Working Paper
ENGLISH ONLY

UNITED NATIONS ECONOMIC COMMISSION FOR EUROPE (UNECE)

CONFERENCE OF EUROPEAN STATISTICIANS

EUROPEAN COMMISSION

STATISTICAL OFFICE OF THE EUROPEAN UNION (EUROSTAT)

Joint UNECE/Eurostat work session on statistical data confidentiality
(Ottawa, Canada, 28-30 October 2013)

Topic (iv): The trade-off between quality, utility and privacy

Assessing the Disclosure Risk of CTA-like Methods

Prepared by Jordi Castro, Universitat Politècnica de Catalunya, Catalonia, Spain

Assessing the disclosure risk of CTA-like methods [1]

Jordi Castro

Department of Statistics and Operations Research, Universitat Politècnica de Catalunya Jordi Girona 1–3, 08034 Barcelona, Catalonia

(jordi.castro@upc.edu)

Abstract. Minimum distance controlled tabular adjustment (CTA) is a recent perturbative approach for statistical disclosure control in tabular data. CTA looks for the closest safe table, using some particular distance. In this talk we provide empirical results to assess the disclosure risk of the method. A set of 33 instances from the literature and four different attacker scenarios are considered. The results show that, unless the attacker has good information about the original table, CTA has low disclosure risk. This talk summarizes results reported in the paper "Castro, J. (2013). On assessing the disclosure risk of controlled adjustment methods for statistical tabular data, *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 20, 921–941."

1 Introduction

Any 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' which is safe and, ideally, with minimum information loss. The inverse map $T = F^{-1}(T')$ should not be available or difficult to compute by any attacker, otherwise the disclosure risk would be high.

CTA (Dandekar and Cox, 2002; Castro, 2006, 2011) is a post-tabular approach which looks for the closest safe table to the original unsafe one. CTA relies on optimization methods, mainly mixed integer linear programming (MILP), and linear programming (LP). This offers a great flexibility when some table properties want to be preserved in the released table, expressed as linear constraints. CTA is one of the methods discussed in the recent monograph Hundepool et al. (2012).

CTA-like methods will have low disclosure risk if no attacker can obtain a good estimate $\hat{T} = \hat{F}^{-1}(T')$, $\hat{F}^{-1}$ being an estimate of F^{-1}. The goodness of $\hat{F}^{-1}$ depends on the amount of information by the attacker. In this talk we will consider four different attacker scenarios —each one associated to a particular $\hat{F}^{-1}$—, providing an exhaustive empirical evaluation of the disclosure risk of these approaches which

[1]This work has been supported by grants MTM2012-31440 of the Spanish research program and SGR-2009-1122 of the Government of Catalonia.

required the solution of more than 2500 optimization attacker problems. As it will be shown, when the attacker has no good information about the original data, the disclosure risk is low. As expected, the computational results also confirmed that the more information by the attacker, the higher is the disclosure risk.

It is worth noting that some authors claimed that protection approaches based on the minimization of information loss are not safe if a *minimality attack* is performed (Chi-Wing et all., 2007). However, minimality attacks have been used for microdata, not for tabular data (e.g., the term table was used for "table in a relational database" not for "statistical table") in Chi-Wing et all. (2007). One of the purposes of this talk is to empirically show that the above assertion can not be generalized, and, depending on the particular attacker problem $\hat{F}^{-1}$, CTA is safe.

This short document summarizes some of the results presented in Castro (2012). Its structure is as follows. The CTA problem will be outlined in Section 2. The different attacker scenarios considered are discussed in Section 3. Finally, computational results are provided in Section 4

2 Outline of minimum distance MILP-CTA

Any CTA instance, either with one table or a number of tables, can be represented by the following parameters:

- A set of cells $a_i, i \in \mathcal{N} = \{1, \ldots, n\}$, that satisfy m linear relations $Aa = b$ (a being the vector of a_i's), and a vector $w \in \mathbb{R}^n$ of positive weights for the deviations of cell values.
- A lower and upper bound for each cell $i \in \mathcal{N}$, respectively l_{a_i} and u_{a_i}, which are considered to be known by any attacker. If no previous knowledge is assumed for cell i $l_{a_i} = 0$ ($l_{a_i} = -\infty$ if $a \geq 0$ is not required) and $u_{a_i} = +\infty$ can be used.
- A set $\mathcal{S} = \{i_1, i_2, \ldots, i_s\} \subseteq \mathcal{N}$ of indices of s confidential cells.
- A lower and upper protection level for each confidential cell $i \in \mathcal{S}$, respectively lpl_i and upl_i, such that the released values must satisfy either $x_i \geq a_i + upl_i$ or $x_i \leq a_i - lpl_i$.

CTA attempts to find the closest values $x_i, i \in \mathcal{N}$, according to some distance ℓ, that makes the released table safe. This involves the solution of the following optimization problem:

$$\begin{array}{ll} \min\limits_{x} & ||x - a||_{\ell(w)} \\ \text{s. to} & Ax = b \\ & l_{a_i} \leq x_i \leq u_{a_i} \quad i \in \mathcal{N} \\ & (x_i \leq a_i - lpl_i) \text{ or } (x_i \geq a_i + upl_i) \quad i \in \mathcal{S}. \end{array} \tag{1}$$

The CTA problem (1) is in general a difficult MILP.

An alternative would be to a priori fix the binary variables $y_i, i \in \mathcal{S}$, thus obtaining a CTA formulation with only continuous variables; the resulting problem would be a LP. Although the information loss of this LP-CTA variant is higher, it can be solved much more efficiently. In the computational experiments performed we considered this LP-CTA approach. It is worth noting that if this variant is shown to be "safe", the problem with binary variables would also be "safe" (even "safer"), since in the former case the decision on the particular value of y_i is governed by a combinatorial optimization procedure. Formulating problem (1) in terms of cell deviations $z = x - a$, $z \in \mathbb{R}^n$, and fixing the binary variables, the resulting continuous CTA problem can be formulated as the general convex optimization problem

$$\begin{array}{ll} \min\limits_{z} & ||z||_{\ell(w)} \\ \text{s. to} & Az = 0 \\ & l(a) \leq z \leq u(a), \end{array} \tag{2}$$

where bounds $l(a)$ and $u(a)$ depend on cell bounds l_a, u_a and protection levels lpl, upl.

Problem (2) can be specialized for several norms, ℓ_1 and ℓ_2 being the two most relevant. For ℓ_1, defining $z = z^+ - z^-$, we obtain the following LP:

$$\begin{array}{ll} \min\limits_{z^+, z^-} & \sum\limits_{i=1}^{n} w_i(a_i)(z_i^+ + z_i^-) \\ \text{s. to} & A(z^+ - z^-) = 0 \\ & l^+(a) \leq z^+ \leq u^+(a) \\ & l^-(a) \leq z^- \leq u^-(a), \end{array} \tag{3}$$

$w(a) \in \mathbb{R}^n$ being a vector of nonnegative cell weights, $z^+ \in \mathbb{R}^n$ and $z^- \in \mathbb{R}^n$ the vector of positive and negative deviations in absolute value, and $l^+(a), l^-(a), u^+(a), u^-(a) \in \mathbb{R}^n$ lower and upper bounds for the positive and negative deviations. For L_2, problem (2) can be directly recast as the following quadratic optimization problem (QP) without introducing additional variables:

$$\begin{array}{ll} \min\limits_{z} & \sum\limits_{i=1}^{n} w_i(a_i) z_i^2 \\ \text{s. to} & Az = 0 \\ & l(a) \leq z \leq u(a). \end{array} \tag{4}$$

3 The attacker scenarios considered

The goal of the attacker is to obtain a *good* estimate $\hat{z}$ of z from the released table T'. In this context, a *good* estimate may be either to obtain the original value z_i for some sensitive cell, or —the weaker condition— a value not too far from z_i. In practice, once the table is published, the attacker only knows

- the released values x;
- the structure of the table, that is, the constraint matrix A.

For the rest of parameters the attacker may only have partial information:

- the particular distance used may be unknown, that is, which of the two problems were solved by the data protector, either (3) or (4); however, providing information about the distance used may be seen as a good practice, so we considered it is known by the attacker;
- cell weights $w(a)$ are unknown, since they depend on the original data;
- the lower and upper bounds ($l^+(a), l^-(a), u^+(a), u^-(a)$ in (3), $u(a)$, $l(a)$ in (4)) are unknown because: (i) they depend on a; (ii) the set of sensitive cells $\mathcal{S}$ is unknown to the attacker; (iii) the a priori assignment of y_i will also be unknown to the attacker.

In general, the optimization problem to be solved by the attacker can be written as:

$$\begin{array}{ll} \min\limits_{\hat{z}} & ||\hat{z}||_{\ell(x)} \\ \text{s. to} & A\hat{z} = 0 \\ & \hat{l}(x) \leq \hat{z} \leq \hat{u}(x). \end{array} \tag{5}$$

We will consider the following four different scenarios according to the knowledge of the attacker for the solution of (5):

B. The attacker has incomplete information about both the bounds and objective function, but he/she knows the subset $\mathcal{S}$ of sensitive cells, and the original cell bounds l_{a_i} and u_{a_i}, $i \in \mathcal{N}$ (which are quite strong assumptions). We have three subscenarios:

   B1. The attacker neither knows the protection levels $upl_i, lpl_i, i \in \mathcal{S}$, nor the protection sense $y_i \in \{0, 1\}, i \in \mathcal{S}$.

   B2. The attacker knows the protection sense $y_i \in \{0, 1\}, i \in \mathcal{S}$, but not the protection levels upl_i, lpl_i, $i \in \mathcal{S}$.

   B3. The attacker knows both the protection sense $y_i \in \{0, 1\}$ and protection levels upl_i, lpl_i, $i \in \mathcal{S}$. The only unknown terms to reproduce the real bounds are then $a_i - l_{a_i}$ and $u_{a_i} - a_i$, $i \in \mathcal{N}$.

C. The attacker has complete information about the bounds, i.e, he/she knows all the parameters for the definition of (5), and the only uncertainty is in the use of $w_i(x_i)$ instead of $w_i(a_i)$ in the objective function. This is a very strong assumption, since it means the attacker knows or has accurate information about the original cell values a.

Table 1: Dimensions of the test instances.

instance	n	s	m	nz
australia_ABS	24420	918	274	13224
bts4	36570	2260	36310	136912
cbs	11163	2467	244	22326
dale	16514	4923	405	33028
destatis	5940	621	1464	18180
hier13	2020	112	3313	11929
hier13x13x13a	2197	108	3549	11661
hier13x13x13b	2197	108	3549	11661
hier13x13x13c	2197	108	3549	11661
hier13x13x13d	2197	108	3549	11661
hier13x13x13e	2197	112	3549	11661
hier13x13x7d	1183	75	1443	5369
hier13x7x7d	637	50	525	2401
hier16	3564	224	5484	19996
hier16x16x16a	4096	224	5376	21504
hier16x16x16b	4096	224	5376	21504
hier16x16x16c	4096	224	5376	21504
hier16x16x16d	4096	224	5376	21504
hier16x16x16e	4096	224	5376	21504
nine12	10399	1178	11362	52624
nine5d	10733	1661	17295	58135
ninenew	6546	858	7340	32920
osorio	10201	7	202	20402
table1	1584	146	510	4752
table3	4992	517	2464	19968
table4	4992	517	2464	19968
table5	4992	517	2464	19968
table6	1584	146	510	4752
table7	624	17	230	1872
table8	1271	3	72	2542
targus	162	13	63	360
toy3dsarah	2890	376	1649	9690
two5in6	5681	720	9629	34310

4 Computational results

For the empirical evaluation we have considered a set of both real and synthetic 25 instances widely used in the literature about statistical data protection (Castro, 2006, 2012). Table 1 shows the main dimensions of these tables: number of cells (n), number or sensitive cells (s), number of tabular constraints (m), and number of nonzero coefficients in the matrix of tabular constraints ("nz").

We first protected the tables using both ℓ_1-CTA and ℓ_2-CTA, to obtain the released values $x = a + z$. Next, we solved the attacker problems for the four different scenarios : B1, B2, B3 and C. For each of the 264 different combinations (33 instances $\times$ 2 distances $\times$ 4 scenarios) we considered ten realizations of the attacker problems for different $\tilde{x}$ values, randomly obtained in an interval around x. This amounts to 2640 optimization attacker problems. From the solution of these problems we computed for each sensitive cell the ten percentage differences between a and $\hat{a}$, the true cell values and the ten attacker estimations, i.e., $|\hat{a}_i - a_i|/a_i \cdot 100$,

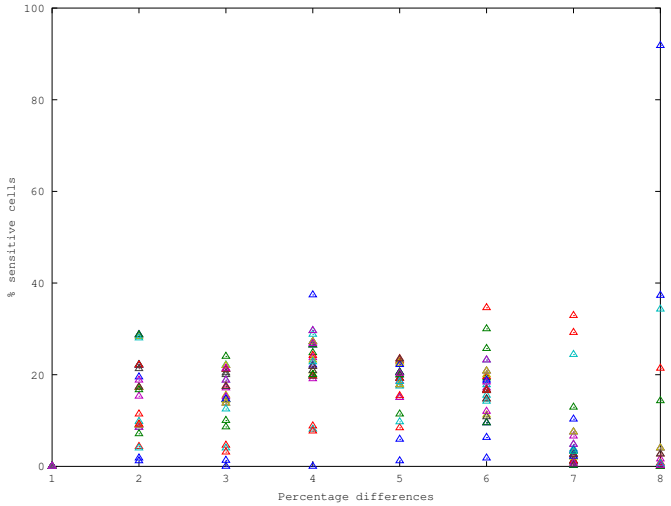

Figure 1: Results for scenario B1 and norm ℓ_1.

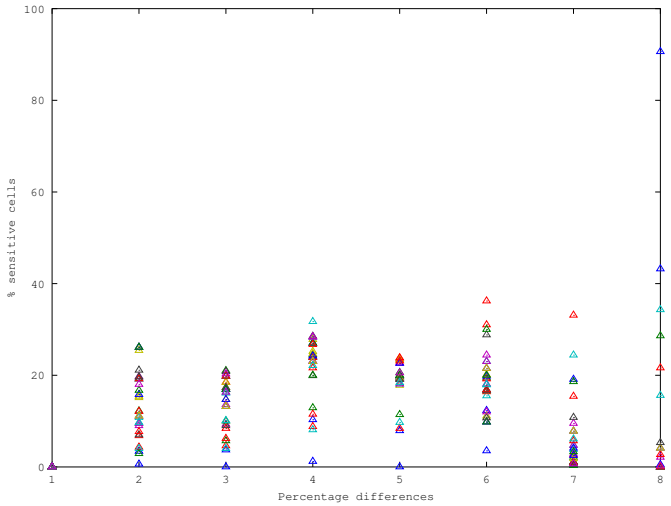

Figure 2: Results for scenario B1 and norm ℓ_2.

$i \in \mathcal{S}$.

Figures 1–8 show the distribution of the percentage differences between $\hat{a}$ and a for all the instances. The eight values of the x-axis are associated to the following intervals for $|\hat{a}_i - a_i|/a_i \cdot 100$: 0, $(0, 5]$, $(5, 10]$, $(10, 20]$, $(20, 30]$, $(30, 50]$, $(50, 100]$ and $(100, -)$. The y-axis is related to the percentage of sensitive cells. Detailed tables with information for each instance can be found in Castro (2012). The following conclusions can be derived from Figures 1–8:

- Scenarios B1 and B2 can be considered safe, in general. The estimate $\hat{a}_i$ was

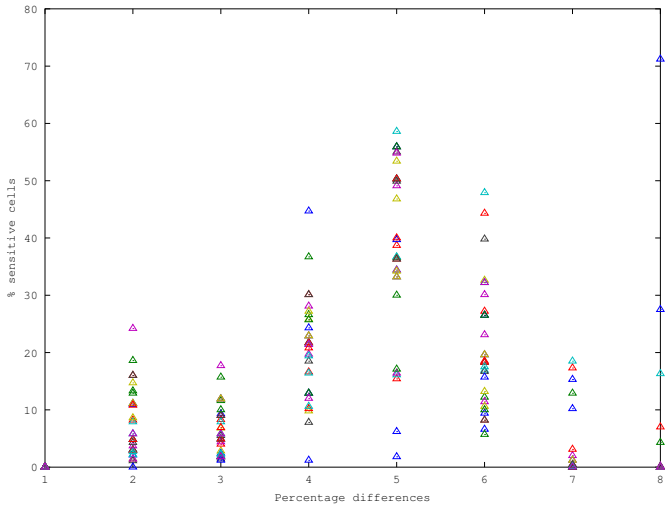

Figure 3: Results for scenario B2 and norm ℓ_1.

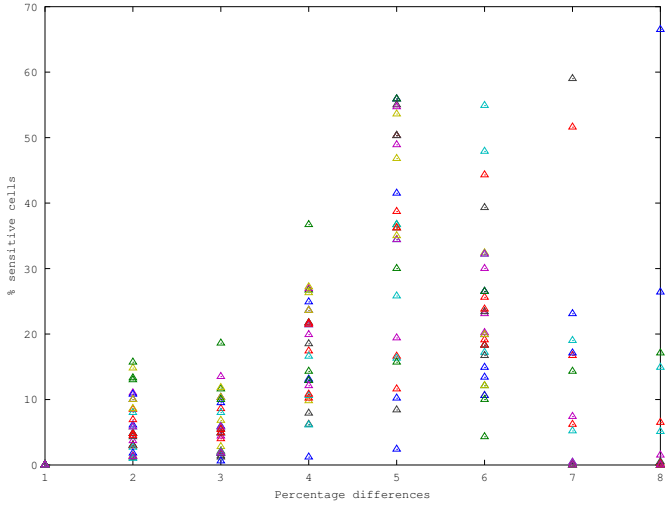

Figure 4: Results for scenario B2 and norm ℓ_2.

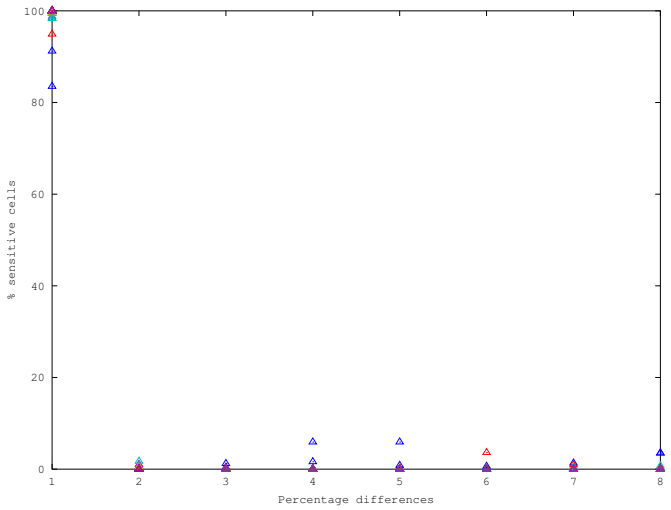

Figure 5: Results for scenario B3 and norm ℓ_1.

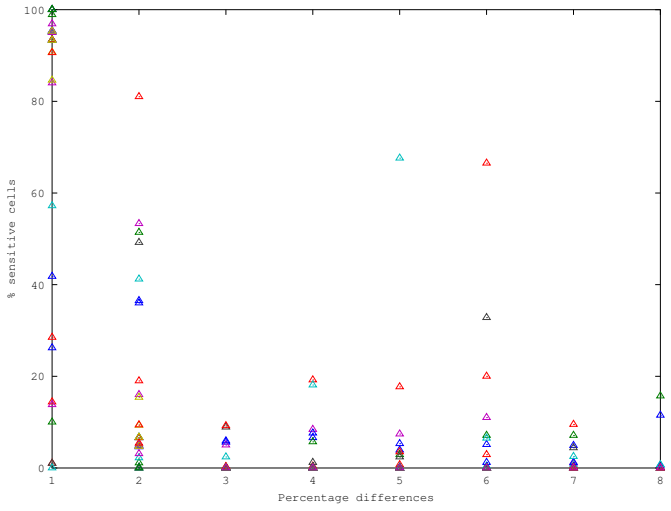

Figure 6: Results for scenario B3 and norm ℓ_2.

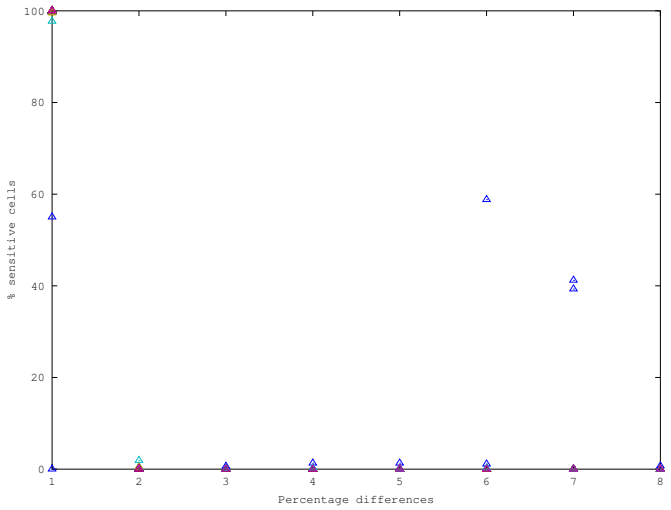

Figure 7: Results for scenario C and norm ℓ_1.

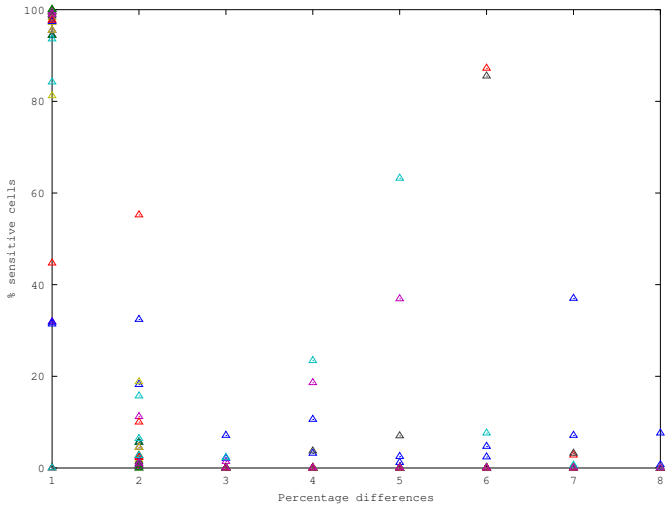

Figure 8: Results for scenario C and norm ℓ_2.

never equal to the true cell value a_i, and the distribution is not concentrated on the left intervals.

- Comparing L_1 and L_2, the latter seems to reduce the disclosure risk: the distribution is more left-skewed for L_2 in scenarios B1 and B2.

- For scenarios B3 and C the attacker was able to re-compute in almost 100% of the cases the original values a. If the attacker has good information about the protection levels, protection senses, set of sensitive cells, and lower and upper bounds, then CTA-like methods exhibit a high disclosure risk. However the knowledge of such big amount of information by the attacker may be a strong assumption.

References

Castro, J. (2006). Minimum-distance controlled perturbation methods for large-scale tabular data protection, *European Journal of Operational Research*, 171, 39–52.

Castro, J. (2011). Recent advances in optimization techniques for statistical tabular data protection, *European Journal of Operational Research*, 216, 257–269.

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

Chi-Wing Wong, R., Wai-Chee Fu, A., Wang, K., and Pei, J. (2007). Minimality attack in privacy preserving data publishing *Proc. 33rd International Conference on Very Large Data Bases*, Vienna, Austria, 553–554.

Dandekar, R.A., and Cox, L.H. (2002). Synthetic tabular data: an alternative to complementary cell suppression, manuscript, Energy Information Administration, U.S. Department of Energy.

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