Computers & Operations Research 58 (2015) 41–52

Contents lists available at ScienceDirect

Computers & Operations Research

journal homepage: www.elsevier.com/locate/caor

Fix-and-relax approaches for controlled tabular adjustment

Daniel Baena, Jordi Castro*, José A. González

Dept. of Statistics and Operations Research, Universitat Politècnica de Catalunya, Campus Nord, Office C5203, Jordi Girona 1–3, 08034 Barcelona, Catalonia, Spain

ARTICLE INFO

Available online 8 January 2015

Keywords:
Fix-and-relax
Block coordinate descent
Mixed-integer linear programming
Controlled tabular adjustment
Primal heuristics
Feasibility pump
Statistical disclosure control

ABSTRACT

Controlled tabular adjustment (CTA) is a relatively new protection technique for tabular data protection. CTA formulates a mixed integer linear programming problem, which is challenging for tables of moderate size. Even finding a feasible initial solution may be a challenging task for large instances. On the other hand, end users of tabular data protection techniques give priority to fast executions and are thus satisfied in practice with suboptimal solutions. This work has two goals. First, the fix-and-relax (FR) strategy is applied to obtain good feasible initial solutions to large CTA instances. FR is based on partitioning the set of binary variables into clusters to selectively explore a smaller branch-and-cut tree. Secondly, the FR solution is used as a warm start for a block coordinate descent (BCD) heuristic (approach named FR+BCD); BCD was confirmed to be a good option for large CTA instances in an earlier paper by the second and third co-authors (Comput Oper Res 2011;38:1826–35 [23]). We report extensive computational results on a set of real-world and synthetic CTA instances. FR is shown to be competitive compared to CPLEX branch-and-cut in terms of quickly finding either a feasible solution or a good upper bound. FR+BCD improved the quality of FR solutions for approximately 25% and 50% of the synthetic and real-world instances, respectively. FR or FR+BCD provided similar or better solutions in less CPU time than CPLEX for 73% of the difficult real-world instances.

1. Introduction

Microdata and tabular data protection are the two main disciplines of statistical disclosure control. The purpose of this field is to avoid that confidential information can be derived from data released. This is one of the main concerns of National Statistical Agencies (NSAs), which have to disseminate a large amount of information minimizing at the same time the disclosure risk of individual respondents. Tabular data is obtained by crossing two or more categorical variables in a microdata file. For each cell, the table may report either the number of individuals (frequency tables) or information about another variable (magnitude tables). More details can be found in the recent survey [5] and the monographs [26,27].

Although cell tables report aggregated information for several respondents—so they could be considered anonymized—there is a risk of disclosing individual data. Fig. 1 illustrates this situation with a simple case. The left table (a) reports the average salary of individuals by age (row variable) and town (column variable), while table (b) provides the number of individuals. If there were only one individual of age between 51 and 55 in town t_2, then any external attacker would know the confidential salary of this person. For two individuals, any of them could disclose the other's salary, becoming an internal attacker. Cells that require protection (such as that of the example) are named sensitive, unsafe, or confidential cells. Sensitive cells are a priori detected by some sensitivity rules. The above example showed the simplest *minimum-frequency* rule, which considers sensitive those cells with very few respondents. The most widely used rule, named p-% rule, considers a cell unsafe if some respondent may obtain an estimate of another respondent contribution within a p-% precision. A detailed description of these rules can be found in [27].

A tabular data protection method can be seen as a map F such that $F(T)=T'$, i.e., table T is transformed to another table T'. Two are the main requirements for F: (1) the output table T' should be "safe", and (2) the quality of T' should be high (or equivalently, the information loss should be small), i.e., T' should be a good replacement for T. The disclosure risk can be analyzed through the inverse map $T=F^{-1}(T')$: if not available or difficult to compute by any *data attacker*, then we may guarantee that F is safe. Controlled Tabular Adjustment (CTA) [3,11] is a recent technique for the protection of any tabular data. It was empirically observed in [6] that estimates $\hat{T}=\hat{F}^{-1}(T')$, with $\hat{F}^{-1}$ being an estimate of F^{-1} for CTA, were not close to T for some real tables. CTA can thus

* Corresponding author.
E-mail addresses: daniel.baena@upc.edu (D. Baena), jordi.castro@upc.edu (J. Castro), jose.a.gonzalez@upc.edu (J.A. González).

http://dx.doi.org/10.1016/j.cor.2014.11.018
0305-0548/

a

		t_1	t_2	
⋮	...	...	...	...
51–55	...	38000 €	40000 €	...
56–60	...	39000 €	42000 €	...
⋮	...	...	...	...

b

		t_1	t_2	
⋮	...	...	...	...
51–55	...	20	1 or 2	...
56–60	...	30	35	...
⋮	...	...	...	...

Fig. 1. Example of disclosure in tabular data. (a) Average salary per age and town. (b) Number of individuals per age and town. If there is only one individual in town t_2 and age interval 51–55, then any external attacker knows the salary of this single person is 40,000€. For two individuals, any of them can deduce the salary of the other, becoming an internal attacker.

be considered a safe method in general. Moreover, the quality of CTA solutions has shown to be high [10], higher than that provided by alternative methods in some real instances [9].

The goal of CTA—which will be formulated in Section 2—is, given a table with any structure, to find the closest *safe* table to the original one. This is achieved by adding the minimum amount of deviations (or perturbations) to the original cell tables that makes the released table safe. Safety is guaranteed by imposing that sensitive cells in the new protected table are far enough from the original value. This means the cell value is either above *or* below some certain values, thus a disjunctive constraint involving a binary variable is needed for each sensitive cell. The minimum amount of above or below perturbations required for each sensitive cell is named, respectively, *upper protection* and *lower protection* levels. Changes in sensitive cells force other changes in the remaining cells to guarantee that the value of total or marginal cells is preserved. Although it is a recent approach, CTA is gaining recognition among NSAs; for instance, CTA is considered a relatively new emerging method in the recent monographs [26,27]. We recently implemented a package for CTA in collaboration with the NSAs of Germany and the Netherlands, within a project funded by Eurostat, the Statistical Office of the European Communities. This package has been largely improved within the FP7-INFRA-2010-262608 project funded by the European Union, with the participation, among others, of the national statistical institutes of Germany, Netherlands, Finland, Sweden and Slovenia. This CTA software is included in the tau-Argus package [25] (http://neon.vb.cbs.nl/casc/tau.htm), used for many European national statistical institutes for the protection of tabular data. Among the recent literature on CTA variants we find [8,24]. In recent specialized workshops on statistical disclosure control, some NSAs stated that perturbative methods, like CTA, are gaining acceptance [31], and perturbative approaches are being used for the protection of national census tables (e.g., [21] for Germany). CTA has also been used within other wider protection schemes, such as the pre-tabular protection method of [20]. In addition, some National Statistical Agencies are questioning current non-perturbative protection methods because "the task of balancing confidentiality and usability […] is nearly impossible" [30]. Therefore there is a need for new methods, and this justifies the research on CTA and other approaches. Indeed, there is no actually any protection method that fits the needs of all NSAs in the world.

From a computational point of view, the size of the CTA optimization problem is by far smaller than for other well-known protection methods, such as the cell suppression problem [4,19]. Despite these nice features, CTA formulates a challenging mixed integer linear problem (MILP) for current state-of-the-art solvers (such as CPLEX or XPress). Optimal (or suboptimal, e.g., with a 5% gap) solutions may require many hours of execution for medium instances; very large or massive tables cannot be tackled with current technology. Several approaches have been tried to speed up the solution time. A straightforward Benders reformulation of the problem was attempted in [7], but promising results were only obtained for two-dimensional tables (i.e., tables obtained by crossing two categorical variables, whose constraints are represented by a node-arc network incidence matrix [5]). Heuristic and metaheuristic methods were attempted in [22], but they only solved small two-dimensional and three-dimensional tables of up to 625 and 8000 cells, respectively, while we consider in this work much more complex synthetic and real tables, from the literature, of up to 200,000 and 36,000 cells, respectively. For instance, we generated a set of 20 two-dimensional and 20 three-dimensional tables with the same characteristics (sizes and number of sensitive cells) than those in [22]. We remark that (1) the tables used in [22] were also randomly generated; (2) the matrix constraints only depend on the table structure (two- or three-dimensional table) so they were the same in our experiments and those in [22]; (3) although the instances are not *exactly* the same, what makes difficult (in general) a problem is the structure of the matrix constraints and the number of sensitive cells (which is associated with the number of binary variables of the optimization problem); those characteristics are the same in our experiments and those of [22]. CPLEX 12.5 found a 0% gap solution for all these two-dimensional tables with an average CPU time of 0.02 s (the maximum time required by an instance was 0.03 s). For the three-dimensional tables, the average CPU time was 0.2 s (the maximum time for an instance was 0.49 s), again for 0% gap solutions. No CPU time comparison with CPLEX was reported in [22]; it was just stated that CPLEX 8.1 could not solve the instances. Therefore, up to now, there is no conclusive evidence that those metaheuristics are helpful for the CTA problem.

We also tried in the past other general metaheuristics as genetic algorithms without success: combinations or modifications of solutions are not expected to satisfy the large number of linear constraints with no particular structure of CTA. Indeed, these constraints are usually complex, and any practical approach must rely on the efficient solution of (usually difficult) linearly constrained problems (either LPs or MILPs). The approaches in this paper rely on decomposing the problem into smaller, thus tractable, MILP instances. It is worth to note that even the LPs obtained from large CTA instances by fixing the binary variables are very difficult for today state-of-the-art solvers. Indeed, some of these instances have been included in standard LP repositories [29].

The purpose of this work is twofold. Its first goal is to apply a fix-and-relax (FR) heuristic [13] to the MILP CTA problem. Briefly, FR partitions the set of binary variables into k clusters, and iteratively optimizes for each cluster $i = 1, \ldots, k$, fixing the binary variables of clusters $j < i$ at the optimal value found in previous iterations, and relaxing the integrality of binary variables of clusters $j > i$. The effect of this partitioning of the set of binary variables is that the nodes of the branch-and-cut tree are selectively explored. Equipping this procedure with a backward repartition strategy (details will be given in Section 3.1), if the MILP is feasible then FR will always provide a feasible, hopefully good and efficient, suboptimal solution. The approach cannot guarantee the optimal solution, but in practice