverse is not true. Since the region of valid patterns in Interval Publication contains the region of valid patterns in Cell Suppression, one could expect to find solutions with smallest loss of information.

Another important advantage of the Interval Publication methodology is that the optimization problem associated to it (called *Interval Publication Problem*, IPP for short) has a much simpler computational complexity (i.e., it admits a polynomial algorithm). Nevertheless, optimal Interval Publication patterns have the disadvantage of containing more intervals than missing values in optimal Cell Suppression patterns.

The next sections presents two mathematical models generalizing the ones presented in Fischetti and Salazar [16] against one attacker, here extended to allow protection against a set of different attackers.

4.1 A first model

Our first LP model for IPP requires the introduction of two continuous variables z_i^+ and z_i^- for each $i \in I$, which will represent the relative increment and decrement, respectively, of the internal

$$\min \sum_{i \in I} w_i^+ z_i^+ + w_i^- z_i^-$$

subject to:

$$\begin{array}{ll} z_i^+ \geq 0 & \text{for all } i \in I \\ z_i^- \geq 0 & \text{for all } i \in I \end{array}$$

and, for all $p \in P$ and all $k \in K$:

$$\begin{array}{ll}
\sum_{i \in I} m_{ij} f_i^{kp} = b_j & \text{for all } j \in J \\
lb_i^k \leq f_i^{kp} \leq ub_i^k & \text{for all } i \in I \\
a_i - z_i^- \leq f_i^{kp} \leq a_i + z_i^+ & \text{for all } i \in I \\
\sum_{i \in I} m_{ij} g_i^{kp} = b_j & \text{for all } j \in J \\
lb_i^k \leq g_i^{kp} \leq ub_i^k & \text{for all } i \in I \\
a_i - z_i^- \leq g_i^{kp} \leq a_i + z_i^+ & \text{for all } i \in I \\
f_p^{kp} \geq upl_p^k & \\
g_p^{kp} \leq lpl_p^k & \\
f_p^{kp} - g_p^{kp} \geq SPL_p^k. &
\end{array}$$

Figure 6: First ILP model for Interval Publication.

from the nominal value a_i. Hence, the published intervals will be defined by $y_i^+ := a_i + z_i^+$ and $y_i^- := a_i - z_i^-$. Moreover, for each attacker $k \in K$ and each sensitive cell $p \in P$, we need auxiliary continuous variables $f^{kp} = [f_i^k : i \in I]$ and $g^{kp} = [g_i^k : i \in I]$ defining tables which are consistent with the published intervals, and certificating the fulfillment of the protection level requirements. The model then reads as in Figure 6.

The model involves a really huge number of auxiliary variables f_i^{kp} and g_i^{kp} and of linking constraints between the (z^+, z^-) and the auxiliary variables.

4.2 A second model

As in previous section, it is again possible to avoid the explicit introduction of the auxiliary variables f^{kp} and g^{kp} $(k \in K$ and $p \in I)$ along with the associated linking constraints, by using standard LP Duality Theory. Indeed, the feasible region of the attacker problems associated to attacker k is now defined by:

$$\left.\begin{array}{rl} \sum_{i\in I} m_{ij} y_i = b_j & \text{for all } j\in J \\ a_i - LB_i^k \le y_i \le a_i + UB_i^k & \text{for all } i\in I \\ a_i - z_i^- \le y_i \le a_i + z_i^+ & \text{for all } i\in I. \end{array}\right\} \tag{9}$$

By similar algebraic considerations as the derived in Section 2, we obtain the second ILP model illustrated in Figure 7, where

$$\left.\begin{array}{rl} \alpha_p^1 + \alpha_p^2 - \beta_p^1 - \beta_p^2 + \sum_{j\in J} m_{pj}\gamma_j = 1 & \\ \alpha_i^1 + \alpha_i^2 - \beta_i^1 - \beta_i^2 + \sum_{j\in J} m_{ij}\gamma_j = 0 & \text{for all } i\in I\setminus\{p\} \\ \alpha_i^1 \ge 0 & \text{for all } i\in I \\ \alpha_i^2 \ge 0 & \text{for all } i\in I \\ \beta_i^1 \ge 0 & \text{for all } i\in I \\ \beta_i^2 \ge 0 & \text{for all } i\in I \\ \gamma_j \text{ unrestricted in sign} & \text{for all } j\in J, \end{array}\right\} \tag{10}$$

and

$$\left.\begin{array}{rl} \alpha_p^1 + \alpha_p^2 - \beta_p^1 - \beta_p^2 + \sum_{j\in J} m_{pj}\gamma_j = -1 & \\ \alpha_i^1 + \alpha_i^2 - \beta_i^1 - \beta_i^2 + \sum_{j\in J} m_{ij}\gamma_j = 0 & \text{for all } i\in I\setminus\{p\} \\ \alpha_i^1 \ge 0 & \text{for all } i\in I \\ \alpha_i^2 \ge 0 & \text{for all } i\in I \\ \beta_i^1 \ge 0 & \text{for all } i\in I \\ \beta_i^2 \ge 0 & \text{for all } i\in I \\ \gamma_j \text{ unrestricted in sign} & \text{for all } j\in J. \end{array}\right\} \tag{11}$$

Constraints in the second model are named the *capacity constraints* in analogy with the similar constraints of the second Cell Suppression model in Figure 4.

4.3 Combining Cell Suppression and Interval Publication

It is possible to embed the Cell Suppression and the Interval Publication methodologies into one. This combined methodology will have advantages of both approaches since, for example, it will benefic from providing the Interval Publication pattern while it could keep control on the maximum number of proper intervals, and also on their minimum width.

In the unified methodology a pattern is a set of intervals, as in the Interval Publication method, Therefore, for the mathematical description, the variables z_i^+ and z_i^- introduced in Section 4.1 are relevant. But now, it is also useful to consider a binary variable x_i for each cell i, which will assume value 1 if and only if $z_i^+ > 0$ or $z_i^- > 0$. These binary variables play a similar role as the ones introduced in Section 3.1.

It is assumed to be given with three input numbers w_i, w_i^+, w_i^- for each cell i from the statistical office, so the cost of a pattern will be defined by

$$w_i^+ z_i^+ + w_i^- z_i^- + w_i x_i.$$

$$\min \sum_{i \in I} w_i^+ z_i^+ + w_i^- z_i^-$$

subject to:

$$\begin{aligned} z_i^+ &\geq 0 \qquad \text{for all } i \in I \\ z_i^- &\geq 0 \qquad \text{for all } i \in I \end{aligned}$$

and, for all $p \in P$ and all $k \in K$:

$$\sum_{i \in I} \alpha_i^1 UB_i^k + \alpha_i^2 z_i^+ + \beta_i^1 LB_i^k + \beta_i^2 z_i^- \geq UPL_p^k$$
$$\text{for all } \alpha^1, \alpha^2, \beta^1, \beta^2, \gamma \text{ satisfying (10)}$$

$$\sum_{i \in I} \alpha_i'^1 UB_i^k + \alpha_i'^2 z_i^+ + \beta_i'^1 LB_i^k + \beta_i'^2 z_i^- \geq LPL_p^k$$
$$\text{for all } \alpha'^1, \alpha'^2, \beta'^1, \beta'^2, \gamma' \text{ satisfying (11)}$$

$$\sum_{i \in I} (\alpha_i^1 + \alpha_i'^1) UB_i^k + (\alpha_i^2 + \alpha_i'^2) z_i^+ + (\beta_i^1 + \beta_i'^1) LB_i^k + (\beta_i^2 + \beta_i'^2) z_i^- \geq SPL_p^k$$
$$\begin{aligned} &\text{for all } \alpha^1, \alpha^2, \beta^1, \beta^2, \gamma \text{ satisfying (10) and} \\ &\text{for all } \alpha'^1, \alpha'^2, \beta'^1, \beta'^2, \gamma' \text{ satisfying (11).} \end{aligned}$$

Figure 7: Second ILP model for Interval Publication.

Then a first ILP model is similar to the one in Figure 6 where also UB_i^k and LB_i^k must be replaced by $UB_i^k x_i$ and $LB_i^k x_i$, respectively. A second model without the f^{kp} and g^{kp} follows in a similar way from the one in Figure 7.

The relevant observation is that on the new models it is possible to add a constraint to keep the number of intervals smaller than a given threshold q, i.e.

$$\sum_{i \in I} x_i \leq q.$$

5 Controlled Rounding Methodology

In *Controlled Rounding Methodology* we are also given with an input base number r_i for each cell i. Let us denote by $\lfloor a_i \rfloor$ the multiple of r_i obtained by rounding down a_i, and by $\lceil a_i \rceil$ the multiple of r_i obtained by rounding up a_i. When r_i is such that $\lfloor a_i \rfloor = \lceil a_i \rceil$, we redefine $r_i := 0$, hence

	A	B	C	Total
Activity I	20	50	10	80
Activity II	10	20	20	50
Activity III	15	30	15	60
Total	45	100	45	190

Figure 8: Controlled Rounding pattern with $r_i = 5$ for all $i \in I$.

$r_i = \lceil a_i \rceil - \lfloor a_i \rfloor$ for all $i \in I$.

A pattern in the Controlled Rounding Methodology is a congruent table $v = [v_i : i \in I]$ such that

$$v_i \in \{\lfloor a_i \rfloor, \lceil a_i \rceil\}. \tag{12}$$

Figure 8 gives an example of pattern when $r_i := 5$ $(i \in I)$ for the instance in Figure 1. The values r_i are assumed to be known by the attackers. The feasible region for the attacker problems associated to attacker k is defined by

$$\begin{aligned} My &= b \\ v_i - r_i \leq y_i &\leq v_i + r_i \qquad && \text{for all } i \in I \\ lb_i^k \leq y_i &\leq ub_i^k && \text{for all } i \in I. \end{aligned}$$

The natural concept of "loss of information" of a cell is defined as the difference between the nominal value and the published value, and then the loss of information of a pattern is the sum of all the individual loss of information.

A main difficulty of this methodology is that a feasible pattern does not always exists, even when all r_i are the same base numbers. A necessary condition for feasibility is that

$$\max\{SPL_p^k, LPL_p^k + UPL_p^k\} \leq 2r_p \qquad \text{for all } k \in K \text{ and } p \in P,$$

but it is not sufficient (even if all protection levels are zero). The combinatorial problem of finding (if any) a protected pattern with minimum information loss is called *Controlled Rounding Problem* (CRP). The problem was first introduced by Bacharach [1] in the context of replacing nonintegers by integers in tabular arrays, and actually it arises in several application contexts. To reduce the complexity of finding a feasible pattern, typically all base numbers are equal, as in the example. Nevertheless, with such hypothesis the existence of a feasible pattern is ensured on 2-dimensional tables, but not on general multi-dimensional tables with marginal totals; Causey, Cox and Ernst [3] showed a simple infeasible $2 \times 2 \times 2$ instance. Kelly, Golden and Assad [20] proposed a branch-and-bound procedure for the case of 3-dimensional tables, based on the LP relaxation of an ILP model. Heuristic methods for CRP on multi-dimensional tables have been proposed by several authors, including Kelly, Golden and Assad [20, 23]. Fischetti and Salazar [14] proposed a branch-and-bound procedure for its resolution based on Linear Programming, and some relaxation of CRP when it is infeasible. We present here a general model for linked and hierarchical tables against different attackers.

Let us consider a binary variable x_i for each cell i, representing

$$x_i = \begin{cases} 0 & \text{if } v_i = \lfloor a_i \rfloor, \\ 1 & \text{if } v_i = \lceil a_i \rceil. \end{cases}$$

Note that when a solution x_i is given, then the published table is determined by $v_i := \lfloor a_i \rfloor + r_i x_i$ for all $i \in I$.

The lost of information of a cell i can now be written as a constant when $x_i = 0$, plus a (positive or negative) difference w_i if $x_i = 1$. Therefore, the loss of information of the pattern defined by x is a constant plus $\sum_{i \in I} w_i x_i$.

To write a first model it is again convenient to consider variables $f^{kp} = [f_i^{kp} : i \in I]$ and $g^{kp} = [g_i^{kp} : i \in I]$ to ensure the existence of consistent tables certificating the fulfillment of the protection level requirements. Then a mathematical model is illustrated in Figure 9.

Once again, by using basic LP Duality Theory, it is possible to eliminate variables f^{kp} and g^{kp} from the first model. This mathematical operation leads to the second model in Figure 10.

6 Cell Perturbation Methodology

The main disadvantage of the Controlled Rounding methodology is that a protected pattern does not always exist due to the tight constraints (12). Therefore, a different way of ensuring the existence of protected patterns is to relax conditions (12) and to look for a congruent table $v = [v_i : i \in I]$ such that

$$v_i \in [\lfloor a_i \rfloor \ldots \lceil a_i \rceil]. \tag{13}$$

where $\lfloor a_i \rfloor$ and $\lceil a_i \rceil$ are given in advance from the statistical office such that $\lfloor a_i \rfloor \leq a_i \leq \lceil a_i \rceil$. This extreme values can be defined as the nearest numbers to a_i which are multiple of a given number (i.e., defined as in the standard Controlled Rounding methodology from a given base number), but they can be also the two values within a given difference respect to a_i (i.e., $\lfloor a_i \rfloor := a_i - t_i$ and $\lceil a_i \rceil := a_i + t_i$ for a given base number $t_i > 0$). Figure 11 shows a possible Cell Perturbation pattern for the nominal table in Figure 1. Table v is then a pattern in the *Cell Perturbation Methodology*. As in the Controlled Rounding methodology, the loss of information of a cell i is

$$\min \sum_{i \in I} w_i x_i$$

subject to:

$$\begin{array}{ll} \sum_{i\in I} m_{ij}(\lfloor a_i \rfloor + r_i x_i) = b_j & \text{for all } j \in J \\ x_i \in \{0,1\} & \text{for all } i \in I \end{array}$$

and, for all $p \in P$ and all $k \in K$:

$$\begin{array}{ll} \sum_{i\in I} m_{ij} f_i^{kp} = b_j & \text{for all } j \in J \\ lb_i^k \leq f_i^{kp} \leq ub_i^k & \text{for all } i \in I \\ \lfloor a_i \rfloor + r_i x_i - r_i \leq f_i^{kp} \leq \lfloor a_i \rfloor + r_i x_i + r_i & \text{for all } i \in I \\ \sum_{i\in I} m_{ij} g_i^{kp} = b_j & \text{for all } j \in J \\ lb_i^k \leq g_i^{kp} \leq ub_i^k & \text{for all } i \in I \\ \lfloor a_i \rfloor + r_i x_i - r_i \leq g_i^{kp} \leq \lfloor a_i \rfloor + r_i x_i + r_i & \text{for all } i \in I \\ f_p^{kp} \geq upl_p^k & \\ g_p^{kp} \leq lpl_p^k & \\ f_p^{kp} - g_p^{kp} \geq SPL_p^k. & \end{array}$$

Figure 9: First ILP model for Controlled Rounding.

defined proportional to $|v_i - a_i|$, and the "loss of information" of a pattern is the sum of the loss of information of all the cells.

Obviously, if all constraints (12) are removed and no one new is required, then the valid pattern with minimum loss of information is the nominal table. Hence, some constraints from (12) must remains (e.g., the one concerning the sensitive cells) or, in a much simple way, it could be required that the published values in each sensitive cell must be equal to some given values (e.g., $v_p = \lceil a_i \rceil$ for all $p \in P$).

Let $r_i := \lceil a_i \rceil - \lfloor a_i \rfloor$ a (possibly) known information for attackers. Then the attacker problems associated to attacker k are now exactly the same as in the Controlled Rounding Methodology, i.e.

$$\begin{array}{ll} My = b & \\ v_i - r_i \leq y_i \leq v_i + r_i & \text{for all } i \in I \\ lb_i^k \leq y_i \leq ub_i^k & \text{for all } i \in I. \end{array}$$

As in the Controlled Rounding methodology, a necessary (but not sufficient) condition for feasibility

$$\min \sum_{i \in I} w_i x_i$$

subject to:

$$\begin{array}{ll} \sum_{i \in I} m_{ij}(\lfloor a_i \rfloor + r_i x_i) = b_j & \text{for all } j \in J \\ x_i \in \{0,1\} & \text{for all } i \in I \end{array}$$

and, for all $p \in P$ and all $k \in K$:

$$\sum_{i \in I} \alpha_i^1 UB_i^k + \alpha_i^2(\lfloor a_i \rfloor + r_i x_i + r_i - a_i) + \beta_i^1 LB_i^k + \beta_i^2(a_i - \lfloor a_i \rfloor - r_i x_i + r_i) \geq UPL_p^k$$
$$\text{for all } \alpha^1, \alpha^2, \beta^1, \beta^2, \gamma \text{ satisfying (10)}$$

$$\sum_{i \in I} \alpha_i'^1 UB_i^k + \alpha_i'^2(\lfloor a_i \rfloor + r_i x_i + r_i - a_i) + \beta_i'^1 LB_i^k + \beta_i'^2(a_i - \lfloor a_i \rfloor - r_i x_i + r_i) \geq LPL_p^k$$
$$\text{for all } \alpha'^1, \alpha'^2, \beta'^1, \beta'^2, \gamma' \text{ satisfying (11)}$$

$$\sum_{i \in I} (\alpha_i^1 + \alpha_i'^1) UB_i^k + (\alpha_i^2 + \alpha_i'^2)(\lfloor a_i \rfloor + r_i x_i + r_i - a_i) +$$
$$(\beta_i^1 + \beta_i'^1) LB_i^k + (\beta_i^2 + \beta_i'^2)(a_i - \lfloor a_i \rfloor - r_i x_i + r_i) \geq SPL_p^k$$
$$\text{for all } \alpha^1, \alpha^2, \beta^1, \beta^2, \gamma \text{ satisfying (10) and}$$
$$\text{for all } \alpha'^1, \alpha'^2, \beta'^1, \beta'^2, \gamma' \text{ satisfying (11).}$$

Figure 10: Second ILP model for Controlled Rounding.

is that $\max_{k \in K}\{SPL_i^k, UPL_i^k + LPL_i^k\} \leq 2r_i$ for all $i \in I$.

There are in literature several methodologies to protect tables by data perturbation (see, e.g., Evans, Zayatz and Slanta [13], Duncan and Fienberg [12]) but, as far as we know, they all concern with modifying the microdata and, therefore, there is less control on the final protection interval of each cell in the published pattern.

To write an LP model for the Cell Perturbation model, it is convenient to introduce two continuous variables z_i^- and z_i^+ for each cell i, with the following meaning:

$$z_i^- := \max\{0, a_i - v_i\}$$

$$z_i^+ := \max\{0, v_i - a_i\}.$$

Let w_i^- be the given cost for each unit of z_i^-, and w_i^+ be the given cost for each unit of z_i^+. Hence

	A	B	C	Total
Activity I	20	50	10	80
Activity II	7	16	26	49
Activity III	18	35	8	61
Total	45	101	44	190

Figure 11: Cell Perturbation pattern.

the objective function is

$$\sum_{i \in I} w_i^+ z_i^+ + w_i^- z_i^-$$

as in the Controlled Rounding methodology.

For a first model it is again convenient to introduce additional variables f^{kp} and g^{kp} for each attacker k and each sensitive cell p. Then Figure 12 shows a first LP model.

Again, as it has been done with the previous methodologies, it is possible to remove the requirement of variables f^{kp} and g^{kp} by using basic Duality Theory on the first model, leading to the second model showed in Figure 13.

As in the example, an optimal valid pattern for Cell Perturbation could request too many modified values. To overcome this disadvantage, as proposed for the Partial Suppression Methodology, it could be possible to bound the maximum number of nominal values to be modified. To this end, an additional binary variable x_i is required for each cell $i \in I$. Variable x_i assumes value 1 when $v_i \neq a_i$ (i.e., $z_i^+ + z_i^- > 0$) and 0 otherwise. This extra variable can be inserted in the models and linked to the other variable with constraints:

$$\begin{array}{ll} 0 \leq z_i^+ \leq (\lceil a_i \rceil - a_i) x_i & \text{for all } i \in I \\ 0 \leq z_i^- \leq (a_i - \lfloor a_i \rfloor) x_i & \text{for all } i \in I. \end{array}$$

The objective function would then include a cost w_i for modifying a nominal value a_i, hence the mathematical models have the following new objective function:

$$\sum_{i \in I} w_i^+ z_i^+ + w_i^- z_i^- + w_i x_i.$$

Then, on the new models, it is again possible to add a constraint to keep the number of intervals smaller than a given threshold q, i.e.

$$\sum_{i \in I} x_i \leq q.$$

7 Conclusion

We have introduced several methods to protect sensitive data when publishing a table. The contribution of this work is a unified mathematical definition of the optimization problem under

$$\sum_{i \in I} w_i^+ z_i^+ + w_i^- z_i^-$$

subject to:

$$\begin{array}{cl}
\sum_{i \in I} m_{ij}(a_i + z_i^+ - z_i^-) = b_j & \text{for all } j \in J \\
z_i^+ = \lceil a_i \rceil - a_i & \text{for all } i \in P \\
z_i^- = 0 & \text{for all } i \in P \\
0 \leq z_i^+ \leq \lceil a_i \rceil - a_i & \text{for all } i \notin P \\
0 \leq z_i^- \leq a_i - \lfloor a_i \rfloor & \text{for all } i \notin P
\end{array}$$

and, for all $p \in P$ and all $k \in K$:

$$\begin{array}{cl}
\sum_{i \in I} m_{ij} f_i^{kp} = b_j & \text{for all } j \in J \\
lb_i^k \leq f_i^{kp} \leq ub_i^k & \text{for all } i \in I \\
a_i + z_i^+ - z_i^- - r_i \leq f_i^{kp} \leq a_i + z_i^+ - z_i^- + r_i & \text{for all } i \in I \\
\sum_{i \in I} m_{ij} g_i^{kp} = b_j & \text{for all } j \in J \\
lb_i^k \leq g_i^{kp} \leq ub_i^k & \text{for all } i \in I \\
a_i + z_i^+ - z_i^- - r_i \leq g_i^{kp} \leq a_i + z_i^+ - z_i^- + r_i & \text{for all } i \in I \\
f_p^{kp} \geq upl_p^k & \\
g_p^{kp} \leq lpl_p^k & \\
f_p^{kp} - g_p^{kp} \geq SPL_p^k. &
\end{array}$$

Figure 12: First ILP model for Cell Perturbation.

four different methodologies: Cell Suppression, Interval Publication, Controlled Rounding, and Cell Perturbation.

For each methodology, we have stablished the concept of protection pattern and information loss, and two mathematical models have been provided for each methodology. A first one contains a polynomial number of variables and constraints. Nevertheless, a serious disadvantage of the first model is that it uses two variables f_i^{kp} and g_i^{kp} for each attacker $k \in K$, sensitive cell $p \in P$ and cell $i \in I$. Basic Dual Theory in Linear Programming provides a procedure to avoid the use of these variables by adding some "capacity constraints", leading to a second mathematical model.

In all the four methodologies, the valid patterns provide feasible ranges of values for the sensitive cells such that there are congruent tables to guarantee all the protection level requirements. Then,

$$\sum_{i \in I} w_i^+ z_i^+ + w_i^- z_i^-$$

subject to:

$$\begin{array}{cl}
\sum_{i \in I} m_{ij}(a_i + z_i^+ - z_i^-) = b_j & \text{for all } j \in J \\
z_i^+ = \lceil a_i \rceil - a_i & \text{for all } i \in P \\
z_i^- = 0 & \text{for all } i \in P \\
0 \leq z_i^+ \leq \lceil a_i \rceil - a_i & \text{for all } i \notin P \\
0 \leq z_i^- \leq a_i - \lfloor a_i \rfloor & \text{for all } i \notin P
\end{array}$$

and, for all $p \in P$ and all $k \in K$:

$$\sum_{i \in I} \alpha_i^1 UB_i^k + \alpha_i^2 (z_i^+ - z_i^- + r_i) + \beta_i^1 LB_i^k + \beta_i^2 (-z_i^+ + z_i^- + r_i) \geq UPL_p^k$$
$$\text{for all } \alpha^1, \alpha^2, \beta^1, \beta^2, \gamma \text{ satisfying (10)}$$

$$\sum_{i \in I} \alpha_i'^1 UB_i^k + \alpha_i'^2 (z_i^+ - z_i^- + r_i) + \beta_i'^1 LB_i^k + \beta_i'^2 (-z_i^+ + z_i^- + r_i) \geq LPL_p^k$$
$$\text{for all } \alpha'^1, \alpha'^2, \beta'^1, \beta'^2, \gamma' \text{ satisfying (11)}$$

$$\begin{array}{l}
\sum_{i \in I} (\alpha_i^1 + \alpha_i'^1) UB_i^k + (\alpha_i^2 + \alpha_i'^2)(z_i^+ - z_i^- + r_i) + \\
\quad (\beta_i^1 + \beta_i'^1) LB_i^k + (\beta_i^2 + \beta_i'^2)(-z_i^+ + z_i^- + r_i) \geq SPL_p^k \\
\qquad \text{for all } \alpha^1, \alpha^2, \beta^1, \beta^2, \gamma \text{ satisfying (10) and} \\
\qquad \text{for all } \alpha'^1, \alpha'^2, \beta'^1, \beta'^2, \gamma' \text{ satisfying (11).}
\end{array}$$

Figure 13: Second ILP model for Cell Perturbation.

the Disclosure Auditing problem does not need to be solved to test validity of the output patterns. This feature represents a large saving of computational effort, specially when the protection is required against several attackers.

When convenient, it is also possible to apply a combination of the different methodologies. Indeed, for example, suppose there is a partition of the cell set I into $I^1 \cup I^2$, and the statistical office is interested in publishing intervals $[y_i^- \ldots y_i^+]$ when $i \in I^1$, using Interval Publication Methodology, and publishing perturbed values v_i when $i \in I^2$, using Cell Perturbation Methodology. Then a combined methodology can be mathematically modelled by observing that the feasible region of the attacker problems associated to attacker k is:

$$\begin{array}{rl} My = b & \\ v_i - r_i \leq y_i \leq v_i + r_i & \text{for all } i \in I^1 \\ y_i^- \leq y_i \leq y_i^+ & \text{for all } i \in I^2 \\ lb_i^k \leq y_i \leq ub_i^k & \text{for all } i \in I. \end{array}$$

Then, by combining the appropriated variables it is also possible the write two models as this paper have showed for each single methodology.

From the first models it is clear that the optimization problems of Interval Publication and Data Perturbation methodologies are both polynomially solvable as they do not require integer variables. The combinatorial problems associated to Cell Suppression and Controlled Rounding methodologies are $\mathcal{NP}$-hard problems, but still the ILP models here presented are suitable for been solved with branch-and-bound approaches (see, e.g., Fischetti and Salazar [15] for computational experiences on the Cell Suppression models).

The second models are more suitable to be solved by cutting-plane approaches. Indeed, the main idea is that not all the capacity constrains must be in the *master* problem since they are in an exponential number. Within an iterative procedure, only some of them are considered, and a missing important one can be computed from the dual variables after solving a linear program (the *subproblem*). Then, the subproblem feeds the master problem with capacity constraints, but it is also important to clean unnecessary constraint from the master problem, so to keep the master problem in a size manageable by an LP solver. See, e.g., Wolsey [29] for details on cutting-plane methods in Mathematical Programming.

An important remark when solving the second models of Cell Suppression and of Controlled Rounding arises by observing that each capacity constraint has the form:

$$\sum_{i \in I} d_i' x_i + d_i''(1 - x_i) \geq d_0$$

where d_i' and d_i'' are non-negative real numbers. Since $x_i \in \{1, 0\}$, the observation allows us to skip the constraint when d_0 is negative and, in other cases, to round down values d_i' and/or d_i'' to d_0 whenever they are bigger than d_0. This simple operation leads to strengthened the LP relaxation of the models, which is of fundamental use to produce lower bounds and speed up enumerative and heuristic approaches. Also others additional inequalities can be inserted to produce a further improvement of the LP relaxation, like the so-called *cover inequalities* (see, e.g., Wolsey [29]).

This paper has presented a unified mathematical framework for the four methodologies, so all differ in the structure of the output patterns, but share the same concept of protection. It is also possible to consider other common features using the presented mathematical models.

Indeed, considering a model with the x_i variables for each methodology, it is easy to observe that the statistical office could also control the number of suppressions, intervals, roundings and perturbations by just including constraint:

$$\sum_{i \in I} x_i \leq q,$$

where q is the desiderated upper bound. Moreover, it is possible to extend the methodologies to consider "conditional protection levels on nonsensitive cells". Indeed, if it is required that an interval $[y_i^- \ldots y_i^+]$ must satisfy $y_i^+ - y_i^- > SPL_i$ for a given non-zero number SPL_i when i is a nonsensitive cell and $y_i^+ \neq y_i^-$, then this conditional request can be inserted in the presented mathematical model by considering i as a sensitive cells with sliding protection level $SPL_i\, x_i$. The same consideration applies to upper and lower protection levels.

References

[1] Bacharach, M. (1966) "Matrix Rounding Problem", *Management Science*, **9**, 732–742.

[2] Carvalho, F. D., Dellaert, N. P. and Osório, M. S. (1994) "Statistical Disclosure in Two-Dimensional Tables: General Tables", *Journal of the American Statistical Association*, **89**, 1547–1557.

[3] Causey, B.D., Cox, L.H. and Ernst, L.R. (1985) "Applications of Transportation Theory to Statistical Problems", *Journal of the American Statistical Association*, **80**, 903–909.

[4] Cox, L. H. (1980) "Suppression Methodology and Statistical Disclosure Control", *Journal of the American Statistical Association*, **75**, 377–385.

[5] Cox, L.H. (1981) "Linear sensitivity measures and statistical disclosure control", *Journal of Statitical Planning and Inference*, **5**, 153–164.

[6] Cox, L.H. (1982) "Controlled Rounding", *INFOR*, **20**, 423–432.

[7] Cox, L.H. (1987) "A Constructive Procedure for Unbiased Controlled Rounding", *Journal of the American Statistical Association*, **82**, 520–524.

[8] Cox, L. H. (1995) "Network Models for Complementary Cell Suppression", *Journal of the American Statistical Association*, **90**, 1453–1462.

[9] Dellaert, N. P. and Luijten, W. A. (1996) "Statistical Disclosure in General Three-Dimensional Tables", Technical Paper TI 96-114/9, Tinbergen Institute.

[10] Domingo-Ferrer, J. (editor) (2002) *Inference Control in Statistical Databases: From Theory to Practice*, Lecture Notes in Computer Science 2316, Springer.

[11] Doyle, P., Lane, J., Theeuwes, J. and Zayatz, L. (editors) (2001) *Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies*, Elsevier Science.

[12] Duncan, G.T. and Fiendberg, S.E. (1998) "Obtaining information while preserving privacy: a Markov perturbation method for tabular data", Proceedings of the *Statistical Data Protection* conference, 351–362.

[13] Evans, T., Zayatz, L. and Slanta, J. (1998) "Using Noise for Disclosure Limitation of Establishment Tabular Data", *Journal of Official Statistics*, **14/4**, 537–551.

[14] Fischetti, M. and Salazar, J. J. (1998) "Computational Experience with the Controlled Rounding Problem in Statistical Disclosure Control", *Journal of Official Statistics*, **14/4**, 553–565.

[15] Fischetti, M. and Salazar, J. J. (2000) "Models and Algorithms for Optimizing Cell Suppression Problem in Tabular Data with Linear Contraints", *Journal of the American Statistical Association*, **95**, 916–928.

[16] Fischetti, M. and Salazar, J. J. (2002) "Partial Cell Suppression: a New Methodology for Statistical Disclosure Control", to appear in *Statistics and Computing.*

[17] Geurts, J. (1992) "Heuristics for Cell Suppression in Tables", Technical Paper, Netherlands Central Bureau of Statistics, Voorburg.

[18] Jewett, R. (1993) "Disclosure Analysis for the 1992 Economic Census", Working paper, U.S.B.C.

[19] Kelly, J. P. (1990) "Confidentiality Protection in Two and Three-Dimensional Tables", Ph.D. dissertation, University of Maryland, College Park, Maryland.

[20] Kelly, J.P., Golden, B.L. and Assad, A.A. (1990) "Using Simulated Annealing to Solve Controlled Rounding Problems", *ORSA Journal on Computing*, **2**, 174–185.

[21] Kelly, J.P., Golden, B.L., Assad, A.A. and Baker, E.K. (1990) "Controlled Rounding of Tabular Data", *Operations Research*, **38**, 760–772.

[22] Kelly, J. P., Golden, B. L. and Assad, A. A. (1992) "Cell Suppression: Disclosure Protection for Sensitive Tabular Data", *Networks*, **22**, 397–417.

[23] Kelly, J.P., Golden, B.L. and Assad, A.A. (1993) "Large-Scale Controlled Rounding Using TABU Search with Strategic Oscillation", *Annals of Operations Research*, **41**, 69–84.

[24] Robertson, D. A. (1994) "Cell Suppression at Statistics Canada", Proceedings of the *Second International Conference on Statistical Confidentiality*, Luxembourg.

[25] Robertson, D. A. (2000) "Improving Statistics Canada's Cell Suppression software (CONFID)", Technical paper, Statistics Canada, Ottawa, Canada.

[26] Sande, G. (1984) "Automated Cell Suppression to preserve confidentiality of business statistics", *Statistical Journal of the United Nations ECE*, **2**, 33–41.

[27] Sande, G. (1995) "ACS documentation", *Sande & Associates*, 600 Sanderling Ct. Secaucus NJ, 07094 U.S.A.

[28] Willenborg, L. C. R. J. and de Waal, T. (2001) *Elements of Statistical Disclosure Control*. Lecture Notes in Statistics 155, Springer.

[29] Wolsey, L.A. (1998) *Integer Programming*, Wiley-Interscience.