Perspective Reformulations of the CTA Problem With L_2 Distances

Jordi Castro	Antonio Frangioni	Claudio Gentile
Dept. of Stat. and O.R.	Dip. di Informatica	IASI
Univ. Politècnica de Catalunya	Univ. di Pisa	C.N.R.
jordi.castro@upc.edu	frangio@di.unipi.it	gentile@iasi.cnr.it

Research Report UPC-DEIO DR 2014-02

previous version also availabe as

Technical Report IASI-CNR 11-19

Perspective Reformulations of the CTA Problem With L_2 Distances*

Jordi Castro
Department of Statistics and Operations Research
Universitat Politècnica de Catalunya
Jordi Girona 1-3, 08034 Barcelona
jordi.castro@upc.edu

Antonio Frangioni
Dipartimento di Informatica
Università di Pisa
Largo B. Pontecorvo 3, 56127 Pisa, Italy
frangio@di.unipi.it

Claudio Gentile
Istituto di Analisi dei Sistemi ed Informatica, C.N.R.
Viale Manzoni 30, 00185 Rome, Italy
gentile@iasi.cnr.it

Abstract

Any institution that disseminates data in aggregated form has the duty to ensure that individual confidential information is not disclosed, either by not releasing data or by perturbing the released data, while maintaining data utility. Controlled tabular adjustment (CTA) is a promising technique of the second type where a protected table that is close to the original one in some chosen distance is constructed. The choice of the specific distance shows a trade-off: while the Euclidean distance has been shown (and is confirmed here) to produce tables with greater "utility", it gives rise to Mixed Integer Quadratic Problems (MIQPs) with pairs of linked semi-continuous variables that are more difficult to solve than the Mixed Integer Linear Problems corresponding to linear norms. We provide a novel analysis of Perspective Reformulations (PRs) for this special structure; in particular, we devise a *Projected* PR (P^2R) which is piecewise-conic but simplifies to a (nonseparable) MIQP when the instance is symmetric. We then compare different formulations of the CTA problem, showing that the ones based on P^2R most often obtain better computational results.

keywords Mixed Integer Quadratic Programming, Perspective Reformulation, Data Privacy, Statistical Disclosure Control, Tabular Data Protection, Controlled Tabular Adjustment

*To appear in *Operations Research*

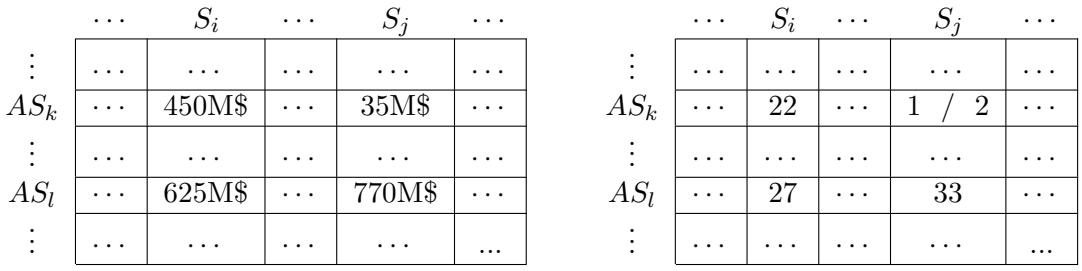

Figure 1: Example of disclosure in tabular data: (a) turnover and (b) number of companies per activity sector and state.

1 Introduction

The most important mission of National Statistical Agencies (NSAs), and a significant mission of several other institutions, is to provide high-quality statistical data. These data are disseminated either in disaggregated (i.e., microdata or microfiles) or aggregated (i.e., tabular data) form. A microdata file is a matrix of individuals by variables, where each cell provides the information of a particular individual for some particular variable. Crossing two or more categorical variables of the microdata file produces tabular data, either a single multiway or multidimensional table, or a set of related tables. There are stringent requirements that no confidential or sensitive information of any individual can be disclosed from the released data; not only is this dictated by law, but also respondents (e.g., of a census) may be tempted to hide or change information if they suspect that their confidential information may be released. This justifies the interest in *statistical disclosure control*, i.e., the set of techniques that can be deployed to protect sensitive information. In particular, the focus of this work is on *tabular data protection*. A seminal work on this field is in Bacharach (1966); the current state-of-the-art is described in the recent surveys of Salazar-González (2008) and Castro (2012a), as well as in the monographs Willenborg and de Waal (2000), Hundepool et al. (2012).

Although tabular data provide aggregated information, the publication of some cells may jeopardize individual information. Consider the small example of Figure 1: if there is only one company with activity sector AS_k in state S_j, then any attacker knows the turnover of this company. For two companies, any of them can deduce the other's turnover, becoming an internal attacker. Clearly, the risk in the example is due to a small number of respondents in cell (AS_k,S_j). However, even if the number of respondents was larger, there could be a disclosure risk if some companies can obtain a tight estimator of another's turnover (for instance by subtracting its own contribution from the cell value). Unsafe or sensitive cells are a priori determined before the application of any tabular data protection method, by applying some "sensitivity rules". These rules are out of the scope of this work; e.g., see Domingo-Ferrer and Torra (2002), Hundepool et al. (2012) for details.

Disclosure limitation techniques for tabular data can be seen as a map $F(T) = T'$,

where T and T' are the original an released tables, respectively. The goal is to obtain a table T' which minimizes the *disclosure risk* while preserving as much as possible the *data utility* of the original table T. According to Hundepool et al. (2012), data utility can be defined as the value of a given data release as an analytical resource; for microdata this means that statistical analyses (e.g., regressions, principal component analysis, etc.) provide similar results with the original and released microfiles, while for tables data utility can be measured (Castro and Giessing 2006) by the number of published cells with large relative changes with respect to their original values. Disclosure risk (Hundepool et al. 2012) occurs when an unacceptably narrow estimate of some confidential respondent's information can be derived from the released table T', that is, when the *attacker problem* $\hat{T} = \hat{F}^{-1}(T')$—where $\hat{F}^{-1}$ is an estimate of the inverse map (Castro 2012b)—produces "close" estimates $\hat{T}$ of T. In practice, disclosure control decisions are a trade-off between data utility and disclosure risk.

Tabular data techniques are classified as *perturbative* if one is allowed to add small perturbations or adjustments to released data, and as *nonperturbative* if released cell values must be exact, and therefore one is only allowed to entirely eliminate cells. Clearly, nonperturbative approaches are more rigid than perturbative ones. Furthermore, the most widely used nonperturbative approach, *cell suppression* (Kelly et al. 1992, Fischetti and Salazar-González 2001, Castro 2007), requires the solution of large-scale optimization problems to identify the optimal set of cells to be suppressed. It is perhaps not surprising, therefore, that perturbative approaches are being considered as emerging technologies for tabular data protection. In particular, *Controlled Tabular Adjustment* (CTA) is gaining recognition and acceptance among NSAs (Zayatz 2009), as testified by the recent monograph Hundepool et al. (2012) and by the fact that it is currently used by Eurostat (Statistical Office of the European Communities) within a wider protection scheme for tabular data (Giessing et al. 2009). Figure 1 can be used to illustrate CTA. If cell (AS_k,S_j) of table (a) is considered sensitive, with *lower* and *upper protection levels* of 5, then the published value of this cell must be in the range $(-\infty, 30] \cup [40, \infty)$. We say that the *protection sense* is "lower" or "upper" if the published value is, respectively, in $(-\infty, 30]$ or in $[40, \infty)$. The remaining cells in the same column and row of the sensitive cell have to be accordingly adjusted to preserve the marginal values, while minimizing the distance between the original and the released values. Since each sensitive cell introduces a disjunctive constraint, which can be formulated by adding one binary variable, when the number of sensitive cells is large CTA is a difficult combinatorial optimization problem.

It is worth remarking that, while the tables of Figure 1 are two-way (two-dimensional) ones, in general the situation can be much more complex. Tables can be classified in (i) k-dimensional tables, which are obtained by crossing k categorical variables; (ii) hierarchical tables, or set of tables that share some variables with hierarchical structure (e.g., "country", "state/province", "city"); (iii) linked tables, the most general situation, which is a set of tables that are obtained from the same microdata. A particularly interesting case for NSAs, which will be tested in this work, are two-dimensional hierarchical tables that share one hierarchical variable (e.g., tables that show the turnover crossing "activity sector" by "country", "activity sector" by "state/province", and "activity sector" by "city"). These are named *one-hierarchical two-dimensional* tables (or 1H2D for short),

and their relations can be represented as a tree of tables. However, table relations for any type of table are represented by linear constraints, where the sum of the inner cells is equal to the marginal cell; thus, the techniques developed in this paper are applicable to the most general case (linked tables) as well.

In all previous works on CTA, the L_1 or Manhattan norm has been used to measure the distance between the original and the protected published data (Dandekar and Cox 2002, Castro 2006). This has the advantage that CTA can then be formulated as a Mixed Integer Linear Problem (MILP) with a number of variables and constraints that is linear in the size of the table, and whose solution can therefore be attempted with general-purpose MILP solvers. By contrast, formulations of the cell suppression problem are much larger and typically require the application of specialized approaches such as Benders decomposition. This is not to say that CTA, even with the L_1 distance, is an easy problem: for large (1H2D) tables MILP solvers may require a long time even to provide a first feasible solution, and therefore heuristic approaches (González and Castro 2011) are required to provide practical solutions in a reasonable time. It can be expected that CTA with L_2 (Euclidean) distance, which results in a Mixed Integer Quadratic Problem (MIQP), is even more difficult to solve; this is likely the reason why this work is, to the best of our knowledge, the first one where such a feat is attempted. Yet, protecting a table using L_2 in CTA has several benefits:

- Weighting the distance between the original and the published cell value by the inverse of the original cell value, the objective function of CTA minimizes the well-known χ^2 distance between the original and the released table, which is useful for the statistical evaluation of the results.

- The L_2 distance more evenly distributes the deviations induced by sensitive cells to other cells. This avoids concentration of deviations in few cells, which improves the overall utility of the published data, as measured, e.g., by the number of cells (possibly counting the non-sensitive ones only) whose published value is "significantly" different from the original data. This is empirically shown in Subsection 4.3 for all the test instances considered in this work.

- It has been recently shown (Castro 2012b) that, in general, L_2 provides solutions with a lower disclosure risk than L_1, i.e., it is more difficult for an attacker to infer good information about the original data if the table was protected with L_2-CTA rather than with L_1-CTA. In particular, the attacker problem $\hat{T} = \hat{F}^{-1}(T')$ was solved in (Castro 2012b) for 25 instances from the literature, considering both L_1-CTA and L_2-CTA in four attacker scenarios, each one corresponding to a different amount of information available to the attacker to compute the estimate of the inverse map $\hat{F}^{-1}$, and taking randomness into account. The empirical distributions of the percentage differences between the original confidential cell values T and the estimates $\hat{T}$ computed by the attacker showed that L_2-CTA provides in general a measurably lower disclosure risk than L_1-CTA.

- From a computational point of view, once the binary variables are fixed (i.e., the protection sense is decided), the solution of the resulting continuous problem can be

more efficient for L_2 than for L_1 if Interior-Point (IP) methods are used, as it was shown in (Castro 2006); while this holds true already for general-purpose solvers, specialized IP approaches can be orders of magnitude faster than state-of-the-art general-purpose ones (Castro and Cuesta 2010). Indeed, from the IP perspective, the resulting QPs and LPs are quite equivalent, both in theory and in practice. Theoretically, the complexity of the IP algorithm is the same, since it only depends on the *self-concordant barrier function* used for the inequality constraints (the variable bounds), which are the same in these QP and LP (Nesterov 2004, Chapter 4). And in practice, since the QP is separable (the Hessian is diagonal), the structure of the linear systems to be solved at each IP iteration is the same for QP and LP. On the other hand, the QPs from L_2-CTA have the additional benefit of being strictly convex, which may actually *improve* IP performances. For instance, in Castro and Cuesta (2010), very large QPs (of about 10 million variables and 200,000 constraints) deriving from L_2-CTA were solved by the IP algorithm of `Cplex` in seven iterations; such good results have never been observed for the LPs deriving from L_1-CTA. We remark that simplex-like approaches are not competitive with IP methods for the LPs derived from L_1-CTA, likely due to the degeneracy of the problem; however, IP methods are not efficient when the original L_1-CTA MILP model is solved, since they lack the reoptimization capabilities of simplex methods.

On the other hand, the protected values provided by CTA with the L_2 distance will likely be more fractional than those provided by the L_1 distance, which has been often observed in practice to provide integer values even without imposing integrality constraints. Yet, this is not a significant drawback since CTA is mainly used for "magnitude" tables which do not provide frequencies but information about a third continuous variable (salary, net profit, turnover, ...) which is most often fractional.

The comparison of the computational results of Section 4 for L_2-CTA with those reported in the literature for L_1-CTA (Castro 2012a) confirms that the former is much harder than the latter. On the other hand, L_2-CTA has been empirically shown to provide tables with higher data utility (Subsection 4.3) and lower disclosure risk (Castro 2012b). Hence, there is a trade-off between computing time and data utility/disclosure risk. Admittedly, L_2-CTA may not be a practical approach for large instances: even if providing better quality tables, the solution time can be unbearable. However, for small-to-medium tables L_2-CTA can be a choice: NSAs may well be willing to spend extra CPU time to obtain tables with higher utility and lower disclosure risk. Unfortunately, the straightforward MIQP formulation of L_2-CTA is computationally prohibitive even for small tables; this paper is the first attempt to derive some practical solution approaches for this problem.

The main structural characteristic of MIQP formulations of CTA with the L_2 distance (from now on, simply "CTA") is very closely related to *convex separable quadratic-cost models with semicontinuous variables*, which are naturally formulated as in the following (fragment of) MIQP

$$\min \left\{ wz^2 + cy \ : \ yl \leq z \leq yu \ , \ y \in \{0,1\} \right\} \tag{1}$$

where $w > 0$ and $l < u$. This is useful because (1) admits the *Perspective Reformulation*

(PR)

$$\min \left\{ wz^2/y + cy \ : \ yl \leq z \leq yu \ , \ y \in \{0,1\} \right\} \ . \tag{2}$$

Despite the weird look and the apparent ill-definiteness at $y = 0$, the objective function in (2) can be extended with continuity in zero and it is convex. Actually, the objective function is the *convex envelope* of an appropriately re-defined version of the objective function in (1), i.e., the best possible convex objective function to have when the integrality constraints $y \in \{0,1\}$ are relaxed to $y \in [0,1]$. Indeed, (2) has at least two possible further reformulations which avoid the fractional term in the objective function with the associated difficulties (nondifferentiability, possible numerical problems) at $y = 0$: one is the Mixed Integer Second-Order Cone Program (SOCP)

$$\min \left\{ v + cy \ : \ yl \leq z \leq yu \ , \ \sqrt{wz^2 + (v-y)^2/4} \ \leq \ (v+y)/2 \ , \ y \in \{0,1\} \right\} \tag{3}$$

(Tawarmalani and Sahinidis 2002, Aktürk et al. 2009, Günlük and Linderoth 2008), and another is the Semi-Infinite (SI) MILP

$$\min \left\{ v + cy \ : \ yl \leq z \leq yu \ , \ y \in \{0,1\}, \quad v \geq w(2\gamma z - \gamma^2 y) \quad \text{for all } \ \gamma \in [l,u] \ \right\} \tag{4}$$

where γ is the index of the infinitely many linear constraints (called *Perspective Cuts* in Frangioni and Gentile (2006)) whose pointwise supremum completely describes the objective function in (2). Either (3) or any finite approximation to (4)—typically, to be iteratively refined—can be used as models of (2), whose continuous relaxation is significantly stronger than the one of (1) and that therefore is a more convenient starting point to develop exact and approximate solution algorithms (Frangioni and Gentile 2006, 2007, Günlük and Linderoth 2008, Aktürk et al. 2009, Frangioni et al. 2009). Somewhat surprisingly, the potentially very large and approximated (4) appears to be most often preferable to the compact and exact (3) in the context of exact or approximate enumerative solution approaches (Frangioni and Gentile 2009), likely due to the better reoptimization capabilities of simplex methods for linear programs compared to those of interior point methods for conic programs.

Yet a different approach, named *Projected Perspective Reformulation* (P^2R), has been recently proposed in Frangioni et al. (2011). The idea is to recast the continuous relaxation of (2) as the minimization over $z \in [0,u]$ of the function

$$\phi(z) = \min_y \left\{ wz^2/y + cy \ : \ ly \leq z \leq uy \ , \ y \in [0,1] \right\} \tag{5}$$

which effectively eliminates the y variable(s) from the model and projects the perspective relaxation on the space of the variable z. The function ϕ is convex, and its closed form can be algebraically computed revealing a piecewise-quadratic function with at most two pieces, at most one of them actually quadratic (and the other linear). When the underlying problem has a useful structure (e.g., network flow or knapsack), the continuous relaxation of (2) obtained in this way retains that structure, which allows to use specialized algorithms to solve it and therefore to outperform both (3) and (4). The application of this approach is only possible under rather restrictive assumptions that are only partially satisfied in

our case; this has been very recently improved upon in Frangioni et al. (2013), where in particular the y variables are re-instated in the formulation thus allowing the use of general-purpose solvers, albeit possibly at the cost of weakening the bound w.r.t. the one of the standard Perspective Reformulation.

In this paper we discuss the application of Perspective Reformulation techniques to the CTA problem. In particular, besides the standard approaches (3) and (4), we develop and test a new reformulation partly inspired by the results of Frangioni et al. (2011). However, since our problem is different and somewhat more complex, the projected version of the PR we obtain is substantially different and trickier to use. Thus, instead of insisting in keeping the equivalence with the original formulation, we adopt an approach conceptually similar to (but technically entirely different from) that of Frangioni et al. (2013) and "drop the nastier pieces" of the projected formulation, ending up with an *approximated* reformulation which is only as tight as the PR in some special cases, and looser otherwise. However, this reformulation results in a simpler (although non-separable) MIQP to be solved, and therefore it is most often preferable to the standard ones (3) and (4); furthermore, it suggests a simple modification to the latter which invariably improves their performances. Armed with these results we show on a large experimental set that CTA for randomly-generated 1H2D and real-world tables of realistic sizes can most often be solved effectively enough.

We remark that the Perspective Reformulation approach is much more widely applicable than the simple quadratic case we consider here; furthermore, it not only applies to the objective function but also to constraints $f(z) \leq 0$ that are "activated" if and only if a binary variable y is 1, where f can be any closed convex (possibly, SOCP-representable) function and z is a vector whose feasible region can be any bounded polyhedron; see e.g. Ceria and Soares (1999), Tawarmalani and Sahinidis (2002), Grossmann and Lee (2003), Frangioni and Gentile (2006), Hijazi et al. (2011) and the recent survey Günlük and Linderoth (2011). It should also be remarked that the computation of convex envelopes for specially-structured functions of "a few" variables is an important field for which several advances are being done; one of the most researched structures is that of functions $\phi(z, y) = f(z)g(y)$ where f is convex and g is concave (Khajavirad and Sahinidis 2013, Tawarmalani et al. 2012, Tawarmalani and Sahinidis 2001). Thus, it is conceivable that some of the ideas developed here, while technically different from these exploited in different settings, could be extended to a possibly large set of more complex situations.

2 Formulations of the CTA problem

Any CTA problem instance, either with one table or with any number of tables, can be represented by the following elements:

- a set of n cells a_i, $i \in \mathcal{N} = \{1, \ldots, n\}$, that satisfy m linear relations $Aa = b$ ($a = [a_i]_{i \in \mathcal{N}}$); these relations impose that the set of *inner* cells has to be equal to the *total* or *marginal* cell, i.e., if $\mathcal{I}_|$ is the set of inner cells of relation $j \in \{1, \ldots, m\}$, and t_j is the index of the total cell of relation j, the constraint associated to this

relation is $\left(\sum_{i\in\mathcal{I}_|} a_i\right) - a_{t_j} = 0$;

- the subset $\mathcal{S} \subseteq \mathcal{N}$ of indices of sensitive cells, and hence its complement $\mathcal{U} = \mathcal{N} \setminus \mathcal{S}$;
- a vector of nonnegative cell weights $w = [w_i]_{i\in\mathcal{N}}$;
- *finite* lower and upper bounds $\bar{l}^a \le a \le \bar{u}^a$ for each cell reasonably known by any attacker;
- nonnegative *lower and upper protection levels* for each confidential cell $i \in \mathcal{S}$, l_i and u_i respectively, such that the released values $x = [x_i]_{i\in\mathcal{N}}$ are considered to be safe if they satisfy

$$\text{either } x_i \ge a_i + u_i \text{ or } x_i \le a_i - l_i \qquad \text{for all } i \in \mathcal{S} \ . \tag{6}$$

Given any weighted distance $\|\cdot\|_w$, CTA can then be formulated as

$$\min \left\{ \|x-a\|_w \ : \ Ax = b \ , \ \bar{l}^a \le x \le \bar{u}^a \ , \ (6) \right\} \tag{7}$$

since one seeks for the released values x that are closest (in the given norm) to the true values a, compatible with the relationships that a is known to have to satisfy, and protected according to (6). Of course, the *disjunctive constraints* (6) are the difficult part of the problem, their feasible region being nonconvex. Formulating them hence requires some nonconvex element, the simplest one being a vector of binary variables $y = [y_i]_{i\in\mathcal{S}} \in \{0,1\}^{|\mathcal{S}|}$. It is also convenient to restate the problem in terms of the *deviations* $z = x - a$ from the true cell values, which therefore have to satisfy $\bar{l}^a - a = \bar{l} \le z \le \bar{u} = \bar{u}^a - a$; this gives the formulation

$$\min \left\{ \|z\|_w \ : \ Az = 0 \ , \ \bar{l} \le z \le \bar{u} \ , \ \bar{l}_i(1-y_i)+u_i y_i \le z_i \le \bar{u}_i y_i - l_i(1-y_i) \ , \ y_i \in \{0,1\} \quad i \in \mathcal{S} \right\} \tag{8}$$

with "natural big-M constraints". Indeed, when $y_i = 1$ one has $z_i \ge u_i$ and thus the *protection sense* is "upper", while when $y_i = 0$ one rather gets $z_i \le -l_i$ and thus the protection sense is "lower". While this formulation is correct, it would provide rather weak bounds when its *continuous relaxation* is formed by replacing the integrality constraints $y_i \in \{0,1\}$ with $y_i \in [0,1]$. The simple example with $n = 1$, "empty" A, $l_1 = u_1 = 10$ and $-\bar{l}_1 = \bar{u}_1 = 100$ shows that for $y_1 = 1/2$ the solution $z_1 = 0$ is feasible to the relaxation, whose optimal value is therefore 0, while the optimal value of the integer problem is $\|10\|_w$. Since weak bounds are very detrimental for the solution of the problem via exact or approximate approaches, we aim at constructing "better" formulations of the problem.

A first step in this direction is to introduce vectors of *positive and negative deviations* $z^+ \in \mathbb{R}^n$ and $z^- \in \mathbb{R}^n$, respectively, thereby redefining $z = z^+ - z^-$; this allows to reformulate the disjunctive constraints in (8) as

$$\begin{array}{rcll} u_i y_i \le & z_i^+ & \le \bar{u}_i y_i & \\ l_i(1-y_i) \le & z_i^- & \le -\bar{l}_i(1-y_i) & \qquad i \in \mathcal{S} \\ y_i & \in & \{0,1\} & \end{array} \tag{9}$$

As before, when $y_i = 1$, the constraints force $u_i \leq z_i^+ \leq \bar{u}_i$ and $z_i^- = 0$, thus the protection sense is "upper"; conversely, when $y_i = 0$ we get $z_i^+ = 0$ and $l_i \leq z_i^- \leq -\bar{l}_i$, thus the protection sense is "lower". This alone is not enough to improve on the bounds, though: in the above example we now have $z_1^+ = z_1^- = 5$ as a feasible solution for $y_1 = 1/2$, which still leads to a null bound. However the advantage of this formulation is that we now have *two semicontinuous variables*, to which we can hope to apply Perspective Reformulation techniques. This is not straightforward: the two semicontinuous variables are governed by the *same integer variable*, and unlike in standard cases—where this is possible, provided that all variables are "active" or "inactive" at the same time—one of them is "active" if and only of the other is not. Furthermore, the objective function is *nonseparable* in z^+ and z^-, and the convex envelope of multilinear functions, even if with only two variables as here, is notoriously a complex object (cf. Luedtke et al. (2010) and the references therein) so that "dirty tricks" have to be used (Frangioni and Gentile 2007) in order to apply PR techniques. Thus, the next section will be devoted to the study of the convex envelope for our particular case.

3 Perspective Reformulations of the CTA problem

In the following we will most often concentrate on a single cell $i \in \mathcal{S}$; thus, to simplify the notation we will consider the index i as fixed and drop it. In order to improve the lower bound provided by the continuous relaxation, one possibility is to compute the convex envelope of the nonconvex function

$$f(\ z^+\ ,\ z^-\ ,\ y\ ) = \begin{cases} w(z^+ - z^-)^2 & \text{if } u \leq z^+ \leq \bar{u}\ \ ,\ z^- = 0 \ \text{ and } y = 1 \\ w(z^+ - z^-)^2 & \text{if } l \leq z^- \leq -\bar{l}\ ,\ z^+ = 0 \ \text{ and } y = 0 \\ +\infty & \text{otherwise} \end{cases} \tag{10}$$

This can be easily accomplished with well-developed tools from the literature, starting from the classical results of Tawarmalani and Sahinidis (2002) that have been along the way extended to large classes of functions, e.g. these whose *generating sets* have appropriate structure (Khajavirad and Sahinidis 2013). In our case, however, this can be very simply obtained by considering two arbitrary points $u \leq \bar{z}^+ \leq \bar{u}$ and $l \leq \bar{z}^- \leq -\bar{l}$ and computing the convex combinations of the two tuples in the epigraphical space

$$(\ \bar{z}^+,\ 0,\ 1,\ w(\bar{z}^+)^2\ ) \qquad\qquad (\ 0,\ \bar{z}^-,\ 0,\ w(\bar{z}^-)^2\ )\ .$$

In other words, taking any arbitrary convex combinator $\theta \in [0, 1]$ and using the shorthand $f(z) = wz^2$ (which also suggests how the approach can be generalized to any convex function f), we have

$$\begin{aligned} &\theta(\ \bar{z}^+\ ,\ 0\ ,\ 1\ ,\ f(\bar{z}^+)\ ) + (1-\theta)(\ 0\ ,\ \bar{z}^-\ ,\ 0\ ,\ f(\bar{z}^-)\ ) \quad = \\ &(\ \theta\bar{z}^+\ ,\ (1-\theta)\bar{z}^-\ ,\ \theta\ ,\ \theta f(\bar{z}^+) + (1-\theta)f(\bar{z}^-)\ ) \end{aligned}$$

Now, identifying $\theta \equiv y$, $z^+ \equiv \theta\bar{z}^+$ and $z^- \equiv (1-\theta)\bar{z}^-$ we can rewrite the above as

$$\left(\ z^+\ ,\ z^-\ ,\ y\ ,\ yf\left(\frac{z^+}{y}\right) + (1-y)f\left(\frac{z^-}{1-y}\right)\ \right)$$

which finally leads to

$$\overline{co}f(\ z^+\ ,\ z^-\ ,\ y\ ) = \begin{cases} w\left(\frac{(z^+)^2}{y} + \frac{(z^-)^2}{1-y}\right) & \text{if} \quad \begin{array}{rcl} uy \leq & z^+ & \leq \bar{u}y \\ l(1-y) \leq & z^- & \leq -\bar{l}(1-y) \end{array} \quad ,\ y \in (0,1) \\ w\frac{(z^+)^2}{y} & \text{if } uy \leq z^+ \leq \bar{u}y\ ,\ z^- = 0\ ,\ y = 1 \\ w\frac{(z^-)^2}{1-y} & \text{if } z^+ = 0\ ,\ l(1-y) \leq z^- \leq -\bar{l}(1-y)\ ,\ y = 0 \\ +\infty & \text{otherwise} \end{cases} \tag{11}$$

and therefore to the following PR of (8):

$$\min \qquad \textstyle\sum_{i\in\mathcal{U}}\ w_i(z_i^+ - z_i^-)^2 + \sum_{i\in\mathcal{S}} \overline{co}f_i(\ z_i^+\ ,\ z_i^-\ ,\ y_i\ ) \tag{12}$$

$$A(z^+ - z^-) = 0 \quad , \quad 0 \leq z^+ \leq \bar{u} \quad , \quad 0 \leq z^- \leq -\bar{l} \quad , \quad (9) \tag{13}$$

In the following, with a little abuse of notation we will often write the simpler

$$w\big(\ (z^+)^2/y + (z^-)^2/(1-y)\ \big) \tag{14}$$

instead of $\overline{co}f(\ z^+\ ,\ z^-\ ,\ y\ )$. This is justified by the fact that even if (14), like (2), looks undefined for $y = 0$ (and $y = 1$), it is easily extended by continuity; this can be seen either by considering that when $y \to 0$ then $z^+ \to 0$ linearly with y, so that $(z^+)^2/y \to 0$ (and symmetrically when $y \to 1$), or by devising reformulations a-la (3) and (4) and verifying that they actually have no problems at all for integer values of y. We also remark that the PR (11) can alternatively be obtained as follows:

1. substitute $(z^+ - z^-)^2$ in the objective function with $(z^+)^2 + (z^-)^2$, which is correct since $z^+z^- = 0$ holds in each *integer* solution;
2. treat z^+ and z^- as two distinct semicontinuous variables with two distinct binary variables, say y^+ and y^-, and apply the standard PR (2);
3. now exploit the fact that $y^+ + y^- = 1$ to replace $y^+ = y$ and $y^- = 1 - y$.

This analysis suggests that one can further improve the PR even regarding the non-sensitive cells $i \in \mathcal{U}$. In fact, these can be considered as sensitive cells with $l = u = 0$, and therefore it is clear that one could have taken

$$\text{(MIQP)} \qquad \min\ \left\{\ \textstyle\sum_{i\in\mathcal{N}}\ w_i\big(\ (z_i^+)^2 + (z_i^-)^2\ \big)\ :\ (13)\ \right\}$$

as the *original* MIQP formulation of CTA, to which then directly apply steps 2. and 3. above, thus obtaining

$$\text{(PR)} \qquad \min\ \left\{\ \textstyle\sum_{i\in\mathcal{U}}\ w_i\big(\ (z_i^+)^2 + (z_i^-)^2\ \big) + \sum_{i\in\mathcal{S}}\ w_i\big(\ (z_i^+)^2/y_i + (z_i^-)^2/(1-y_i)\ \big)\ :\ (13)\ \right\}\ .$$

Note how (MIQP) has already improved the lower bound: for our example of Section 2 (with $w_1 = 1$), $z_1^+ = z_1^- = 5$ and $y_1 = 1/2$, (MIQP) gives a bound of 50 instead of 0. Yet,

(PR) is even better: for the same solution it gives a bound of 100, which (as expected) is the optimal solution to the problem. One can then apply the standard SOCP and SI reformulation tricks to (PR), i.e., formulae (3) and (4), to express the objective function of (PR) in terms of one SOCP constraint/infinitely many linear constraints, respectively; we denote the two thus obtained PRs of CTA as (SOCP) and (P/C), respectively.

Conversely, applying the projection approach of Frangioni et al. (2011) following the same guidelines is not possible. The reason is that the main condition required for that to work is that *the binary variable corresponding to one semicontinuous variable only appears in the corresponding constraints (9) and nowhere else*, or, in other words, that there are no constraints directly linking the binary variables to one another. This is clearly *not* the case here, as the constraint $y^+ + y^- = 1$ is crucial. While the separability requirement on the integer variables has been somewhat relaxed in Frangioni et al. (2013), the corresponding formulations are shown to be substantially weaker than the (PR) when the "linking" constraints are important in the formulation, which clearly is the case here. Hence, in order to extend the projection approach of Frangioni et al. (2011) to CTA we elected to explicitly carry out the analysis for our case. This is done by considering the function

$$g(z^+, z^-) = \min_y \left\{ \overline{co} f(\ z^+\ ,\ z^-\ ,\ y\ )\ :\ y \in [0,1] \right\} \tag{15}$$

(clearly convex, being the partial minimization of a convex function) and carrying out a case-by-case analysis of its shape. This is significantly more complex and rather tedious, so the details are best relegated to the Appendix. These can be summarized by the following Theorem.

Theorem 1 *The function $g(z^+, z^-)$ is piecewise-conic-quadratic with at most three pieces. If cell i is* reasonably balanced*, i.e.,* $\max\{\ l_i\ ,\ u_i\ \} < \min\{\ \bar{u}_i\ ,\ -\bar{l}_i\ \}$*, then $g(z^+, z^-)$ has exactly three pieces, the "central" one of which is*

$$w_i(z_i^+ + z_i^-)^2 \tag{16}$$

that is also the lower approximation *to $g(z^+, z^-)$ corresponding to the relaxation of the bounds constraints (9). If, furthermore, cell i is* totally symmetric*, i.e., $\bar{u}_i = -\bar{l}_i$ and $l_i = u_i$, then (16) actually* coincides *with $g(z^+, z^-)$.*

It would be then possible to derive a projected model analogous to those of Frangioni et al. (2011) for CTA, but the prospects of doing so are not particularly encouraging. First of all, the corresponding model would be a SOCP with up to three SOCP constraints for each sensitive cell; the standard formulation (SOCP), which already has only two of them, is typically not competitive with (P/C) (Frangioni and Gentile 2009), a fact that we directly verified to be true for CTA also. Furthermore, the rationale of Frangioni et al. (2011) is to exploit structural properties in the original problem, which are absent here for general tabular data since the matrix A lacks exploitable characteristics.

Yet, the analysis readily suggests a workable alternative: use the model

$$\text{(MIQP+)} \qquad \min \left\{ \ \sum_{i \in \mathcal{N}} w_i \big(\ z_i^+ + z_i^-\ \big)^2\ :\ (13)\ \right\}$$

instead of (MIQP), (SOCP) or (P/C). This is possible since (16) is a lower approximation to (15); furthermore, the two objective functions obviously coincide on integer solutions. The model is clearly stronger than (MIQP); on sensitive cells its objective function is weaker than that of (SOCP) or (P/C), unless in the *totally symmetric* case, in which they are equivalent. However, on non-sensitive cells its objective function is stronger than that of (SOCP) or (P/C). Note that the objective functions of (MIQP) and (MIQP+), on non-sensitive cells, could seem to actually be equivalent on the constraints (13), since these can all be written in terms of $z = z^+ - z^-$. In other words, the coefficient of z^- in every constraint is always the opposite to that of z^+. Hence, one could always assume that $z^+z^- = 0$ in the optimal solution of each continuous relaxation, since if this were not the case then one could reduce both variables at the same rate, keeping feasibility and improving the objective function value. However, this line of reasoning fails when *valid inequalities* are added to the formulation. These, typically, do not obey to the condition that the coefficients of z^+ and z^- are opposite, and therefore $z^+z^- > 0$ can (and indeed does) happen. So, in terms of strength of the continuous relaxation (and after introduction of valid inequalities) the models (MIQP+) and (PR) are not comparable. The (MIQP+) model is somewhat simpler than (SOCP), not requiring SOCP constraints; however, it has a nonseparable (albeit only slightly so) objective function. It is also more compact than (P/C), which however is a separable quadratic model.

Note that, as in the previous case, there is no need to distinguish between sensitive and non-sensitive cells: the reformulation (16) of the objective function can be applied to either, and this actually has—as it can be expected—positive results. Indeed, since $(z_i^+ + z_i^-)^2$ dominates $(z_i^+)^2 + (z_i^-)^2$, and they coincide in a minimizer, as previously seen, the analysis suggests to rather consider

$$\text{(PR+)} \qquad \min \left\{ \sum_{i\in\mathcal{U}} w_i(z_i^+ + z_i^-)^2 + \sum_{i\in\mathcal{S}} w_i\big((z_i^+)^2/y_i + (z_i^-)^2/(1-y_i) \big) \ : \ (13) \right\}$$

as the "starting" Perspective Relaxation. Thus, other than (MIQP), (SOCP), (P/C) and (MIQP+), there are two further possible models: (SOCP+) and (P/C+), obtained from (PR+) exactly as (SOCP) and (P/C) are obtained from (MIQP), respectively. Compared to (SOCP) and (P/C), these new models have (slightly) nonseparable objective function but may provide better results. The relative strengths and weaknesses of these six models can only be gauged computationally, which is done in the next section.

4 Computational Tests

We performed a large computational experiment to compare the six models (MIQP), (P/C), (SOCP), (MIQP+), (P/C+), and (SOCP+). All models have been solved with `Cplex 12.2` in single-threaded mode on a computer with 2.2 GHz AMD Opteron 6174 CPUs and 32 GB of RAM, under a GNU/Linux operating system (Ubuntu 10.10). In addition, models (MIQP), (SOCP), (MIQP+), and (SOCP+) have been solved, for some real-world difficult instances, with `Cplex 12.1` in multi-threaded mode (up to 24 parallel threads) on a computer with 3.33GHz Intel Xeon X5680 CPUs and 144 GB of RAM, under a GNU/Linux operating system (Suse 11.4). A few details are noteworthy:

- (SOCP) and (SOCP+) have been tested but were regularly worse than (P/C) and (P/C+), respectively, for single-threaded executions, confirming the results of (Frangioni and Gentile 2009); therefore, the corresponding results have not been reported.

- (P/C) and (P/C+) could not be considered for the multi-threaded executions, since the addition of perspective cuts deactivates the parallel capabilities of `Cplex`. (SOCP) and (SOCP+), which are inefficient for single-threaded executions, allow `Cplex` to exploit its parallel features, and then are considered for the multi-threaded executions.

- The large values of $\bar{l}_i$ and $\bar{u}_i$ in the instances created substantial numerical problems, whereby a variable (say z_i^-) that should have been zero (say because $y_i = 1$) actually had a "substantial" nonzero value (say because $1 - y_i \approx$ `1e-6`, and therefore $-\bar{l}_i(1 - y_i)$ was still "large"), leading to some of the cells in the table not actually being protected. This has been solved as specified below.

- The default `Cplex` parameters have been used for the computational results save for the parameter `CPX_PARAM_NUMERICALEMPHASIS` that was set to 1 to avoid numerical difficulties (see point above), likely affecting the solution speed. In general, all the versions tested were very sensitive to this particular parameter, and for many instances no solution could be found if not set to 1. In addition, (P/C) methods require also the `Cplex` parameter `PRELINEAR` set to 0 and `REDUCE` set to 1 (`PRIMALONLY`).

- The (P/C) and (P/C+) models have been implemented by means of a CPLEX `cutcallback` procedure that allows to dynamically add *user cuts* during the execution of the Branch&Cut. In particular, given the continuous solution $(\tilde{z}^+, \tilde{z}^-, \tilde{y})$ of the current node relaxation, for each $i \in \mathcal{U}$ one tests if the perspective cut for $\gamma = \tilde{z}^+/\tilde{y}$ (cf. inequality in formula (4)) is violated, and the same is done for $\gamma = \tilde{z}_i^-/(1 - \tilde{y}_i)$; all the violated perspective cuts are returned to be added to the model. Additional details about the (P/C) procedure can be found in Frangioni and Gentile (2006), Frangioni and Gentile (2009).

- The runs were performed with a time limit of 10000 seconds (wall-clock time) and, unless otherwise specified, with the `Cplex` default gap of 0.01%.

4.1 Test instances

For our tests we have considered both synthetic hierarchical instances and real-world ones. Hierarchical instances were obtained with a generator of 1H2D synthetic tables (Castro 2007) that was retrieved from `http://www-eio.upc.es/~jcastro/generators_csp.html`. This is a relevant class of instances, since a significant fraction of the tables released by NSAs are 1H2D. The generator produces a set of two-dimensional subtables with hierarchical structure according to the setting of several parameters, among which the mean number of rows per subtable, the number of columns per subtable, the depth of the hierarchical tree, the percentage of sensitive cells, the minimum and maximum number of rows with hierarchies per subtable, and the random seed. We fixed all these parameters,

but three: the mean number of rows per subtable ("r"$\in \{10, 20\}$), the number of columns per subtable ("c"$\in \{20, 30\}$), and the percentage of sensitive cells ("s"$\in \{3, 5, 10\}$). In addition, we generated both symmetric and asymmetric instances. The former have the property that $u_i = l_i$; note that in general this does *not* imply $\bar{u}_i = -\bar{l}_i$, since in many cases one has to ensure non-negativity of the perturbed values, which usually leads to $\bar{u}_i > -\bar{l}_i$. Asymmetric instances were instead obtained by considering $u_i = \mathtt{a} \cdot l_i$, "a"$\in \{2, 5, 10\}$ being the *asymmetry parameter*. Instances are thus named by the particular combination of parameters used for its generation, i.e., "r-c-s" for symmetric instances and "r-c-s-a" for asymmetric ones. For each combination of parameters we generated 5 instances varying the random generator seed, and all the reported results are averaged on these five instances.

We also dealt with a set of real-world instances. These are a subset of public instances that have been previously used in the literature (Castro 2006, Fischetti and Salazar-González 2001), and some confidential ones provided by Eurostat and the Australian NSA. Of the available real-world instances, we selected those that are neither too easy, i.e., solved by every model in a few seconds, nor too difficult, i.e., very large (up to millions of cells) and such that one cannot even find the first feasible solution—and often even solve the continuous relaxation at the root node—within the allotted timeframe. Unlike the synthetic 1H2D instances, the real-world ones have symmetric protection levels (i.e., $u_i = l_i$); as we shall see, this turns out to be a questionable modeling choice from the computational viewpoint.

Tables 1, 2, and 3 report the characteristics of, respectively, the 1H2D symmetric, 1H2D asymmetric, and real-world instances: the number of cells $|\mathcal{N}|$, the number of sensitive cells $|\mathcal{S}|$, the number of table relations m, the percentage of nonzeros in matrix A (showing that these matrices are very sparse), the number of variables and constraints in the resulting (MIQP) or (MIQP+) models, and the percentage of pure binary variables (that are in one-to-one correspondence with sensitive cells). As already mentioned, these data is averaged over the 5 instances of the same type for synthetic tables. Note that (P/C),(P/C+), (SOCP), and (SOCP+) models have more variables and constraints than these due to the reformulation tricks (3) and (4); in particular, (P/C) and (P/C+) formulations in theory have infinitely many constraints, but only finitely many ones are dynamically generated in order to approximate the objective function value of (PR) or (PR+), respectively, with the same GAP required to the solution of the problem.

4.2 Computational Results

The computational results obtained with models (MIQP+), (P/C+), (MIQP) and (P/C) in single-threaded executions are reported in Tables 4 and 5 for the symmetric and asymmetric 1H2D instances, respectively. In the tables, the column "gap" reports the gap between the value of the best feasible solution (UB) and the lower bound provided (LB) by the algorithm at termination (i.e., gap = (UB − LB)/ LB); this is the optimality gap "perceived" by the algorithm. The column "pgap" reports the analogous measure, only using the *best known lower bound* ever computed in our tests (on the same architecture) in place of LB; this is our best measure of the actual optimality gap of the feasible solution

Table 1: Size and properties of symmetric instances.

instance	$\lvert\mathcal{N}\rvert$	$\lvert\mathcal{S}\rvert$	m	%nnz	vars.	cons.	%bin
10-20-3	2877	81	452	0.47	5835	777	1.39
10-20-5	3163	150	466	0.45	6475	1064	2.31
10-20-10	2772	262	447	0.47	5806	1495	4.51
10-30-3	4569	131	612	0.34	9270	1137	1.42
10-30-5	4185	201	600	0.35	8571	1403	2.34
10-30-10	4706	452	617	0.34	9864	2426	4.59
20-20-3	6607	188	630	0.32	13401	1381	1.40
20-20-5	6426	305	621	0.33	13157	1841	2.32
20-20-10	6212	590	611	0.34	13013	2969	4.53
20-30-3	9145	264	760	0.27	18554	1816	1.42
20-30-5	8947	431	754	0.27	18324	2478	2.35
20-30-10	9164	884	761	0.27	19211	4296	4.60

produced by the algorithm, and the difference between "gap" and "pgap" gives a sense of how much weaker the lower bound attained at termination is w.r.t. the best among the four models. The columns "time" and "nodes" report, respectively, the total CPU time and the number of Branch&Cut nodes expended by the algorithm. For the sake of clarity, gaps below 0.01% are represented by a "−" and instances that hit the time limit of 10000 seconds are marked by a "∗". Note that Tables 4 and 5 show average results for the five instances of each set of parameters `r-c-s` and `r-c-s-a`. This explains that in some cases (e.g., 20-30-5 of Table 4) the average gap is positive whereas the average CPU time is below the time limit.

The results show that, as it could be expected, (MIQP) attains by far the worst results. Similarly to what has been reported several times (Frangioni and Gentile 2006, 2007, Günlük and Linderoth 2008, Aktürk et al. 2009, Frangioni et al. 2009, Hijazi et al. 2011), the use of "standard" PR techniques, i.e. (P/C) (and (SOCP), which is always worse) significantly improve on (MIQP) by delivering much better lower bounds, which in turn dramatically reduce the number of required B&C nodes. Note that typically (P/C) enumerates fewer nodes than (MIQP) in the same time, which is reasonable since adding valid inequalities requires repeated solutions of the continuous relaxation. This is true consistently both for symmetric and asymmetric instances.

In many cases (P/C+) is even more efficient than (P/C), showing that the trade-off between the (slightly) non-separable objective function and the higher bound is often favorable. This is true for all symmetric instances, and for roughly half of the asymmetric ones, in particular the smallest ones. Furthermore, most often (MIQP+) performs better than (P/C+). This is true for all symmetric instances, and for most of the asymmetric instances except some of those with large asymmetry parameter $\mathtt{a} \in \{5, 10\}$ (e.g., 10_30_10_5, 10_30_10_10, 20_30_10_5, and 20_30_10_10). This is consistent with our theoretical results: (MIQP+) and (P/C+) should provide the same lower bound on fully symmetric instances,

Table 2: Size and properties of asymmetric instances.

instance	$\lvert\mathcal{N}\rvert$	$\lvert\mathcal{S}\rvert$	m	%nnz	vars.	cons.	%bin
10-20-3-2	2877	81	452	0.47	5835	777	1.39
10-20-3-5	3163	89	466	0.45	6414	822	1.39
10-20-3-10	2919	82	454	0.46	5920	784	1.39
10-20-5-2	3095	146	462	0.45	6337	1048	2.31
10-20-5-5	2835	134	450	0.47	5804	986	2.31
10-20-5-10	3188	151	467	0.45	6526	1070	2.31
10-20-10-2	3230	306	469	0.45	6765	1691	4.52
10-20-10-5	3146	298	465	0.45	6589	1655	4.52
10-20-10-10	3024	286	459	0.46	6334	1603	4.52
10-30-3-2	4476	129	609	0.34	9081	1124	1.42
10-30-3-5	4383	126	606	0.35	8893	1110	1.41
10-30-3-10	4452	128	609	0.34	9031	1121	1.42
10-30-5-2	4439	213	608	0.35	9091	1460	2.34
10-30-5-5	4427	212	608	0.35	9066	1457	2.34
10-30-5-10	3999	192	594	0.36	8190	1360	2.34
10-30-10-2	4334	416	605	0.35	9084	2270	4.58
10-30-10-5	4204	404	601	0.35	8811	2216	4.58
10-30-10-10	4545	437	612	0.34	9526	2359	4.59
20-20-3-2	5985	170	600	0.34	12140	1280	1.40
20-20-3-5	6556	186	627	0.33	13299	1372	1.40
20-20-3-10	6737	192	636	0.32	13665	1402	1.40
20-20-5-2	5905	280	596	0.34	12091	1717	2.32
20-20-5-5	6573	312	628	0.33	13458	1876	2.32
20-20-5-10	6409	304	620	0.33	13123	1837	2.32
20-20-10-2	6082	577	605	0.34	12740	2913	4.53
20-20-10-5	6094	578	605	0.34	12767	2919	4.53
20-20-10-10	6577	624	628	0.33	13779	3126	4.53
20-30-3-2	8804	254	749	0.27	17862	1767	1.42
20-30-3-5	9219	266	762	0.27	18705	1828	1.42
20-30-3-10	9176	265	761	0.27	18617	1822	1.42
20-30-5-2	9126	440	759	0.27	18693	2519	2.35
20-30-5-5	8661	417	744	0.28	17740	2414	2.35
20-30-5-10	8996	434	755	0.27	18426	2490	2.35
20-30-10-2	9170	884	761	0.27	19224	4298	4.60
20-30-10-5	9151	883	760	0.27	19185	4291	4.60
20-30-10-10	9033	871	756	0.27	18938	4241	4.60

Table 3: Size and properties of real-world instances.

instance	$\lvert\mathcal{N}\rvert$	$\lvert\mathcal{S}\rvert$	m	%nnz	vars.	cons.	%bin
australia_ABS	24420	918	274	0.20	49758	3946	1.84
cbs	11163	2467	244	0.82	24793	10112	9.95
hier13	2020	112	3313	0.18	4152	3761	2.70
hier13x13x13a	2197	108	3549	0.15	4502	3981	2.40
hier13x13x13b	2197	108	3549	0.15	4502	3981	2.40
hier13x13x13c	2197	108	3549	0.15	4502	3981	2.40
hier13x13x13d	2197	108	3549	0.15	4502	3981	2.40
hier13x13x13e	2197	112	3549	0.15	4506	3997	2.49
hier13x13x7d	1183	75	1443	0.31	2441	1743	3.07
osorio	10201	7	202	0.99	20409	230	0.03
sbs2008_C	4212	1135	2580	0.13	9559	7120	11.87
sbs2008_E	1430	382	991	0.33	3242	2519	11.78
table7	624	17	230	1.30	1265	298	1.34
table8	1271	3	72	2.78	2545	84	0.12
targus	162	13	63	3.53	337	115	3.86

Table 4: Results for symmetric instances.

	MIQP+				P/C+				MIQP				P/C			
instance	gap	pgap	time	nodes	gap	pgap	time	nodes	gap	pgap	time	nodes	gap	pgap	time	nodes
10-20-3	–	–	442	474	–	–	486	357	6.49	–	9686	10365	–	–	1331	1973
10-20-5	–	–	765	690	–	–	1016	611	67.62	0.05	∗	2649	0.16	–	6695	8675
10-20-10	–	–	3852	10507	2.21	0.07	7660	2676	72.75	0.14	∗	5536	12.39	0.14	∗	3230
10-30-3	–	–	1470	760	–	–	1749	457	127.03	0.02	∗	778	0.98	–	9070	3022
10-30-5	–	–	4850	4003	0.07	–	7102	4769	118.53	0.12	∗	1422	15.80	0.03	∗	1853
10-30-10	2.44	2.44	∗	3512	8.26	2.53	∗	889	128.67	2.62	∗	1619	35.30	2.54	∗	643
20-20-3	–	–	1710	260	–	–	1874	291	158.64	–	∗	636	17.84	0.04	8559	596
20-20-5	–	–	3543	1507	1.27	–	7237	1185	138.59	0.12	∗	625	12.33	–	8808	481
20-20-10	7.10	7.10	∗	1968	24.51	7.21	∗	504	142.82	7.60	∗	777	38.22	7.39	∗	262
20-30-3	0.40	0.40	6113	738	3.60	0.41	6800	458	138.85	0.47	∗	726	27.17	0.45	∗	379
20-30-5	7.39	7.39	8791	751	15.19	7.46	8885	379	156.73	9.37	∗	801	32.83	8.02	∗	406
20-30-10	19.92	19.92	∗	674	32.04	21.13	∗	102	153.79	23.08	∗	496	44.06	21.20	∗	56

Table 5: Results for asymmetric instances.

	MIQP+				P/C+				MIQP				P/C			
instance	gap	pgap	time	nodes	gap	pgap	time	nodes	gap	pgap	time	nodes	gap	pgap	time	nodes
10-20-3-2	–	–	23	9	–	–	58	1	–	–	1218	7823	–	–	106	17
10-20-3-5	–	–	19	1	–	–	82	1	–	–	322	197	–	–	111	1
10-20-3-10	–	–	15	7	–	–	55	1	–	–	270	124	–	–	78	1
10-20-5-2	–	–	58	30	–	–	119	9	0.04	–	∗	113601	–	–	152	32
10-20-5-5	–	–	21	15	–	–	79	1	–	–	1293	2332	–	–	81	1
10-20-5-10	–	–	20	2	–	–	106	1	–	–	1483	660	–	–	111	1
10-20-10-2	–	–	438	556	–	–	637	181	0.04	–	∗	67541	1.49	–	2904	370
10-20-10-5	–	–	4315	31344	–	–	142	1	0.08	–	∗	102641	–	–	142	1
10-20-10-10	–	–	416	2135	–	–	120	1	0.04	–	5044	26508	–	–	109	1
10-30-3-2	–	–	115	28	–	–	271	5	0.02	–	∗	55266	–	–	391	35
10-30-3-5	–	–	40	4	–	–	220	1	–	–	2447	1333	–	–	237	1
10-30-3-10	–	–	31	1	–	–	232	1	–	–	1468	565	–	–	258	1
10-30-5-2	–	–	193	103	–	–	377	19	0.05	–	∗	28721	–	–	455	72
10-30-5-5	–	–	119	39	–	–	333	1	–	–	4055	24181	–	–	258	1
10-30-5-10	–	–	63	46	–	–	207	1	–	–	1855	1104	–	–	216	1
- 10-30-10-2	–	–	1158	1035	–	–	1905	230	7.03	–	∗	27461	0.82	–	3066	986
10-30-10-5	–	–	6489	38818	–	–	401	1	8.53	–	∗	60347	–	–	311	1
10-30-10-10	–	–	4806	22519	–	–	522	1	0.09	–	∗	52141	–	–	372	1
20-20-3-2	–	–	136	25	–	–	393	1	0.03	–	∗	13721	–	–	502	9
20-20-3-5	–	–	72	1	–	–	625	1	–	–	4074	1207	–	–	691	1
20-20-3-10	–	–	76	1	–	–	574	1	2.18	–	5356	465	–	–	644	1
20-20-5-2	–	–	257	47	–	–	601	4	1.40	–	∗	14362	–	–	598	24
20-20-5-5	–	–	117	10	–	–	690	1	1.19	–	∗	15635	–	–	638	1
20-20-5-10	–	–	128	54	–	–	736	1	0.52	–	6434	2076	–	–	623	1
20-20-10-2	–	–	1448	212	–	–	2802	138	63.41	0.04	∗	1006	–	–	2525	228
20-20-10-5	0.02	–	9203	22462	–	–	943	1	3.40	–	∗	9950	–	–	634	1
20-20-10-10	0.03	–	7910	19421	–	–	1327	1	7.33	–	∗	9801	–	–	801	1
20-30-3-2	–	–	439	28	–	–	1477	1	13.94	–	∗	1203	–	–	1649	16
20-30-3-5	–	–	140	1	–	–	1597	1	5.39	–	8400	1767	–	–	1510	1
20-30-3-10	–	–	157	8	–	–	1601	1	8.34	–	9321	691	–	–	1547	1
20-30-5-2	–	–	777	65	–	–	2160	17	48.34	–	∗	612	–	–	2111	34
20-30-5-5	–	–	618	462	–	–	1800	1	19.74	–	∗	1692	–	–	1622	1
20-30-5-10	–	–	622	243	–	–	1988	1	2.14	–	9815	2623	–	–	1625	1
20-30-10-2	1.23	1.23	7575	1454	3.67	1.24	8407	297	79.80	1.39	∗	422	4.16	1.23	7705	262
20-30-10-5	0.52	–	∗	12890	–	–	2784	1	36.91	0.03	∗	718	–	–	1915	1
20-30-10-10	0.04	–	∗	17526	–	–	2619	1	27.08	0.03	∗	1441	–	–	1817	1

and although this is not really the case even for our symmetric instances (cf. §4.1), it appears that the bounds are close enough to be roughly equivalent within the B&C approach. Indeed, the same phenomenon observed for (MIQP) and (P/C) shows off once again here: (P/C+) most often enumerates less nodes than (MIQP+), which means that the (P/C+) bound is usually somewhat stronger. However, most often (MIQP+) is faster on the instances that are solved within 10000 seconds, and it provides better gaps on the ones that stop at the time limit. This is due to the fact that, by not requiring constraint generation to compute the (approximated) PR bound, its time-per-node is lower.

The results show that, as it could be expected, the main driver of the difficulty of an instance is the percentage of sensitive cells: while instances with up to 5% of sensitive cells are routinely solved within the time limit, instances with 10% of sensitive cells are typically more difficult. However, this is only true for symmetric instances: as the asymmetry parameter "a" grows, the instances become easier. Indeed, almost all asymmetric instances are solved within 10000 seconds by (P/C) and (P/C+), and the easiest ones are associated with values $\mathtt{a} > 2$. This is not unreasonable, as a high degree of symmetry (albeit in a technically different sense) is well-known to be detrimental for combinatorial problems. Remarkably, a trade-off shows off for (MIQP+). While that model is almost invariably the best for $\mathtt{a} = 2$, it is typically worse than (P/C+) and (P/C), often by a relevant margin, when $\mathtt{a} > 2$ and $\mathtt{s} = 10$. Indeed, $\mathtt{s} = 10$ are the most difficult instances, since they involve a higher percentage of binary variables. Also, these are the cases where most often (P/C) bests (P/C+). This seems to indicate that the approximation (16) of the objective function only makes sense, both for sensitive and non-sensitive cells, when a reasonably degree of symmetry is present (which is, however, the most difficult case).

It should be remarked that protection levels, and therefore their (a)symmetry, are a choice of the modeler. Indeed, in practice NSAs derive the upper protection levels u_i from the sensitivity rules (Hundepool et al. 2012), and, as a rule of thumb, this value is assigned to the lower protection level l_i, too. Since asymmetric instances are more efficiently solved than symmetric ones, however, such a practice should be discouraged in favor of choosing decidedly asymmetric values with any appropriate heuristic. This will likely keep the same confidentiality protection and data usability in the disclosed tables while making their computation more efficient.

Tables 6 and 7 show the results on the real-world instances for, respectively, single- and multi-threaded executions. Note that the column "pgap" in each table is computed considering only the lower bounds of the four algorithms of the table, since the others were solved on a different computer and by a different `Cplex` release. As it is customary, column "time" in Table 7 reports wall-clock time.

The single-threaded results in Table 6 basically confirm these on the synthetic instances: (MIQP) is the worst model, (P/C) is significantly better, (P/C+) is usually (but not always) better yet, (MIQP+) is (at least on our test set) invariably the best. Yet, in several cases the obtained results can hardly be deemed satisfactory, with several gaps larger than 20%, and one as high as 50%. It thus makes sense to investigate if the problems can be solved with reasonable precision when more computational power is available.

The 24-threads results of Table 7 show mixed success for (SOCP) and (SOCP+);

Table 6: Single-threaded results for real instances.

	MIQP+				P/C+				MIQP				P/C			
instance	gap	pgap	time	nodes	gap	pgap	time	nodes	gap	pgap	time	nodes	gap	pgap	time	nodes
australia_ABS	2.29	2.29	∗	2401	8.08	3.15	∗	401	113.69	2.60	∗	1501	9.45	2.87	∗	763
cbs	1.24	0.61	∗	32281	0.77	0.77	∗	1801	96.38	1.22	∗	16400	0.77	0.74	∗	1801
hier13	24.77	24.77	∗	145	129.83	118.55	∗	15	111.65	27.21	∗	126	74.02	28.40	∗	73
hier13x13x13a	30.52	29.87	∗	123	29.87	29.87	∗	119	114.41	29.74	∗	200	55.41	29.74	∗	171
hier13x13x13b	30.54	29.88	∗	114	29.88	29.88	∗	99	114.41	29.74	∗	200	55.41	29.74	∗	171
hier13x13x13c	32.30	29.90	∗	100	29.90	29.90	∗	90	114.41	29.77	∗	197	55.41	29.77	∗	171
hier13x13x13d	–	–	3357	89	–	–	3479	78	–	–	9135	162	–	–	4960	100
hier13x13x13e	–	–	3293	149	–	–	3830	90	–	–	9887	185	–	–	5489	111
hier13x13x7d	–	–	1458	805	–	–	3052	1033	–	–	4995	4310	–	–	2928	2595
osorio	–	–	3754	255	–	–	6493	252	27.42	–	∗	83	0.72	–	∗	145
sbs2008_C	4.97	4.97	∗	33839	219.57	26.69	∗	389	49.66	5.24	∗	11332	12047.9	2594.78	∗	110
sbs2008_E	50.15	50.15	∗	401380	68.49	47.97	∗	11988	55.82	48.16	∗	507901	31.75	17.30	∗	7937
table7	–	–	0.55	1	–	–	5.74	1	–	–	76.61	12	–	–	3.86	1
table8	–	–	1.84	15	–	–	2.85	15	–	–	1.22	9	–	–	3.01	15
targus	–	–	0.16	3	–	–	0.28	13	–	–	0.21	16	–	–	0.32	3

Table 7: Multi-threaded results for real instances.

	MIQP+				SOCP+				MIQP				SOCP			
instance	gap	pgap	time	nodes	gap	pgap	time	nodes	gap	pgap	time	nodes	gap	pgap	time	nodes
australia_ABS	1.41	1.41	∗	30870	13.77	3.25	∗	1780	137.12	12.98	∗	9010	14.45	3.79	∗	221
cbs	1.12	0.59	∗	737465	71.70	39.80	∗	4828	94.98	1.09	∗	198867	1.00	1.00	∗	5990
hier13	–	–	4290	3185	143.47	2.99	∗	829	–	–	9403	6256	174.56	5.39	∗	0
hier13x13x13a	–	–	7038	5092	169.62	3.67	∗	0	12.15	–	∗	4609	169.08	1.79	∗	5308
hier13x13x13b	–	–	7018	5122	169.62	3.67	∗	0	–	–	9015	6750	167.00	0.83	∗	5866
hier13x13x13c	–	–	7165	5591	63.44	32.99	∗	0	–	–	7771	6425	172.30	3.04	∗	5910
hier13x13x13d	–	–	154	139	62.73	–	∗	0	–	–	409	261	109.16	0.02	∗	0
hier13x13x13e	–	–	148	169	62.73	–	∗	0	–	–	429	251	89.86	0.02	∗	0
hier13x13x7d	–	–	160	1704	120.07	9.71	∗	0	–	–	1258	11812	113.26	3.42	∗	0
hier13x7x7d	–	–	34.08	2029	56.28	3.99	∗	0	–	–	91.52	3801	56.74	2.33	∗	4
osorio	–	–	363	255	–	–	250	505	–	–	2439	255	–	0.00	224	509
sbs2008_C	2.39	2.39	∗	726715	17.37	2.82	∗	4979	4.82	2.21	∗	292576	8.09	3.20	∗	5131
sbs2008_E	39.34	0.82	∗	12640158	–	–	6097	20602	43.08	–	∗	13122442	184.90	2.49	∗	184041
table7	–	–	0.26	0	–	–	64.23	1283	–	–	28.31	9	–	–	73.89	1280
table8	–	–	1.52	15	–	–	10.40	15	–	–	0.96	9	–	–	9.85	15
targus	–	–	0.46	3	–	–	11.82	316	–	–	0.32	25	–	–	6.17	177

Table 8: Results for symmetric instances with 5% gap.

	MIQP+				P/C+				
instance	gap	pgap	time	nodes	gap	pgap	time	nodes	PCs
10-20-3	3.93	3.30	52	47	3.58	3.42	74	55	819
10-20-5	3.92	3.83	82	96	3.99	3.80	127	109	1559
10-20-10	4.96	4.96	175	344	5.41	4.97	2485	757	6734
10-30-3	4.79	4.71	230	165	4.98	4.84	249	113	1194
10-30-5	4.07	3.89	194	132	4.22	4.05	321	152	2263
10-30-10	5.14	5.04	1172	1320	9.89	5.16	5545	1123	8709
20-20-3	3.97	3.72	352	159	3.70	3.68	436	124	1815
20-20-5	3.75	3.73	594	310	4.84	3.73	1073	364	4460
20-20-10	12.37	11.17	5328	1377	13.04	11.13	8987	1944	21451
20-30-3	4.12	3.54	1265	357	3.92	3.55	1614	257	2449
20-30-5	9.43	5.96	5145	618	5.98	5.97	5248	1122	7279
20-30-10	14.72	14.72	$*$	2190	28.24	15.08	$*$	593	16350

sometimes they are better than (MIQP), sometimes worse. In general (MIQP+) is by far the best option, as in previous tables, although it is very occasionally bested by (SOCP+) (cf. sbs2008_E, one of the most difficult confidential instances). What is perhaps more relevant is that, coupled with a relatively powerful—but by no means "super"—24-threads machine, (MIQP+) is capable of providing solutions with pretty low gap for all the real-world instances in our test bed.

We finally explore the other obvious approach for reducing the required running time, i.e., accepting less accurate solutions. This is also reasonable, since in practice solutions that are a few percent off the optimal one should be more than acceptable for end users of CTA. In Tables 8, 9, and 10 we show results for, respectively, symmetric, asymmetric and real instances with a 5% gap, a setting that has also been used in the literature for L_1-CTA (Castro 2012a). Due to the previous results we only report data for the formulations (MIQP+) and (P/C+); for the latter, columns "PCs" report the number of perspective cuts. Executions were performed in the same hardware and software environment than for Table 7, except in single-threaded mode.

Clearly, solution time is now significantly reduced, providing for most instances satisfactory solutions with a moderate CPU time. For symmetric 1H2D instances, (MIQP+) is clearly the most efficient approach, being almost always faster (sometimes by more than an order of magnitude) and, with only a few exceptions, obtaining comparable or better bounds. Things are slightly more complex for asymmetric 1H2D instances: while (MIQP+) is still faster, and often significantly so, the relationships between the bounds is more erratic, with both methods displaying significantly better bounds at times and significantly worse in other cases. However, this may well depend on the fact that, with such a large gap, one model may basically stumble upon the good feasible solution needed to terminate the execution much sooner than the other. Finally, results are less clear for real instances: the two models exhibit a more similar behavior, although in a few cases

Table 9: Results for asymmetric instances with 5% gap.

	MIQP+				P/C+				
instance	gap	pgap	time	nodes	gap	pgap	time	nodes	PCs
10-20-3-2	1.57	–	4	0	2.49	2.49	23	0	450
10-20-3-5	2.90	–	4	0	0.38	0.38	29	0	459
10-20-3-10	3.37	–	3	0	1.07	1.07	22	0	401
10-20-5-2	2.47	0.03	6	0	2.82	2.82	34	0	849
10-20-5-5	2.87	–	4	0	0.84	0.84	27	0	680
10-20-5-10	1.93	–	4	0	1.38	1.38	33	0	722
10-20-10-2	3.96	0.02	85	144	4.35	4.35	123	87	3268
10-20-10-5	3.55	0.03	10	0	0.02	0.02	51	0	1890
10-20-10-10	3.26	0.03	7	0	0.26	0.26	43	0	1628
10-30-3-2	1.34	–	21	9	2.41	2.41	70	3	729
10-30-3-5	3.34	–	8	0	0.11	0.11	58	0	613
10-30-3-10	2.20	–	7	0	2.76	2.76	61	0	596
10-30-5-2	2.78	0.02	29	18	2.69	2.69	85	2	1264
10-30-5-5	2.52	–	14	0	0.05	0.05	78	0	1191
10-30-5-10	2.56	–	9	0	1.30	1.30	62	0	953
10-30-10-2	4.39	0.09	77	50	2.88	2.88	234	87	4116
10-30-10-5	3.05	0.02	23	0	0.07	0.07	89	0	2551
10-30-10-10	3.91	0.02	19	0	0.21	0.21	112	0	2409
20-20-3-2	1.98	–	14	0	2.44	2.44	100	0	931
20-20-3-5	1.96	–	14	0	3.77	3.77	150	0	983
20-20-3-10	0.38	–	13	0	0.97	0.97	135	0	887
20-20-5-2	2.76	0.02	21	0	3.24	3.24	123	0	1578
20-20-5-5	0.97	–	16	0	1.18	1.18	157	0	1673
20-20-5-10	0.58	–	16	0	1.27	1.27	156	0	1519
20-20-10-2	4.94	0.02	314	132	4.46	4.46	393	56	6304
20-20-10-5	2.52	–	25	0	1.04	1.04	180	0	3612
20-20-10-10	2.28	–	126	47	0.22	0.22	216	0	3647
20-30-3-2	2.68	–	41	0	2.28	2.28	287	0	1400
20-30-3-5	0.98	–	31	0	3.47	3.47	307	0	1340
20-30-3-10	0.29	–	31	0	1.05	1.05	323	0	1232
20-30-5-2	3.36	–	111	24	2.73	2.73	348	0	2636
20-30-5-5	2.30	–	39	0	2.02	2.02	335	0	2336
20-30-5-10	1.43	–	41	0	0.88	0.88	376	0	2258
20-30-10-2	5.13	0.55	1313	503	4.58	4.58	1555	208	10454
20-30-10-5	3.37	–	72	0	0.12	0.12	495	0	5924
20-30-10-10	2.15	–	58	0	0.31	0.31	484	0	5030

Table 10: Results for real instances with 5% gap.

	MIQP+				P/C+				
instance	gap	pgap	time	nodes	gap	pgap	time	nodes	PCs
australia_ABS	4.94	4.94	695	300	5.09	4.58	5486	1747	35395
cbs	2.43	1.93	295	2130	2.04	2.04	291	0	13917
hier13	5.26	5.23	3712	1016	5.23	5.23	5511	1059	3660
hier13x13x13a	5.25	5.20	7904	1788	5.20	5.20	6966	1620	3178
hier13x13x13b	5.25	5.20	7900	1788	5.20	5.20	7023	1620	3178
hier13x13x13c	5.25	5.20	7908	1788	5.20	5.20	6907	1620	3178
hier13x13x13d	5.02	2.12	683	56	2.12	2.12	660	50	1761
hier13x13x13e	4.97	4.69	645	71	4.69	4.69	745	73	1800
hier13x13x7d	5.21	5.21	312	433	5.22	5.21	666	581	1567
osorio	1.85	1.85	3	0	1.81	1.81	64	1	27
sbs2008_C	5.26	5.26	4740	56479	56.94	6.17	$*$	1081	31708
sbs2008_E	44.25	44.25	$*$	1156201	55.40	42.75	$*$	44615	14305
table7	–	–	0	0	–	–	3	0	75
table8	0.71	0.71	0	0	1.12	0.71	1	2	12
targus	1.11	1.11	0	0	1.12	1.12	0	0	66

(e.g. australia_ABS, osorio, and sbs2008_C) (MIQP+) has a clear edge, and where it is slower it's not so by much more than 10%. It is worth noting that some real instances exhibit a gap greater than 5%; since each row corresponds to an instance this is not, like it happened in other cases, due to averaging, but rather to the (somewhat questionable) habit of `Cplex` to define the gap as (UB − LB)/UB.

To summarize, it is fair to say that (MIQP+) obtains good results for all the types of instances, showing that appropriate modeling techniques combined with state-of-the-art, general-purpose MIQP solvers can provide accurate solutions to real-life (and realistic) instances within a reasonable timeframe, especially if one is ready to throw at the problem a slightly more substantial amount of computational resources and/or accept solutions with a somewhat larger gap.

4.3 About the utility of protected data

As stated in the introduction, one of the possible benefits of using L_2-CTA, as opposed to L_1-CTA, is the expected higher utility of published data. This is explored in this section, where we empirically compare the data utility provided by both approaches on our test bed.

We measure "data utility" as in Castro and Giessing (2006), i.e., by counting the number of published cells which changed "too much" with respect to their original values. To measure the change in a cell we consider the percentage deviation, i.e., $\%\text{dev} = |z|/|a| \cdot 100$; cells with percentage deviations above a threshold value are considered to have changed "too much". We set the threshold value to one fourth of the maximum percentage deviation of L_1-CTA, which was almost (but on two instances) always greater than the maximum percentage deviation of L_2-CTA.

Table 11: Utility of protected tables for symmetric instances.

	L_1			L_2		
instance	mean	max	large	mean	max	large
sym-10-20-3	27.8	9425.3	6	6.3	318.4	0
sym-10-20-5	36.6	9900.0	11	7.3	624.6	0
sym-10-20-10	36.0	9420.0	11	6.5	337.4	0
sym-10-30-3	31.6	9197.4	9	7.9	270.4	0
sym-10-30-5	44.5	9900.0	12	8.3	429.0	0
sym-10-30-10	36.7	9900.0	10	7.1	284.0	0
sym-20-20-3	60.9	9900.0	19	15.0	555.0	0
sym-20-20-5	82.5	9900.0	28	15.5	430.2	0
sym-20-20-10	82.6	9900.0	25	14.3	842.2	0
sym-20-30-3	30.7	9754.7	13	6.9	504.9	0
sym-20-30-5	28.5	9900.0	10	5.6	349.1	0
sym-20-30-10	33.0	9900.0	12	5.7	463.6	0

Tables 11, 12, and 13 show the results for, respectively, symmetric, asymmetric and real-world instances. As in previous sections, results for symmetric and asymmetric synthetic tables are averaged over five instances. The tables show, both for L_1 and L_2, the average (columns "mean") and maximum (columns "max") percentage deviation over all the cells in the table, as well as the number of cells with "large" percentage deviations (columns "large") defined as discussed above.

The tables clearly show that L_2-CTA typically provides much more "useful" data. Indeed, for symmetric and asymmetric synthetic tables the average and maximum percentage deviations of L_2-CTA were always smaller than those of L_1-CTA, and the number of cells with large deviations was always 0 for L_2-CTA. For real-world tables, L_2-CTA always provided smaller average deviations and, excluding instances "hier13x13x13d" and "hier13x13x13e", the same can be said for the maximum deviations. The number of cells with large deviations was also smaller for L_2 than for L_1, although, unlike for synthetic tables, it is not always zero. Therefore, it can be concluded that, at least using as criteria the number of cells with large percentage deviations, the utility of tables protected with L_2-CTA is higher than that of tables protected with L_1-CTA.

It is fair to mention that L_1-CTA may outperform L_2-CTA for other measures. For instance, the number of (non-sensitive) cells with zero deviations (nonperturbed cells) provided by L_1-CTA is typically greater than for L_2-CTA, as more variables can be expected to be fixed at their bounds when minimizing a linear objective instead of a quadratic one. However this is not a main inconvenience; indeed, CTA is used in practice as a second stage after the introduction of stochastic noise (Giessing 2012), so original cell values will anyway be modified.

Table 12: Utility of protected tables for asymmetric instances.

	L_1			L_2		
instance	mean	max	large	mean	max	large
asym-10-20-3-2	27.8	9425.3	6	6.3	318.4	0
asym-10-20-3-5	36.6	9900.0	11	7.3	624.6	0
asym-10-20-3-10	36.0	9420.0	11	6.5	337.4	0
asym-10-20-5-2	31.6	9197.4	9	7.9	270.4	0
asym-10-20-5-5	44.5	9900.0	12	8.3	429.0	0
asym-10-20-5-10	36.7	9900.0	10	7.1	284.0	0
asym-10-20-10-2	60.9	9900.0	19	15.0	555.0	0
asym-10-20-10-5	82.5	9900.0	28	15.5	430.2	0
asym-10-20-10-10	82.6	9900.0	25	14.3	842.2	0
asym-10-30-3-2	30.7	9754.7	13	6.9	504.9	0
asym-10-30-3-5	28.5	9900.0	10	5.6	349.1	0
asym-10-30-3-10	33.0	9900.0	12	5.7	463.6	0
asym-10-30-5-2	42.9	9900.0	18	9.0	442.5	0
asym-10-30-5-5	52.9	9900.0	23	9.6	621.8	0
asym-10-30-5-10	45.5	9900.0	17	9.0	449.4	0
asym-10-30-10-2	62.1	9900.0	25	14.4	665.8	0
asym-10-30-10-5	74.8	9900.0	31	14.6	641.8	0
asym-10-30-10-10	91.9	9900.0	43	15.0	469.3	0
asym-20-20-3-2	31.6	9900.0	18	5.9	623.3	0
asym-20-20-3-5	38.0	9900.0	23	6.6	495.7	0
asym-20-20-3-10	31.8	9900.0	21	5.4	515.5	0
asym-20-20-5-2	40.7	9900.0	20	7.2	285.3	0
asym-20-20-5-5	40.6	9900.0	23	7.5	573.6	0
asym-20-20-5-10	42.9	9900.0	24	8.1	559.3	0
asym-20-20-10-2	67.6	9900.0	41	14.7	798.5	0
asym-20-20-10-5	88.7	9900.0	55	15.5	866.4	0
asym-20-20-10-10	91.1	9900.0	61	15.9	916.1	0
asym-20-30-3-2	30.8	9900.0	26	6.4	799.6	0
asym-20-30-3-5	36.0	9900.0	33	5.8	562.2	0
asym-20-30-3-10	36.2	9900.0	32	5.4	480.0	0
asym-20-30-5-2	41.0	9900.0	35	8.4	837.5	0
asym-20-30-5-5	51.6	9900.0	44	8.9	578.8	0
asym-20-30-5-10	54.2	9900.0	51	8.9	532.5	0
asym-20-30-10-2	72.8	9900.0	70	15.8	966.4	0
asym-20-30-10-5	87.7	9900.0	84	13.4	804.9	0
asym-20-30-10-10	88.3	9900.0	79	14.8	771.5	0

Table 13: Utility of protected tables for real instances.

	L_1			L_2		
instance	mean	max	large	mean	max	large
australia_ABS	70.5	153000.0	7	5.8	545.5	0
cbs	115.6	55400.0	33	9.3	156.5	0
hier13	0.9	69.7	11	0.8	18.8	2
hier13x13x13a	0.8	69.7	11	0.7	18.8	2
hier13x13x13b	0.8	69.7	11	0.7	18.8	2
hier13x13x13c	0.8	69.7	11	0.7	18.8	2
hier13x13x13d	1.1	82.5	16	0.9	99.6	6
hier13x13x13e	1.1	82.5	16	0.9	99.6	6
hier13x13x7d	1.1	46.5	15	0.7	12.2	1
hier13x7x7d	1.0	28.3	22	0.8	27.7	15
osorio	0.1	100.0	8	0.1	100.0	2
sbs2008_C	44279.7	13158600.0	21	1280.8	728200.0	0
sbs2008_E	3964.4	2023400.0	2	482.9	197875.0	0
table7	3.0	200.0	13	2.4	140.0	9
table8	0.3	150.0	3	0.1	11.1	0
targus	3.3	50.0	15	2.8	37.5	13

5 Conclusions

This paper studies the CTA problem with L_2 distance. The peculiar structure of the problem are *pairs of alternative semicontinuous variables*, such that exactly one of them is nonzero in any feasible solution. Exploiting ideas from the Perspective Reformulation approach, we developed and analyzed several MIQP, SOCP, and Semi-Infinite LP strong formulations for the problem, which provide different degrees of approximation to the objective function of the classical PR. We show that one particularly simple MIQP model, derived applying a relaxation of the P^2R technique, is often preferable, from the computational viewpoint, at least on instances that are not "too much asymmetric". Yet, other models are better on highly asymmetric and with a large percentage of binary variables instances, which are usually easier to solve; this also provides indications to practitioners about setting the protection levels to the cells in order to make the instances more easily solvable. The right choice of the model allows to solve real-life instances in reasonable time with off-the-shelf, general-purpose MIQP solvers, at least on relatively powerful multi-core computers and/or if a relatively large optimality gap is allowed. The published tables can be expected to be "more useful" than those obtained with the L_1 distance, albeit at a greater computational effort. Indeed, since CTA is a difficult problem, instances with a large number of sensitive cells and/or a high degree of symmetry remain difficult to solve with high accuracy; further research will then be required to improve the effectiveness of the solution methods for these cases. Also, the specific structure of CTA may show up,

perhaps with non-quadratic functions, in other applications: the techniques developed in this paper could be adaptable to these cases.

Acknowledgements

This work has been supported by grants MTM2012-31440 of the Spanish Ministry of Economy and Competitiveness, and SGR-2009-1122 of the Government of Catalonia. Claudio Gentile has been partly supported by the Short-Term-Mobility program of the Italian National Research Council.

Appendix. Proof of Theorem 1.

As in Section 3 we will concentrate on a fixed cell $i \in \mathcal{S}$ and therefore drop the index i. Also, in the development we assume w.l.o.g. $w = 1$, because it is a multiplicative factor which just goes untouched through the derivation. It is easy to see that the constraint

$$\min\{\, l \,,\, u \,\} \leq z^+ + z^- \leq \max\{\, \bar{u} \,,\, -\bar{l} \,\} \tag{17}$$

is implied by (9): in all integral solutions one has either $z^+ \leq \bar{u}$ and $z^- = 0$, or $z^- \leq -\bar{l}$ and $z^+ = 0$, and, analogously, either $z^+ \geq u$ and $z^- = 0$ or $z^- \geq l$ and $z^+ = 0$. Therefore, we can consider (17) as explicitly added to the formulation if we need it. Furthermore, the constraints $0 \leq z^+ \leq \bar{u}$ and $0 \leq z^- \leq -\bar{l}$ are always valid.

From (9) we immediately obtain

$$\begin{aligned} 0 \leq z^+/\bar{u} \quad &\leq y \leq \quad z^+/u \\ (l - z^-)/l \quad &\leq y \leq \quad (z^- + \bar{l})/\bar{l} \leq 1 \end{aligned}$$

which yields

$$\delta(z^+, z^-) = \max\left\{\frac{z^+}{\bar{u}}, 1 - \frac{z^-}{l}\right\} \leq y \leq \min\left\{\frac{z^+}{u}, 1 + \frac{z^-}{\bar{l}}\right\} = \Delta(z^+, z^-) \ . \tag{18}$$

We now want to develop a closed-form formula for the optimal solution $y(z^+, z^-)$ of (15). We therefore need to find the value of y such that

$$\frac{\partial h(z^+, z^-, y)}{\partial y} = -\frac{(z^+)^2}{y^2} + \frac{(z^-)^2}{(1-y)^2} = 0,$$

where $h(\, z^+ \,,\, z^- \,,\, y \,) = \overline{co} f(\, z^+ \,,\, z^- \,,\, y \,)$ in (11), which leads to

$$\begin{aligned} (1-y)^2(z^+)^2 = y^2(z^-)^2 \quad &\Leftrightarrow \quad (1 - 2y + y^2)(z^+)^2 = y^2(z^-)^2 \\ y^2((z^+)^2 - (z^-)^2) - 2y(z^+)^2 + (z^+)^2 = 0 \quad &\Leftrightarrow \quad y = z^+/(z^+ + z^-) = \tilde{y} \end{aligned}$$

as $0 \leq y \leq 1$, $z^+ \geq 0$ and $z^- \geq 0$. In fact, the other root of the quadratic equation, $z^+/(z^+ - z^-)$, coincides with $\tilde{y}$ when $z^- = 0$, is > 1 when $z^+ > z^- > 0$, is indefinite when

$z^+ = z^-$ and is < 0 when $z^- > z^+$, and therefore is never relevant. Moreover, the second derivative

$$\frac{\partial^2 h(z^+, z^-, y)}{\partial y^2} = 2\frac{(z^+)^2}{y^3} + 2\frac{(z^-)^2}{(1-y)^3}$$

is greater then zero in $y = \tilde{y}$ when $0 < \tilde{y} < 1$. Me must now distinguish three cases:

1) $\tilde{y} \leq \delta(z^+, z^-) \quad \Rightarrow \quad y(z^+, z^-) = \delta(z^+, z^-)$;

2) $\delta(z^+, z^-) < \tilde{y} < \Delta(z^+, z^-) \quad \Rightarrow \quad y(z^+, z^-) = \tilde{y}$;

3) $\Delta(z^+, z^-) \leq \tilde{y} \quad \Rightarrow \quad y(z^+, z^-) = \Delta(z^+, z^-)$.

For case 2), plugging $y = \tilde{y} = z^+/(z^+ + z^-)$ into (9) gives

$$u \leq z^+ + z^- \leq \bar{u} \quad \text{and} \quad l \leq z^+ + z^- \leq -\bar{l} \ . \tag{19}$$

Therefore, under these conditions, the optimal objective function value $f^*(z^+, z^-) = h(z^+, z^-, \tilde{y})$ takes the particularly simple form

$$f^*(\ z^+\ ,\ z^-\ ) = h(\ z^+\ ,\ z^-\ ,\ z^+/(z^+ + z^-)\ ) = (z^+ + z^-)^2 \ ,$$

i.e., (16). Hence, in the totally symmetric case $\bar{u} = -\bar{l}$, $l = u$ one has $\max\{\ \bar{u}\ ,\ -\bar{l}\ \} = \min\{\ \bar{u}\ ,\ -\bar{l}\ \}$ and $\max\{\ u\ ,\ l\ \} = \min\{\ u\ ,\ l\ \}$, only case 2) can happen: $g(z^+, z^-) = f^*(z^+, z^-)$. Note that, as claimed in the Theorem, (16) $\equiv f^*(z^+, z^-) \leq g(z^+, z^-)$ as it corresponds to *unconstrained* minimization over y.

With non-symmetric data, cases 1) and 3) has to be taken into account. The analysis has to be divided into several sub-cases.

1) $\tilde{y} \leq \delta(z^+, z^-)$. Because $\delta(z^+, z^-) = \max\{z^+/\bar{u}, 1 - z^-/l\}$, two sub-cases have to be separately considered:

   1.1) $z^+/\bar{u} \geq 1 - z^-/l$ and $\tilde{y} \leq z^+/\bar{u}$; by simple algebraic manipulations, these two conditions boil down to

$$\begin{aligned} lz^+ + \bar{u}z^- &\geq \bar{u}l & (20)\\ z^+ + z^- &\geq \bar{u} & (21) \end{aligned}$$

   By rewriting (20) in the equivalent form

$$z^+ + z^-(\bar{u}/l) \geq \bar{u}$$

   it is immediately evident that one among (20) and (21) is redundant when the other is imposed; this depends on which of the two conditions

$$\begin{aligned} \bar{u} &\leq l & (22)\\ l &\leq \bar{u} & (23) \end{aligned}$$

   holds. In particular,

* (22) $\Rightarrow$ (20) dominates (21);
* (23) $\Rightarrow$ (21) dominates (20).

In either case we have $y(z^+, z^-) = z^+/\bar{u}$, which finally leads to

$$g(\ z^+\ ,\ z^-\ ) = h(\ z^+\ ,\ z^-\ ,\ z^+/\bar{u}\ ) = \bar{u}((z^-)^2/(\bar{u} - z^+) + z^+)\ . \tag{24}$$

Note that the objective function value is always positive, as $z^+ \leq \bar{u}$.

1.2) $z^+/\bar{u} \leq 1 - z^-/l$ and $\tilde{y} \leq 1 - z^-/l$; this gives

$$\begin{aligned} lz^+ + \bar{u}z^- &\leq \bar{u}l & (25)\\ z^+ + z^- &\leq l & (26) \end{aligned}$$

Again, by rewriting (25) in the equivalent form

$$z^+(l/\bar{u}) + z^- \leq l$$

we see that one of these is redundant when the other is imposed, depending on *the same* conditions (22)/(23); that is,

* (22) $\Rightarrow$ (25) dominates (26);
* (23) $\Rightarrow$ (26) dominates (25).

In either case we have $y(z^+, z^-) = 1 - z^-/l$, which finally leads to

$$g(\ z^+\ ,\ z^-\ ) = h(\ z^+\ ,\ z^-\ ,\ 1 - z^-/l\ ) = l((z^+)^2/(l - z^-) + z^-)\ . \tag{27}$$

Note that the objective function value is always positive, as $z^- \leq z^+ + z^- \leq l$.

3) $\Delta(z^+, z^-) \leq \tilde{y}$. Because $\Delta(z^+, z^-) = \min\{z^+/u, 1 + z^-/\bar{l}\}$, again this can happen in two different ways:

3.1) $z^+/u \leq 1 + z^-/\bar{l}$ and $\tilde{y} \geq z^+/u$; this is equivalent to

$$\begin{aligned} -\bar{l}z^+ + uz^- &\leq -\bar{l}u & (28)\\ z^+ + z^- &\leq u & (29) \end{aligned}$$

where as usual (28) can be rewritten as $z^+ + z^-(u/-\bar{l}) \leq u$. Thus, according to which among

$$\begin{aligned} -\bar{l} &\leq u & (30)\\ u &\leq -\bar{l} & (31) \end{aligned}$$

holds, one of the constraints is useless; indeed,

* (30) $\Rightarrow$ (28) dominates (29);
* (31) $\Rightarrow$ (29) dominates (28).

In either case we have $y(z^+, z^-) = z^+/u$, which finally leads to

$$g(\ z^+\ ,\ z^-\ ) = h(\ z^+\ ,\ z^-\ ,\ z^+/u\ ) = u((z^-)^2/(u - z^+) + z^+)\ . \tag{32}$$

Note that the objective function value is always positive, as $z^+ \leq z^+ + z^- \leq u$.

3.2) $z^+/u \geq 1 + z^-/\bar{l}$ and $\tilde{y} \geq 1 + z^-/\bar{l}$; one has

$$\begin{aligned} -\bar{l}z^+ + uz^- &\geq -\bar{l}u & (33)\\ z^+ + z^- &\geq -\bar{l} & (34) \end{aligned}$$

According to which among (30)/(31) holds, one of the above (considering that (33) can be rewritten as $z^+(-\bar{l}/u) + z^- \geq -\bar{l}$) is irrelevant; that is,

- (30) $\Rightarrow$ (33) dominates (34);
- (31) $\Rightarrow$ (34) dominates (33).

In either case we have $y(z^+, z^-) = 1 + z^-/\bar{l}$, which finally leads to

$$g(\ z^+\ ,\ z^-\ ) = h(\ z^+\ ,\ z^-\ ,\ 1 + z^-/\bar{l}\ ) = (-\bar{l})((z^+)^2/(-\bar{l} - z^-) + z^-)\ . \tag{35}$$

Again, the objective function value is always positive, as $z^- \leq -\bar{l}$.

From the above discussion we conclude, remembering that $0 \leq u \leq \bar{u}$ and $0 \leq l \leq -\bar{l}$, that the (z^+, z^-) space can be partitioned into several subsets, in each of which the objective function is uniquely determined. Again this requires a case-by-case discussion:

- If $\bar{u} \leq l$ (cf. (22)), then $\max\{l, u\} = l \geq \min\{\bar{u}, -\bar{l}\} = \bar{u}$; therefore, case 2) is not significant (cf. 19). Because (20) dominates (21) and (25) dominates (26), we have that for all $u \leq z^+ + z^- \leq -\bar{l}$

$$g(z^+, z^-) = \begin{cases} \bar{u}((z^-)^2/(\bar{u} - z^+) + z^+) & \text{if } lz^+ + \bar{u}z^- \geq \bar{u}l \\ l((z^+)^2/(l - z^-) + z^-) & \text{if } lz^+ + \bar{u}z^- \leq \bar{u}l \end{cases}\ .$$

- Analogously, if $-\bar{l} \leq u$ (cf. (30)), then $\max\{l, u\} = u \geq \min\{\bar{u}, -\bar{l}\} = -\bar{l}$; therefore, case 2) does not happen (cf. 19). Because (28) dominates (29) and (33) dominates (34), we have that for all $l \leq z^+ + z^- \leq \bar{u}$

$$g(z^+, z^-) = \begin{cases} u((z^-)^2/(u - z^+) + z^+) & \text{if } -\bar{l}z^+ + uz^- \leq -\bar{l}u \\ (-\bar{l})((z^+)^2/(-\bar{l} - z^-) + z^-) & \text{if } -\bar{l}z^+ + uz^- \geq -\bar{l}u \end{cases}\ .$$

If none of the above two "extreme" cases occur, then the "simple" inequalities (21), (26), (29), and (34) all dominate their "complex" companions (20), (25), (28), and (33), respectively. We can thus continue the discussion listing all other possible ways in which l, u, $-\bar{l}$ and $\bar{u}$ can be arranged along the line:

- If $l \leq u \leq \bar{u} \leq -\bar{l}$, then $\max\{l,u\} = u$ and $\min\{\bar{u}, -\bar{l}\} = \bar{u}$. Thus,

$$g(z^+, z^-) = \begin{cases} u((z^-)^2/(u - z^+) + z^+) & \text{if } l \leq z^+ + z^- \leq u \\ (z^+ + z^-)^2 & \text{if } u \leq z^+ + z^- \leq \bar{u} \\ \bar{u}((z^-)^2/(\bar{u} - z^+) + z^+) & \text{if } \bar{u} \leq z^+ + z^- \leq -\bar{l} \end{cases}$$

- If $l \leq u \leq -\bar{l} \leq \bar{u}$, then $\max\{l,u\} = u$ and $\min\{\bar{u}, -\bar{l}\} = -\bar{l}$. Thus,

$$g(z^+, z^-) = \begin{cases} u((z^-)^2/(u - z^+) + z^+) & \text{if } l \leq z^+ + z^- \leq u \\ (z^+ + z^-)^2 & \text{if } u \leq z^+ + z^- \leq -\bar{l} \\ (-\bar{l})((z^+)^2/(-\bar{l} - z^-) + z^-) & \text{if } -\bar{l} \leq z^+ + z^- \leq \bar{u} \end{cases}$$

- If $u \leq l \leq -\bar{l} \leq \bar{u}$, then $\max\{l,u\} = l$ and $\min\{\bar{u}, -\bar{l}\} = -\bar{l}$. Thus,

$$g(z^+, z^-) = \begin{cases} l((z^+)^2/(l - z^-) + z^-) & \text{if } u \leq z^+ + z^- \leq l \\ (z^+ + z^-)^2 & \text{if } l \leq z^+ + z^- \leq -\bar{l} \\ (-\bar{l})((z^+)^2/(-\bar{l} - z^-) + z^-) & \text{if } -\bar{l} \leq z^+ + z^- \leq \bar{u} \end{cases}$$

- If $u \leq l \leq \bar{u} \leq -\bar{l}$, then $\max\{l,u\} = l$ and $\min\{\bar{u}, -\bar{l}\} = \bar{u}$. Thus,

$$g(z^+, z^-) = \begin{cases} l((z^+)^2/(l - z^-) + z^-) & \text{if } u \leq z^+ + z^- \leq l \\ (z^+ + z^-)^2 & \text{if } l \leq z^+ + z^- \leq \bar{u} \\ \bar{u}((z^-)^2/(\bar{u} - z^+) + z^+) & \text{if } \bar{u} \leq z^+ + z^- \leq -\bar{l} \end{cases}$$

Thus, we have a total of 6 possible cases; in 4 of them the function has three pieces, two SOCP ones and a quadratic one, while in the remaining 2 the function has two pieces, all of them being SOCP. We have therefore completed the proof of Theorem 1.

References

Aktürk, S., A. Atamtürk, S. Gürel. 2009. A strong conic quadratic reformulation for machine-job assignment with controllable processing times. *Oper. Res. Lett.* **37**(3) 187–191.

Bacharach, M. 1966. Matrix rounding problems. *Management Sci.* **9** 732–742.

Castro, J. 2006. Minimum-distance controlled perturbation methods for large-scale tabular data protection. *Eur. J. Oper. Res.* **171** 39–52.

Castro, J. 2007. A shortest paths heuristic for statistical disclosure control in positive tables. *INFORMS J.Comput.* **19** 520–533.

Castro, J., S. Giessing. 2006. Testing variants of minimum distance controlled tabular adjustment. *Monographs of Official Statistics*, Eurostat-Office for Official Publications of the European Communities, Luxembourg, 333–343.

Castro, J., J. Cuesta. 2011. Quadratic regularizations in an interior-point method for primal block-angular problems. *Math. Prog.* **130** 415–445.

Castro, J. 2012a. Recent advances in optimization techniques for statistical tabular data protection. *Eur. J. Oper. Res.* **216** 257–269.

Castro, J. 2012b. On assessing the disclosure risk of controlled adjustment methods for statistical tabular data. *Internat. J. of Uncertainty, Fuzziness and Knowledge-Based Systems* **20** 921–941.

Ceria, S., J. Soares. 1999. Convex programming for disjunctive convex optimization. *Math. Prog.* **86** 595–614.

Dandekar, R.A., L.H. Cox. 2002. Synthetic tabular Data: an alternative to complementary cell suppression. Manuscript, Energy Information Administration, US.

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

Fischetti, M., J.J. Salazar-González. 2001. Solving the cell suppression problem on tabular data with linear constraints. *Management Sci.* **47** 1008–1026.

Frangioni A., F. Furini, C. Gentile. 2013. Approximated Perspective Relaxations: a Project&Lift Approach. *Tech. Report*, **13-04**, Dipartimento di Informatica, Università di Pisa.

Frangioni, A., C. Gentile. 2006. Perspective cuts for 0-1 mixed integer programs. *Math. Prog.* **106(2)** 225–236.

Frangioni, A., C. Gentile. 2007. SDP diagonalizations and perspective cuts for a class of nonseparable MIQP. *Oper. Res. Lett.* **35(2)** 181–185.

Frangioni, A., C. Gentile. 2009. A computational comparison of reformulations of the perspective relaxation: SOCP vs. cutting planes. *Oper. Res. Lett.* **37**(3) 206–210.

Frangioni A., C. Gentile, E. Grande, A. Pacifici. 2011. Projected perspective reformulations with applications in design problems. *Oper. Res.*, **59**(5) 1225–1232.

Frangioni, A., C. Gentile, F. Lacalandra. 2009. Tighter approximated MILP formulations for unit commitment problems. *IEEE Trans. Power Systems* **24**(1) 105–113.

Giessing, S. 2012. Federal Statistical Office of Germany. Personal communication in the scope of the "DwB. Data without Boundaries" Project INFRA-2010-262608, VII Mark Program of the European Union.

Giessing, S., A. Hundepool, J. Castro. 2009. Rounding methods for protecting EU-aggregates. *Worksession on statistical data confidentiality. Eurostat methodologies and working papers*, Eurostat-Office for Official Publications of the European Communities, Luxembourg, 255–264.

González, J.A., J. Castro. 2011. A heuristic block coordinate descent approach for controlled tabular adjustment. *Comput. Oper. Res.* **38** 1826-1835.

Grossmann, I., S. Lee. 2003. Generalized convex disjunctive programming: nonlinear convex hull relaxation. *Comput. Optim. App.* **26** 83-100.

Günlük, O., J. Linderoth. 2008. Perspective relaxation of MINLPs with indicator variables. A. Lodi, A. Panconesi, G. Rinaldi, eds., *Proceedings 13th IPCO*, *Lecture Notes in Computer Science*, vol. 5035. 1–16.

Günlük, O., J. Linderoth. 2009. Perspective reformulation and applications. *Research Report* **RC24858**, IBM.

Hijazi, H., P. Bonami, G. Cornuejols, A. Ouorou. 2010. Mixed integer nonlinear programs featuring "On/Off" constraints: convex analysis and applications. *Electronic Notes in Discrete Mathematics* **36** 1153–1160.

Hundepool, A., J. Domingo-Ferrer, L. Franconi, S. Giessing, E. Schulte-Nordholt, K. Spicer, P.P. de Wolf. 2012. *Statistical Disclosure Control*, Wiley, Chichester.

Khajavirad, A., N.V. Sahinidis. 2013. Convex envelopes generated from finitely many compact convex sets. *Math. Prog.* **137**(1-2) 371–408.

Kelly, J.P., B.L. Golden, A.A. Assad. 1992. Cell suppression: disclosure protection for sensitive tabular data. *Networks* **22** 28–55.

Luedtke, J., M. Namazifar, J.T. Linderoth. 2010. Some results on the strength of relaxations of multilinear functions *Technical Report* **#1678**, Computer Sciences Department, University of Wisconsin-Madison.

Nesterov, Y. 2004. *Introductory Lectures on Convex Optimization*, Kluwer, Boston.

Salazar-González, J.J. 2008. Statistical confidentiality: Optimization techniques to protect tables, *Comput. Oper. Res.* **35** 1638–1651.

Tawarmalani, M., J.P.P. Richard, C. Xiong. 2012. Explicit convex and concave envelopes through polyhedral subdivisions. *Math. Prog.* **138**(1-2) 531–577.

Tawarmalani, M., N.V. Sahinidis. 2001. Semidefinite Relaxations of Fractional Programs via Novel Convexification Techniques. *J. Glob. Opt.* **20**(2) 133–154.

Tawarmalani, M., N.V. Sahinidis. 2002. Convex extensions and envelopes of lower semi-continuous functions. *Math. Prog.* **93** 515–532.

Willenborg, L., T. de Waal. 2000. *Lecture Notes in Statistics. Elements of Statistical Disclosure Control* 155, Springer, New York.

Zayatz, L. 2009. U.S. Census Bureau. Communication at Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality, Bilbao (Basque Country, Spain).