Josep Domingo-Ferrer
Yücel Saygın (Eds.)

LNCS 5262

Privacy in Statistical Databases

UNESCO Chair in Data Privacy
International Conference, PSD 2008
Istanbul, Turkey, September 2008, Proceedings

Springer

Lecture Notes in Computer Science 5262

Commenced Publication in 1973
Founding and Former Series Editors:
Gerhard Goos, Juris Hartmanis, and Jan van Leeuwen

Josep Domingo-Ferrer Yücel Saygın (Eds.)

Privacy in Statistical Databases

UNESCO Chair in Data Privacy
International Conference, PSD 2008
Istanbul, Turkey, September 24-26, 2008
Proceedings

Springer

Volume Editors

Josep Domingo-Ferrer
Rovira i Virgili University, Department of Computer Engineering and Mathematics
Av. Països Catalans 26, 43007 Tarragona, Catalonia, Spain
E-mail: josep.domingo@urv.cat

Yücel Saygın
Sabancı University, Faculty of Engineering and Natural Sciences
Orhanli, 34956 Tuzla, Istanbul, Turkey
E-mail: ysaygin@sabanciuniv.edu

Library of Congress Control Number: 2008934858

CR Subject Classification (1998): D.4.6, G.3, H.2.8, K.4.1

LNCS Sublibrary: SL 3 – Information Systems and Application, incl. Internet/Web and HCI

ISSN 0302-9743
ISBN-10 3-540-87470-4 Springer Berlin Heidelberg New York
ISBN-13 978-3-540-87470-6 Springer Berlin Heidelberg New York

Springer is a part of Springer Science+Business Media

springer.com

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India
Printed on acid-free paper SPIN: 12521944 06/3180 5 4 3 2 1 0

Preface

Privacy in statistical databases is a discipline whose purpose is to provide solutions to the tension between the increasing social, political and economical demand of accurate information, and the legal and ethical obligation to protect the privacy of the various parties involved. Those parties are the respondents (the individuals and enterprises to which the database records refer), the data owners (those organizations spending money in data collection) and the users (the ones querying the database, who would like their queries to stay confidential). Beyond law and ethics, there are also practical reasons for data collecting agencies to invest in respondent privacy: if individual respondents feel their privacy guaranteed, they are likely to provide more accurate responses. Data owner privacy is primarily motivated by practical considerations: if an enterprise collects data at its own expense, it may wish to minimize leakage of those data to other enterprises (even to those with whom joint data exploitation is planned). Finally, user privacy results in increased user satisfaction, even if it may curtail the ability of the database owner to profile users.

There are at least two traditions in statistical database privacy, both of which started in the 1970s: one stems from official statistics, where the discipline is also known as statistical disclosure control (SDC), and the other originates from computer science and database technology. In official statistics, the basic concern is respondent privacy. In computer science, one started with respondent privacy but, from 2000 onwards, growing attention has been devoted to owner privacy (privacy-preserving data mining) and user privacy (private information retrieval). In the last few years, the interest and the achievements of computer scientists in the topic have substantially increased.

Privacy in Statistical Databases 2008 (PSD 2008) was held under the sponsorship of the UNESCO Chair in Data Privacy, which intends to act as a stable umbrella for the PSD biennial conference series from now on. PSD 2008 was a successor of PSD 2006, the final conference of the Eurostat-funded CENEX-SDC project, held in Rome in 2006, and PSD 2004, the final conference of the EU FP5 CASC project (IST-2000-25069), held in Barcelona in 2004. Proceedings of PSD 2006 and PSD 2004 were published by Springer in LNCS 4302 and LNCS 3050, respectively. The three PSD conferences held so far are a follow-up of a series of high-quality technical conferences on SDC which started one decade ago with “Statistical Data Protection-SDP 1998”, held in Lisbon in 1998 and with proceedings published by OPOCE, and continued with the AMRADS project SDC Workshop, held in Luxemburg in 2001 and with proceedings published by Springer in LNCS 2316.

For PSD 2008, the Program Committee accepted 27 papers out of 37 submissions from 16 different countries in five different continents. Each submitted paper received at least two reviews. These proceedings contain the revised versions of the accepted papers, which are a fine blend of contributions from the areas of official statistics and computer science. Covered topics include tabular data protection, methods and case studies for microdata protection, disclosure risk assessment in microdata protection, on-line databases and remote access, privacy-preserving data mining, private information retrieval and legal issues.

We are indebted to many people. First, to the Government of Catalonia for financial support of the UNESCO Chair in Data Privacy, which enabled the latter to sponsor PSD 2008. Also, to the Organizing Committee for making the conference possible and especially to Jesús Manjón, who helped with these proceedings. In evaluating the papers we were assisted by the Program Committee and the following external reviewers: Javier Herranz, Marlow Lemons, Thomas B. Pedersen, Adam Persing and Elizabeth Ransom.

We also wish to thank all the authors of submitted papers and apologize for possible omissions.

July 2008

Josep Domingo-Ferrer
Yücel Saygın

Organization

Program Committee

John Abowd	Cornell University, USA
Elisa Bertino	CERIAS, Purdue University, USA
Jordi Castro	Polytechnical University of Catalonia, Spain
Lawrence Cox	Nat. Center for Health Statistics, USA
Josep Domingo-Ferrer	Rovira i Virgili University, Catalonia, Spain
Mark Elliot	Manchester University, UK
Elena Ferrari	University of Insubria, Italy
Stephen Fienberg	Carnegie Mellon University, USA
Luisa Franconi	ISTAT, Italy
Sarah Giessing	Destatis, Germany
Anco Hundepool	Statistics Netherlands, The Netherlands
Ramayya Krishnan	Carnegie Mellon University, USA
Julia Lane	NORC/University of Chicago, USA
Jane Longhurst	Office for National Statistics, UK
Bradley Malin	Vanderbilt University, USA
Josep M. Mateo-Sanz	Rovira i Virgili University, Catalonia, Spain
Krish Muralidhar	University of Kentucky, USA
Jean-Marc Museux	EUROSTAT, European Union
Silvia Polettini	University of Naples, Italy
Yosef Rinott	Hebrew University, Israel
Gerd Ronning	University of Tübingen, Germany
Juan José Salazar	University of La Laguna, Spain
Yücel Saygın	Sabancı University, Turkey
Eric Schulte-Nordholt	Statistics Netherlands, The Netherlands
Francesc Sebé	Rovira i Virgili University, Catalonia, Spain
Natalie Shlomo	University of Southampton, UK
Julian Stander	University of Plymouth, UK
Vicenç Torra	IIIA-CSIC, Catalonia, Spain
William E. Winkler	Census Bureau, USA
Laura Zayatz	Census Bureau, USA

Program Chair

Josep Domingo-Ferrer	Rovira i Virgili University, Catalonia, Spain

General Chair

Yücel Saygın	Sabancı University, Turkey

Organizing Committee

Jordi Castellà-Roca	Rovira i Virgili University, Catalonia, Spain
Aysegul Cayci	Sabancı University, Turkey
Ercument Cicek	Sabancı University, Turkey
Aras Sami Kubilay	Sabancı University, Turkey
Jesús Manjón	Rovira i Virgili University, Catalonia, Spain
Antoni Martínez-Ballesté	Rovira i Virgili University, Catalonia, Spain
Glòria Pujol	Rovira i Virgili University, Catalonia, Spain

Table of Contents

Tabular Data Protection

Microdata Protection: Methods and Case Studies

Microdata Protection: Disclosure Risk Assessment

On-Line Databases and Remote Access

Privacy-Preserving Data Mining and Private Information Retrieval

Legal Issues

Using a Mathematical Programming Modeling Language for Optimal CTA

Jordi Castro[1,⋆,⋆⋆] and Daniel Baena[2]

[1] Department of Statistics and Operations Research,
Universitat Politècnica de Catalunya,
Jordi Girona 1–3, 08034 Barcelona, Catalonia
jordi.castro@upc.edu
http://www-eio.upc.es/~jcastro
[2] Institut d'Estadística de Catalunya,
Via Laietana 58, 08003 Barcelona, Catalonia
dbaena@idescat.net

Abstract. Minimum-distance controlled tabular adjustment methods (CTA) have been formulated as an alternative to the cell suppression problem (CSP) for tabular data. CTA formulates an optimization problem with fewer variables and constraints than CSP. However, the inclusion of binary decisions about protection sense of sensitive cells (optimal CTA) in the formulation, still results in a mixed integer-linear problem. This work shows how mathematical programming modeling languages can be used to develop a prototype for optimal CTA based on Benders method. Preliminary results are reported for some medium size two-dimensional tables. For this type of tables, the approach is competitive with other general-purpose algorithms implemented in commercial solvers.

Keywords: statistical disclosure control, controlled tabular adjustment, mixed-integer linear programming, Benders decomposition.

1 Introduction

Minimum-distance controlled tabular adjustment methods (CTA) were suggested in [2,7] as an alternative to the difficult cell suppression problem (CSP) [3,8]. In some instances, the quality of CTA solutions has shown to be higher than that provided by CSP ones [4].

Although CTA formulates an optimization problem with fewer variables and constraints than CSP, it is also a mixed integer-linear problem (MILP) if the binary decision of protection sense of sensitive cells is included in the model (optimal CTA). Therefore, for some instances, the solution time of optimal CTA by a general purpose solver, like CPLEX or XPress, can still be large. (Some

[⋆] Supported by grant MTM2006-05550 of the Spanish Ministry of Science and Education.

[⋆⋆] Corresponding author.

J. Domingo-Ferrer and Y. Saygın (Eds.): PSD 2008, LNCS 5262, pp. 1–12, 2008.

metaheuristics approaches have been used, but only for small-medium instances [6].) For MILP models there are some specialized algorithms. One of them is Benders method [1]. In this work we show how a mathematical programming modeling language can be used for a prototype for optimal CTA based on Benders decomposition. Preliminary results with this prototype are reported, using a battery of two-dimensional tables. For these instances, the algorithm is more efficient that the general purpose solver implemented in CPLEX.

The paper is organized as follows. Section 2 reviews the CTA method. Section 3 outlines the Benders decomposition algorithm for the non mathematical programming experts. Section 4 shows how this approach can be implemented in the AMPL mathematical programming language. Section 5 illustrates the approach in the solution of a small example. Finally, Section 6 reports computational results in the solution of a set of two-dimensional tables.

2 The Optimal CTA Problem

Given (i) a set of cells $a_i, i = 1, \ldots, n$, that satisfy some linear relations $Aa = b$ (a being the vector of a_i's); (ii) a lower and upper bound for each cell $i = 1, \ldots, n$, respectively l_{a_i} and u_{a_i}, which are considered to be known by any attacker; (iii) a set $\mathcal{P} = \{i_1, i_2, \ldots, i_p\} \subseteq \{1, \ldots, n\}$ of indices of sensitive cells; (iv) and a lower and upper protection level for each sensitive cell $i \in \mathcal{P}$, respectively lpl_i and upl_i, such that the released values satisfy either $x_i \geq a_i + upl_i$ or $x_i \leq a_i - lpl_i$; the purpose of CTA is to find the closest safe values $x_i, i = 1, \ldots, n$, according to some distance L, that makes the released table safe. This involves the solution of the following optimization problem:

$$\begin{array}{ll} \min\limits_{x} & ||x-a||_L \\ \text{s. to} & Ax = b \\ & l_{a_i} \leq x_i \leq u_{a_i} \quad i = 1, \ldots, n \\ & x_i \leq a_i - lpl_i \text{ or } x_i \geq a_i + upl_i \quad i \in \mathcal{P}. \end{array} \tag{1}$$

Problem (1) can also be formulated in terms of deviations from the current cell values. Defining $z_i = x_i - a_i, \quad i = 1, \ldots, n$ —and similarly $l_{z_i} = l_{x_i} - a_i$ and $u_{z_i} = u_{x_i} - a_i$—, (1) can be recast as:

$$\begin{array}{ll} \min\limits_{z} & ||z||_L \\ \text{s. to} & Az = 0 \\ & l_{z_i} \leq z_i \leq u_{z_i} \quad i = 1, \ldots, n \\ & z_i \leq -lpl_i \text{ or } z_i \geq upl_i \quad i \in \mathcal{P}, \end{array} \tag{2}$$

$z \in \mathbb{R}^n$ being the vector of deviations. Using the L_1 distance, and after some manipulation, (2) can be written as

$$\begin{array}{rl} \min\limits_{z^+,z^-,y} & \sum\limits_{i=1}^{n} w_i(z_i^+ + z_i^-) \\ \text{s. to} & A(z^+ - z^-) = 0 \\ & 0 \leq z_i^+ \leq \quad u_{z_i} \quad i \notin \mathcal{P} \\ & 0 \leq z_i^- \leq -l_{z_i} \quad i \notin \mathcal{P} \\ & upl_i \; y_i \leq z_i^+ \leq u_{zi} \; y_i \quad i \in \mathcal{P} \\ & lpl_i(1-y_i) \leq z_i^- \leq -l_{zi}(1-y_i) \quad i \in \mathcal{P}, \end{array} \tag{3}$$

$w \in \mathbb{R}^n$ being the vector of cell weights, $z^+ \in \mathbb{R}^n$ and $z^- \in \mathbb{R}^n$ the vector of positive and negative deviations in absolute value, and $y \in \mathbb{R}^p$ being the vector of binary variables associated to protections senses. When $y_i = 1$ the constraints mean $upl_i \leq z_i^+ \leq u_{zi}$ and $z_i^- = 0$, thus the protection sense is "upper"; when $y_i = 0$ we get $z_i^+ = 0$ and $lpl_i \leq z_i^- \leq -l_{z_i}$, thus protection sense is "lower". Model (3) is a (difficult) MILP.

3 Outline of Benders Method for MILP Problems

Benders decomposition method [1] was suggested for problems with two types of variables, one of them considered as "complicating variables". In MILP models complicating variables are the binary/integer ones. Consider the following MILP primal problem (P) in variables (x, y)

$$(P) \qquad \begin{array}{ll} \min & c^T x + d^T y \\ \text{s. to} & A_1 x + A_2 y = b \\ & x \geq 0 \\ & y \in Y, \end{array}$$

where y are the binary/complicating variables, $c, x \in \mathbb{R}^{n_1}$, $d, y \in \mathbb{R}^{n_2}$, $A_1 \in \mathbb{R}^{m \times n_1}$ and $A_2 \in \mathbb{R}^{m \times n_2}$. For binary problems, as in optimal CTA, we have $Y = \{0,1\}^{n_2}$. Fixing some $y \in Y$, we obtain:

$$(Q) \qquad \begin{array}{ll} \min & c^T x \\ \text{s. to} & A_1 x = b - A_2 y \\ & x \geq 0. \end{array}$$

The dual of (Q) is:

$$(Q_D) \qquad \begin{array}{ll} \max & u^T(b - A_2 y) \\ \text{s. to} & A_1^T u \leq c \\ & u \in \mathbb{R}^m. \end{array}$$

It is known that if (Q_D) has a solution then (Q) has a solution too, and both objective functions coincide; if (Q_D) is unbounded, then (Q) is infeasible. Let assume that (Q_D) is never infeasible (indeed, this is the case in optimal CTA). If, as notation convention, we consider that the objective of (Q) is $+\infty$ when it is infeasible, then (P) can be written as

$$(P') \qquad \begin{array}{ll} \min & \left\{ d^T y + \max \left\{ u^T(b - A_2 y) | A_1^T u \leq c, u \in \mathbb{R}^m \right\} \right\} \\ \text{s. to} & y \in Y. \end{array}$$

Let $U = \left\{u | A_1^T u \leq c, u \in \mathbb{R}^m\right\}$ be the convex feasible set of (Q_D). By Minkowski representation we know that every point $u \in U$ may be represented as a convex combination of the vertices $u^1, \ldots, u^s$ and extreme rays $v^1, \ldots, v^t$ of the convex polytope U. Therefore any $u \in U$ may be written as

$$\begin{array}{c} u = \sum_{i=1}^{s} \lambda_i u^i + \sum_{j=1}^{t} \mu_j v^j \\ \sum_{i=1}^{s} \lambda_i = 1 \\ \lambda_i \geq 0 \quad i = 1, \ldots, s \\ \mu_j \geq 0 \quad j = 1, \ldots, t. \end{array}$$

If $v^{jT}(b - A_2 y) > 0$ for some $j \in \{1, \ldots, t\}$ then (Q_D) is unbounded, and thus (Q) is infeasible. We then impose

$$v^{jT}(b - A_2 y) \leq 0 \quad j = 1, \ldots, t.$$

The optimal solution of (Q_D) is then known to be in a vertex of U, and (P') may be rewritten as

$$(P'') \qquad \begin{array}{ll} \min & d^T y + \max\limits_{i=1,\ldots,s} (u^{iT}(b - A_2 y)) \\ \text{s. to} & v^{jT}(b - A_2 y) \leq 0 \quad j = 1, \ldots, t \\ & y \in Y. \end{array}$$

Introducing variable θ, (P'') is equivalent to the Benders problem (BP):

$$(BP) \qquad \begin{array}{lll} \min & \theta & \\ \text{s. to} & \theta \geq d^T y + u^{iT}(b - A_2 y) & i = 1, \ldots, s \\ & v^{jT}(b - A_2 y) \leq 0 & j = 1, \ldots, t \\ & y \in Y. & \end{array}$$

Problem (BP) is impractical since s and t can be very large, and in addition the vertices and extreme rays are unknown. Instead, the method considers a relaxation (BP_r) with a subset of the vertices and extreme rays. The relaxed Benders problem (or master problem) is thus:

$$(BP_r) \qquad \begin{array}{lll} \min & \theta & \\ \text{s. to} & \theta \geq d^T y + u^{iT}(b - A_2 y) & i \in I \subseteq \{1, \ldots, s\} \\ & v^{jT}(b - A_2 y) \leq 0 & j \in J \subseteq \{1, \ldots, t\} \\ & y \in Y. & \end{array}$$

Initially $I = J = \emptyset$, and new vertices and extreme rays provided by the subproblem (Q_D) are added to the master problem, until the optimal solution is found. In summary, the steps of the Benders algorithm are:

Benders Algorithm

0. Initially $I = \emptyset$ and $J = \emptyset$. Let (θ_r^*,y_r^*) be the solution of current master problem (BP_r), and (θ^*,y^*) the optimal solution of (BP).
1. Solve master problem (BP_r) obtaining θ_r^* and w_r^*. At first iteration, $\theta_r^* = -\infty$ and y_r is any feasible point in Y.

2. Solve subproblem (Q_D) using $y = y_r^*$. There are two cases:
   (a) (Q_D) has finite optimal solution in vertex u^{i0}.
      - If $\theta_r^* = d^T y_r^* + u^{i_0 T}(b - A_2 y_r^*)$ then **STOP**. Optimal solution is $y^* = y_r^*$ with cost $\theta^* = \theta_r^*$.
      - If $\theta_r^* < d^T y_r^* + u^{i_0 T}(b - A_2 y_r^*)$ then this solution violates constraint of (BP) $\theta > d^T y + u^{i_0 T}(b - A_2 y)$. Add this new constraint to (BP_r): $I \leftarrow I \cup \{i_0\}$.

   (b) (Q_D) is unbounded along segment $u^{i_0} + \lambda v^{j_0}$ (u^{i_0} is current vertex, v^{j_0} is extreme ray). Then this solution violates constraint of (BP) $v^{j_0 T}(b - A_2 w) \le 0$. Add this new constraint to (BP_r): $J \leftarrow J \cup \{j_0\}$; vertex may also be added: $I \leftarrow I \cup \{i_0\}$.
3. Go to step 1 above.

Convergence is guaranteed since at each iteration one or two constraints are added to (BP_r), no constraints are repeated, and the maximum number of constraints is $s + t$.

4 Prototype of Benders Method for Optimal CTA

It can be shown that, applying Benders method to the optimal CTA problem (3), the formulation subproblem (Q_D) is given by (see [5] for details):

$$\begin{array}{ll} \max\limits_{\mu_u^+,\mu_u^-,\mu_l^+,\mu_l^-} & -{\mu_u^+}^T u^+ - {\mu_u^-}^T u^- + {\mu_l^+}^T l^+ + {\mu_l^-}^T l^- \\ & \\ \text{s. to} & \begin{pmatrix} A^T \\ -A^T \end{pmatrix} \lambda - \begin{pmatrix} \mu_u^+ \\ \mu_u^- \end{pmatrix} + \begin{pmatrix} \mu_l^+ \\ \mu_l^- \end{pmatrix} = \begin{pmatrix} w \\ w \end{pmatrix} \\ & \mu_u^+, \mu_u^-, \mu_l^+, \mu_l^- \ge 0, \lambda \text{ free }, \end{array} \tag{4}$$

where $\mu_u^+, \mu_u^-, \mu_l^+, \mu_l^- \in \mathbb{R}^n$, $\lambda \in \mathbb{R}^m$, and l^+, l^-, u^+, u^- provide the lower and upper bounds of z^+ and z^- once binary variables $y \in \mathbb{R}^p$ are fixed.

Similarly, the formulation of the master (BP_r) is

$$\begin{array}{ll} \min\limits_{\theta,y} & \theta \\ \text{s. to} & \sum\limits_{h \notin \mathcal{P}} (-\mu_{u_h}^{+,i} u_{z_h} + \mu_{u_h}^{-,i} l_{z_h}) + \sum\limits_{h \in \mathcal{P}} (\mu_{u_h}^{-,i} l_{z_h} + \mu_{l_h}^{-,i} lpl_h) + \\ & + \sum\limits_{h \in \mathcal{P}} (-\mu_{u_h}^{+,i} u_{z_h} - \mu_{u_h}^{-,i} l_{z_h} + \mu_{l_h}^{+,i} upl_h - \mu_{l_h}^{-,i} lpl_h) y_h \le \theta \quad i \in I \\ & \sum\limits_{h \notin \mathcal{P}} (-v_{u_h}^{+,j} u_{z_h} + v_{u_h}^{-,j} l_{z_h}) + \sum\limits_{h \in \mathcal{P}} (v_{u_h}^{-,j} l_{z_h} + v_{l_h}^{-,j} lpl_h) + \\ & + \sum\limits_{h \in \mathcal{P}} (-v_{u_h}^{+,j} u_{z_h} - v_{u_h}^{-,j} l_{z_h} + v_{l_h}^{+,j} upl_h - v_{l_h}^{-,j} lpl_h) y_h \le 0 \quad j \in J \\ & y_h \in \{0, 1\} \quad h \in \mathcal{P}. \end{array} \tag{5}$$

Constraint

$$\theta \ge \sum_{h \in \mathcal{P}} \min(lpl_h, upl_h) w_h \tag{6}$$

10	15	11	9	45
8	**10**$^{(3)}$	**12**$^{(4)}$	15	45
10	12	**11**$^{(2)}$	**13**$^{(5)}$	46
28	37	34	37	136

Fig. 1. Small table for optimal CTA by Benders method. Sensitive cells are in boldface. Symmetric protection limits lpl_i and upl_i are in brackets. Weights are cell values ($w_i = a_i$).

may also be imposed to (5) since the master provides a lower bound on the optimal objective, and the right-hand-side of (6) provides a known lower bound on θ. Problems (4) and (5), together with the Benders algorithm were implemented in the AMPL modeling language [9]. Appendix A shows an extract of this implementation for (4) and (5).

5 Illustrative Example

Benders method is applied to the small two-dimensional table of Figure 1:

- Initialization (Step 0): $I = J = \emptyset$, lower bound (6) is 165.
- Iteration 1
  - Step 1: Solve master problem (5) with only constraint (6), obtaining $\theta_r^* = 165$, and $y_{r_i}^* = 1$ for all $i = 0, \ldots, 4$.
  - Step 2: Solve subproblem (4), with optimal objective $458 \geq 165 = \theta_r^*$. Add first optimality cut

$$60y_1 + 298160000y_2 + 350776000y_3 + 70155300y_4 + \theta \geq 458$$

    to (5).
- Iteration 2:
  - Step 1: Solve master problem (5):

$$\begin{array}{ll} \min & \theta \\ \text{s. to} & 60y_1 + 298160000y_2 + 350776000y_3 + 70155300y_4 + \theta \geq 458 \\ & \theta \geq 165. \end{array}$$

    obtaining $\theta_r = 165$, $y_i = 1$ for all $i = 1, \ldots, 4$.
  - Step 2: Solve subproblem (4), with optimal objective $458 \geq 165 = \theta_r^*$. Add second optimality cut

$$-60y_1 - 368y_2 - 304y_3 - 202y_4 + \theta \geq -476$$

    to (5).

$\vdots$

Table 1. Summary of illustrative example

iter.	$f^*_{Q_D}$	θ^*_r	y_1	y_2	y_3	y_4
1	458	165	0	0	0	0
2	458	165	1	1	1	1
3	330	165	1	0	0	1
4	334	165	1	1	0	1
5	346	165	0	1	0	0
6	303	165	0	1	0	1
7	334	238	0	0	1	0
8	303	274	1	0	1	0
9	303	303	1	0	1	0

Table 2. Instance dimensions

Instance	n	p	m	n. coef.
dale	16514	4923	405	33028
osorio	10201	7	202	20402
table8	1271	3	72	2542
targus	162	13	63	360
random1	22801	15000	302	45602
random2	30351	12000	352	60702
random3	40401	10000	402	80802
random4	40401	20000	402	80802
random5	35376	10000	377	70752
random6	10201	6000	202	20402
random7	10201	7000	202	20402
random8	20301	15000	302	40602
random9	20301	10000	302	40602
random10	40401	30000	402	80802
random11	30351	25000	352	60702
random12	10251	8500	252	20502
random13	37901	20000	402	75802
random14	22801	20000	302	45602
random15	25351	10000	352	50702
random16	22801	10000	302	45602
random17	22801	18500	302	45602
random18	15251	13000	252	30502
random19	15251	11000	252	30502
random20	22801	18500	302	45602

- Iteration 9:
  - Step 1: Solve master problem (5):

$$\begin{array}{ll}
\min & \theta \\
\text{s.to} & 60y_1+298160000y_2+350776000y_3+70155300y_4+\theta \geq 458 \\
& -60y_1-368y_2-304y_3-202y_4+\theta \geq -476 \\
& 36y_1+298160000y_2+44y_3-90y_4+\theta \geq 276 \\
& -320y_1-368y_2-44y_3+30y_4+\theta \geq -324 \\
& -36y_1-72y_2+36y_3+280621000y_4+\theta >= 274 \\
& 54y_1+24y_2-44y_3-418y_4+\theta >= -91 \\
& 350776000y_1+298160000y_2+44y_3-30y_4+\theta \geq 378 \\
& -54y_1-24y_2+44y_3+280621000y_4+\theta \geq 293 \\
& \theta \geq 165
\end{array}$$

    obtaining $\theta_r = 303$, and $y_1 = y_3 = 1$ and $y_2 = y_4 = 0$.
  - Step 2: Solve subproblem (4), with optimal objective $303 = 303 = \theta_r$. Solution found: $y^* = y^*_r$.

Table 1 summarizes the example, showing for each iteration the optimal objective function of the subproblem "$f^*_{Q_D}$" and the master problem "θ^*_r", and the values of y^*_r.

Table 3. Results with Benders method

Instance	CPU	iter.	MIP iter.	Simp. iter.	f_{Q_D}	θ
dale	8.47	20	2591	2783	3581.03	3549.53
osorio	73.94	123	13890	50903	6.0317	6.0073
table8	0.24	9	60	43	3.0848	3.0848
targus	0.62	16	449	399	59.3295	58.8393
random1	2.38	4	342	494	48477.7	47993
random2	3.48	5	463	775	38726.3	38398.8
random3	4.9	7	596	510	32170	31907.4
random4	4.22	4	338	466	64127.5	63522.7
random5	4.1	6	577	991	32159.7	31868
random6	1.73	7	1554	1395	12963.4	12835.9
random7	1.68	7	574	722	30250.1	29979.5
random8	3.46	6	461	661	48469	48088.3
random9	3.00	7	590	779	8852.87	8769.04
random10	5.90	5	372	642	64720.7	64111.1
random11	3.15	4	218	397	107088	106025
random12	4.07	12	1261	1282	18388	18210
random13	4.84	5	362	131	128188	127164
random14	3.15	5	337	495	170645	169344
random15	3.67	7	586	932	85189.8	84441.4
random16	4.26	9	985	1246	32339.3	32028
random17	3.24	5	468	635	59720.9	59170.5
random18	1.69	4	353	411	42052.8	41658.4
random19	2.02	5	462	553	35507.2	35209.6
random20	3.18	5	468	635	59720.9	59170.5

Table 4. Results with CPLEX

Instance	CPU	MIP iter.	f^*
dale	3.53	11559	3562.11
osorio	93.09	2093	6.0316
table8	1.08	147	3.0848
targus	0.13	107	59.3295
random1	14.26	33717	48074.07
random2	11.38	27526	38474.2
random3	8.75	23210	31998.93
random4	23.83	45053	63616.5
random5	9.41	23302	31953.81
random6	3.71	13754	12872.3
random7	4.69	16004	30063.69
random8	15.33	33749	48161.70
random9	8.2	22849	8791.66
random10	36.41	66306	64170.65
random11	26.8	55080	106138.14
random12	5.92	19329	18262.87
random13	22.42	44789	127343.58
random14	19.86	44124	169538.78
random15	9.06	23186	84688.72
random16	8.29	23228	32096.94
random17	19.08	41301	59240.23
random18	9.16	28959	41731.95
random19	8.53	24856	35280.94
random20	17.33	41301	59240.23

6 Computational Results

Benders algorithm for optimal CTA has been implemented using the AMPL mathematical programming modeling language [9]. This implementation has been applied to a set of four small pseudo-real and 20 random larger two-dimensional instances obtained with the generator used in [3]. All runs were carried on a Sun Fire V20Z server with two AMD Opteron processors (without exploiting parallelism capabilities), 8 GB of RAM, and under the Linux operating system. Table 2 show the instance dimensions: number of cells n, number of sensitive cells p (which is the number of binary variables), number of linear relations m, and number of coefficients in linear relations “n. coef.”.

Table 3 shows the results obtained with AMPL implementation of Benders method. Column “CPU” provides the CPU time for solution of the master and subproblems using CPLEX. Column “Benders iter.” gives the total number of Benders iterations. Columns “MIP iter.”and “Simp. iter.” show the total number of MIP and simplex iterations for, respectively, all masters and subproblems. Columns f_{Q_D} and θ show, respectively, the upper and lower bounds found. An optimality tolerance of 1% was used for all runs.

Table 4 shows the results with the CPLEX branch-and-cut algorithm. Column “CPU” provides the CPU time. Column “MIP iter.” gives the overall number of MIP simplex iterations. Column f^* provides the optimal objective function found. As for Benders, an optimality tolerance of 1% was used for all runs. It can be observed that in all instances, but for “dale” and “targus”, Benders method is faster than CPLEX. In particular, efficiency of Benders increases with the number of sensitive cells (i.e., binary variables), as in instances “random10”, “random11”, “random13”, “random14”, “random17” and “random18”. This makes it a promising approach for large tables.

7 Conclusions

This work presented an AMPL implementation of Benders decomposition for optimal CTA. The main benefit of this prototype code is to have a tool for ease testing with alternative cuts. Preliminary results for some small-medium two-dimensional tables show it can be a promising approach for more complex tables, if Benders can be appropriately tuned to efficiently deal with them. The development of a more efficient code, and applying it to larger two-dimensional tables, and more complex structures, is part of the further work to be done.

References

1. Benders, J.F.: Partitioning procedures for solving mixed-variables programming problems. Computational Management Science 2, 3–19 (2005); English translation of the original paper appeared in Numerische Mathematik 4, 238–252 (1962)
2. Castro, J.: Minimum-distance controlled perturbation methods for large-scale tabular data protection. European Journal of Operational Research 171, 39–52 (2006)
3. Castro, J.: A shortest paths heuristic for statistical disclosure control in positive tables. INFORMS Journal on Computing 19, 520–533 (2007)
4. Castro, J., Giessing, S.: Testing variants of minimum distance controlled tabular adjustment. In: Monographs of Official Statistics, Eurostat-Office for Official Publications of the European Communities, Luxembourg, pp. 333–343 (2006)
5. Castro, J., González, J.A.: A Benders decomposition approach to CTA, working paper, Dept. of Statistics and Operations Research, Universitat Politècnica de Catalunya (2008)
6. Cox, L.H.: Confidentiality protection by CTA using metaheuristic methods. In: Monographs of Official Statistics, Eurostat-Office for Official Publications of the European Communities, Luxembourg, pp. 267–276 (2006)
7. Dandekar, R.A., Cox, L.H.: Synthetic tabular Data: an alternative to complementary cell suppression, manuscript, Energy Information Administration, U.S, Available from the first author on request (Ramesh.Dandekar@eia.doe.gov) (2002)
8. Fischetti, M., Salazar, J.J.: Solving the cell suppression problem on tabular data with linear constraints. Management Science 47, 1008–1026 (2001)
9. Fourer, R., Gay, D.M., Kernighan, D.W.: AMPL: A Modeling Language for Mathematical Programming. Duxbury Press (2002)

A AMPL Models for Benders Subproblem and Master

A.1 Extract of AMPL Implementation of (4)

```
###############################################
# Definiton of Benders subproblem for CTA
###############################################
param lp{1..ncells} default 0;
param up{i in 1..ncells} default (ub[i]-a[i]);;
param ln{1..ncells} default 0;
param un{i in 1..ncells}default (a[i]-lb[i]);

var la_up {1..ncells} >= 0;
var la_un {1..ncells} >= 0;
var la_lp{1..ncells} >= 0;
var la_ln {1..ncells} >= 0;
var lambda {1..nconstraints};

maximize QD:
    sum {i in 1..ncells} (-la_up[i]*up[i] -la_un[i]*un[i]
    + la_lp[i]*lp[i] + la_ln[i]*ln[i]);
subj to R1_QD {i in 1..ncells}:
    sum{l in Trasbegconst[i]..Trasbegconst[i+1]-1}
        Trascoef[l]*lambda[Trasxcoef[l]]
        -la_up[i] + la_lp[i] =c[i];
subj to R2_QD {i in 1..ncells}:
    sum{l in Trasbegconst[i]..Trasbegconst[i+1]-1}
        -Trascoef[l]*lambda[Trasxcoef[l]]
        -la_un[i]+la_ln[i] =c[i];
```

A.2 Extract of AMPL Implementation of (5)

```
##################################################
# Definition of Benders master for CTA
##################################################
param nCUT >= 0 integer;
param iter >= 0 integer;
param mipgap;
param const {1..nCUT} default 0;
param consty {1..npcells,1..nCUT} default 0;
param cut_type {1..nCUT} symbolic within {"point","ray"};
param MinTheta;
```

```
var y {1..npcells} binary;
var Theta;

minimize BPr: Theta;
#Feasibility/optimality cuts
subj to Cut_Point {j in 1..nCUT}:
    if (cut_type[j]="point") then Theta else 0 >=
    const[j] + sum {i in 1..npcells} consty[i,j]*y[i];
subj to RMinTheta:  Theta >= MinTheta;
```

A Data Quality and Data Confidentiality Assessment of Complementary Cell Suppression

Lawrence H. Cox

National Center for Health Statistics, Centers for Disease Control and Prevention
3311 Toledo Road, Hyattsville, MD 20782 USA
lcox@cdc.gov

Abstract. Complementary cell suppression has been used for disclosure limitation of magnitude data such as economic censuses data for decades. This paper examines data quality and data confidentiality characteristics of cell suppression. We demonstrate that when cell suppression is not performed using a proper mathematical model, it can fail to protect. Moreover, we demonstrate that properly executed suppression based on standard disclosure definitions can be vulnerable to other attacks, sometimes fatally.

Keywords: alternating cycle, releasing exact intervals, p/q-ambiguity rule.

1 Introduction

Tabular data are aggregated data organized into tables. Each tabular dimension corresponds to a variable of interest partitioned into mutually exclusive characteristics. The simplest tabular structures are one- and two-way tables. A typical one-way table could present the number of students at a U.S. university receiving the grade A, B, C, D or F, respectively, in a particular course. A typical two-way table could present the aggregate of monthly sales from retail stores comprising 0-100, 101-200, or 200+ employees within each county of a state.

The tabular structure defines a partition of the population of interest--each subject in the population is assigned to a unique cross-classification (*cell*) of all of the variables. In a table of counts (*contingency table*), each subject contributes one unit to its partition cell and zero otherwise. If instead the subject contributes its particular value for a statistic (e.g., monthly retail sales), the data are *magnitude data*. Percentages, profit/loss, etc., also can be organized into tables, but interest here is restricted to nonnegative count and magnitude data. Tabular structures can be simple (two-way tables) or large (hierarchies of two-way tables), complex (multi-way tables, linked tables), or large and complex (linked multi-way tables).

Statistical disclosure in tabular data is defined by a disclosure rule (*sensitivity measure*) that identifies the disclosure (*sensitive*) cells. These are the cells achieving a positive value for the sensitivity measure. The measure is a continuous function that also can be used to indicate how "far" a sensitive cell is from being nonsensitive--this "distance" is its *protection limit*. Statistical disclosure limitation (**SDL**) of tabular data is *complete* if and only if only the tightest (*exact*) interval estimate of each sensitive cell value computable from the released tabulations is nonsensitive. These notions are made precise in [1] and elaborated in Section 2.

J. Domingo-Ferrer and Y. Saygın (Eds.): PSD 2008, LNCS 5262, pp. 13–23, 2008.

Complementary cell suppression (**CCS**) is a methodology for statistical disclosure limitation in tabular data. CCS replaces the value of each sensitive cell by a symbol (**D** for “disclosure”). Typically, these *primary suppressions* are insufficient to ensure complete SDL, and additional, nonsensitive cells (*complementary suppressions*) also have their values replaced by **D**. CCS methodology is focused on assuring good choices for the complementary cells, viz., a collection of cells that assures SDL while suppressing as little useful data as possible. See Section 2, and [2, 3] for a complete presentation.

We examine the data quality and confidentiality characteristics of complementary cell suppression. Section 2 provides SDL preliminaries. Section 3 examines usability of suppressed data and other data quality characteristics of CCS, and describes an oft-discussed alternative to full suppression--releasing exact intervals in lieu of symbols for suppressed values. In Section 4 we examine CCS mathematically. Section 5 presents vulnerabilities of CCS and a widely used sensitivity measure, and demonstrates serious vulnerabilities associated with releasing exact intervals. Section 6 contains concluding comments.

2 Statistical Disclosure Limitation Preliminaries

2.1 Sensitivity Measures

SDL methods such as *rounding* and adding zero-mean random noise (*random perturbation*) are alternatives to suppression for count data, but can be ineffective or inappropriate for magnitude data. Suppression has been applied to major, important data collections of economic and other magnitude data in the US, Canada and the European Union, in some cases for decades, based on software developed at the US Census Bureau, the US National Center for Health Statistics, Statistics Canada, and the EU CASC Project. For these reasons, we focus on magnitude data.

A simple, widely used disclosure rule for magnitude data is the *p-percent rule* which, in simplified form, states: a tabulation cell **X** is sensitive if, after subtracting the second largest contribution from the cell value, the remainder is within p-percent of the largest contribution. This rule is designed to prevent narrow estimation of any contribution to a cell value by a second contributor or third party. Note: protecting the largest from the second largest assures protection for all others.

X denotes a tabulation cell, and its cell value is x . Order the contributions to x from largest to smallest and denote these contributions x_i, so that $x = \sum_i x_i$; $x_1 \geq x_2 \geq .... x_i \geq ....$ Express p as a decimal; e.g., 20% = 0.20. Sensitivity for the p-percent rule is expressed: $S_p(X) = x_1 - (1/p)\sum_{i \geq 3} x_i > 0$.

An extension of the p-percent rule that incorporates prior information in the hands of the intruder is the *p/q-ambiguity rule*: the releaser assumes that an intruder can estimate any contribution to within q-percent, $1 \geq q >> p$. Express q as decimal. The sensitivity measure for the p/q-ambiguity rule is $S_{p/q}(X) = x_1 - (q/p)\sum_{i \geq 3} x_i > 0$. When

$q = 1$, the p-percent and p/q-ambiguity rules are identical; otherwise, it is evident that the p/q-ambiguity rule is *stricter* than the p-percent rule—it identifies as sensitive all p-sensitive cells and possibly more.

A sensitivity measure is a continuous function. If it is normalized with the leading coefficient = 1 (as above), then its value r measures the distance from the sensitive cell value x to larger values that would not be sensitive under the same circumstances. r is called the upper protection limit. Typically, the lower protection limit is set to $-r$, and the open interval $(x - r, x + r)$ is the *protection interval.* See [1].

2.2 Complementary Cell Suppression

Complementary cell suppression is a very difficult problem theoretically and computationally. CCS usually is accomplished using mathematical programming. A mathematical program for CCS is obtained as follows.

Represent the tabular structure as **Ay** = **b.** Entries of **A** are 0 or 1. The original data is **a** = $(a_1, \ldots, a_n)$, so that **Aa** = **b**. Denote the sensitive cell values $a_{d(i)}$, i = 1, …, s, and the protection limits $r_{d(i)}$, $0 \le r_{d(i)} \le a_{d(i)}$, with $r_k = 0$ otherwise.
(Note: in general, upper and lower protection limits can be unequal). A mathematical programming model for CCS is given by (1). See also [4].

The first constraint of (1) preserves the tabular structure. The second and third enforce the sensitivity measure. $M \geq 1$ is a suitable constant. The choice of objective function is used to enforce some notion of data quality: to minimize number of cells suppressed, set $c_k = 1$; to minimize total value suppressed, $c_k = a_k$; and, to minimize Berg entropy (a compromise between number and total value suppressed), $c_k = \log(1 + a_k)$. Note that these are geometrical, not necessarily statistical, measures of quality.

$$\begin{gathered} \min \sum_k c_k z_k \\ i=1,\ldots,s;\ j=1,2;\ k=1,\ldots,n: \\ Ay_{i,j} = b \\ a_k(1-z_k) \le y_{i,1,k} \le a_k - r_k z_k \\ a_k + M z_k \ge y_{i,2,k} \ge a_k + r_k z_k \\ z_j = 0,1;\quad z_{d(i)} = 1 \end{gathered} \tag{1}$$

2.3 Releasing Exact Intervals in Place of Suppressions

Magnitude data are treated as continuous data, and therefore exact interval estimates of suppressed cell values y_k can be obtained via linear programming: min y_k (respectively, max y_k) subject to **Ay** = **b.** The exact interval for a_k is [min y_k, max y_k]. By Sec. 2.2, model constraints assure exact intervals contain protection intervals.

For the p-percent rule, sophisticated users can use compute exact intervals, albeit at some effort. Why, it has been argued, doesn't the releaser release exact intervals in lieu of suppressed data? Doing so improves the ability of analysts to manipulate disclosure-limited tabular data, and perhaps analytical precision as well. From a data quality and usability perspective, this is a sound argument, raised in the 1970s.

Releasing intervals recently reemerged as *partial cell suppression* [5, 6]. We examine this and related issues from a data confidentiality standpoint in Section 5.

3 Data Quality Characteristics of CCS

We consider two dimensions of data quality

- *local quality*: focused on individual values and comparing values
- *global quality:* focused on distributions, inferences and statistics

Local quality characteristics of tables with suppressions are as follows. Unsuppressed cell values are released unchanged. Users of tabular data are often interested in specific values, and users interested in any of the unsuppressed values are able to work with the "true" values. On the contrary, sensitive and complementary values are suppressed. Unsophisticated users interested in any of these values will be thwarted as true values are unavailable. If the releaser released exact interval estimates of suppressed values, the unsophisticated user could impute interval midpoints for suppressed data and analyze the "midpoint tables." Users are likely to do this because it is simple. However, the resulting tables may fail to be additive, but at additional effort could be *adjusted* ([7]) to restore additivity. Alternatively, additivity is preserved if iterative proportional fitting (IPF) is used to impute suppressed values. The sophisticated user can invoke linear programming to do so even if the releaser fails to provide exact intervals.

Exact intervals can be very broad as the position of values in the table may force complementary suppressions that are large. Also, the mechanics of the simplex algorithm forces maximal masking of suppressed data. This is called *overprotection* or *oversuppression.* In the working example to follow (Table 3), r(X) = 2 units of protection is required but 5 (or 8) units are actually provided.

Prominent among global quality characteristics of tables with suppressions is that missing (suppressed) data thwarts analysis, somewhat so for regression and to a considerable degree for analysis of trend. Release of exact intervals enables all users to proceed with analyses such as via midpoint or IPF imputation.. Releasing exact intervals spares sophisticated users the trouble of computing the intervals directly. And, providing the intervals can demonstrate that the disclosure limitation was successful.

4 Mathematical Properties of CCS

4.1 CCS Can Be Vulnerable

If complementary cell suppression is performed using a mathematical model that incorporates protection constraints explicitly, such as (1), exact intervals for suppressed sensitive cells must be nonsensitive and disclosure limitation is complete. Model (1) is an integer linear program, which can be difficult to impossible to solve computationally by direct means such as branch and bound, except for small problems. Recent research has focused on solving medium to large CCS problems using branch and cut and specialized techniques [4]. Unfortunately, many organizations

Table 1. 3x3 Table With Internal Entries Suppressed

$$\begin{array}{ccc} \begin{pmatrix} * & * & * \\ * & * & * \\ * & * & * \end{pmatrix}\begin{pmatrix} 11 \\ 5 \\ 5 \end{pmatrix} & \begin{pmatrix} * & * & * \\ * & * & * \\ * & * & * \end{pmatrix}\begin{pmatrix} 5 \\ 11 \\ 5 \end{pmatrix} & \begin{pmatrix} * & * & * \\ * & * & * \\ * & * & * \end{pmatrix}\begin{pmatrix} 5 \\ 5 \\ 11 \end{pmatrix} \\ \begin{pmatrix} 11 & 5 & 5 \end{pmatrix}(21) & \begin{pmatrix} 5 & 11 & 5 \end{pmatrix}(21) & \begin{pmatrix} 5 & 5 & 11 \end{pmatrix}(21) \end{array}$$

$$\begin{pmatrix} 1 & 10 & 10 \\ 10 & 1 & 10 \\ 10 & 10 & 1 \end{pmatrix}$$

continue to solve CCS problems "by hand" or using computer programs based on "by hand" reasoning. These programs are faster than humans, but in the absence of CCS methodology, offer little improvement in terms of protection or data quality. As an example, we use Table 1, and a typical disclosure rule for counts that defines an unsafe interval to be the closed interval [1, 4] or smaller (*5-threshold rule*).

Table 1 is a 3x3x3 contingency table with all internal entries suppressed for confidentiality. In lieu of releasing the internal entries (3-dimensional cells), all 2-dimensional cells ("line" marginal totals) are released. This approach is believed to limit disclosure, such as in this example.

Table 1 is not a realistic confidentiality example because it contains published marginal totals with value = 1. We address this momentarily. Meantime, compute 2-dimensional Frechet lower bounds for cells (1,1,1), (2,2,2) and (3,3,3) within planes k = 1, 2, 3, respectively. Each of these lower bounds = 1. Each of these three cells is constrained by a marginal total (vertical) = 1, and consequently these cells cannot achieve value > 1. Hence, each has value =1, which has been revealed and is sensitive--disclosure limitation has been entirely unsuccessful.

Now the issue of realism. Replace Table 1 by a table comprising 5 copies of Table 1 stacked vertically, viz., a 3x3x15 table with two sets of planar marginals unchanged and the third (vertical) set with values five times those of Table 1 (table not shown here). This is a realistic example (no marginals < 5) for which 15 cells are revealed to have value = 1.

Table 2, a two-dimensional table with suppressions, is a second example illustrating the failure of non-mathematical CCS methods.

Table 2. 4x5 Table With Suppressions

18	**21**	**18**	**23**	**80**
D_{11}	D_{12}	D_{13}	9	**20**
6	D_{22}	D_{23}	6	**20**
D_{31}	5	5	D_{34}	**15**
D_{41}	5	6	D_{44}	**25**

CCS in Table 2 may appear successful, as each suppressed cell is contained in a row and a column containing one or two additional suppressions and corresponding sums are $\geq$ 5. But, in fact, D_{11} = 1 can be deduced: add the first two rows: $D_{11}+D_{12}+D_{13}+D_{23}+D_{33}=19$; add the second and third columns: $D_{12}+D_{13}+D_{23}+D_{33}=18$; subtract the latter from the former, to obtain D_{11} = 1.

Both examples illustrate that CCS should be done algorithmically and NOT "by hand" or by software based, in essence, on "by hand" reasoning.

4.2 The Mechanics of CCS

Complementary cell suppression replaces all sensitive and selected nonsensitive values, which are fixed, by symbols, which can be treated as variables. By definition, a particular CCS solution is complete if all exact interval for variables corresponding to sensitive values *x* contain the cell's protection interval (*x* – *r, x* + *r*). Table 3 will be used as a working example.

Table 3. 4x5 Working Example

T	O	T	A	L	T
					O
	X(10)		**B(5)**		T
					A
	C(7)		**A(8)**		L

X denotes a sensitive cell, and **A, B, C** denote **X**'s complementary suppressions. Table 3 is the simplest example possible, and it can be misleading to generalize from a simple example to the general case. However, as complex suppression patterns can be decomposed into simple patterns of (elaborations of) this form, for CCS any information derivable from Table 3 will be valid in general for tables with suppressions, viewed as a composite of smaller, simpler tables resembling (elaborations of) Table 3. For our analysis, we extract the essence of Table 3 and provide hypothetical values for the reduced marginal totals (Table 4).

Table 4. Essentials of Table 3

17	13	30
x=10	b=5	15
c=7	a=8	15

Let *r* = 2. Then **X** is protected if and only if no interval derivable for *x* is finer than (*x*-*r, x*+*r*) = (10-2, 10+2) = (8, 12). This condition holds if **X** is in an *alternating cycle* of suppressed cells - the cycle permits a *flow* of *r* = 2 units from x = 10 in both **+** and **–** directions. The alternating cycle is given by

17	13	30
X (10)+/-	**B** (5)-/+	15
C (7)-/+	**A** (8)+/-	15

In the + direction, can move up to 5 units into **X**--more would force *b* < 0.

17	13	30
X (15)	**B** (0)	15
C (2)	**A** (13)	15

In the – direction, we can move up to 8 units out of **X**--more would force *a* < 0.

17	13	30
X (2)	**B** (13)	15
C (15)	**A** (0)	15

Verification that Table 4 protects **X** is demonstrated by exact intervals below. As we can move *r* = 2 units in either direction, **X** is protected.

17	13	30
X [2, 15]	**B** [0, 13]	15
C [2, 15]	**A** [0, 13]	15

CCS is also data dependent--the exact intervals above are much broader than the protection limit, whereas the same pattern fails to protect the table below.

17	6	23
X (10)+/-	**B** (5)-/+	15
C (7)-/+	**A** (1)+/-	8

5 Confidentiality Characteristics of CCS

5.1 CCS, Cycles and Protection

Movement of up to 5 (respectively, 8) units through sensitive cell **X** may be represented by the alternating cycle below.

17	13	30
X (10)+/-	**B** (5)-/+	15
C (7)-/+	**A** (8)+/-	15

17	13	30
x+/-	**b**-/+	15
c-/+	**a**+/-	15

Cells marked with +/- have the *same parity* as *x;* those with -/+ have *opposite parity* to *x*. In general,

- maximum increase to *x* = minimum value with opposite parity
  (here, b = 5)
- maximum decrease to *x* = minimum value with same parity
  (here, a = 8*)*
- exact interval for *x* = [*x-a, x+b*]
  (here, = [2, 15])
- *width* of exact interval = (*b+a*)
  (here, = 13)
- *radius* of exact interval = *(b+a)/2*
  (here, = 6.5*)*
- interval *midpoint* = *x* + (*b-a*)/2
  (here, = 8.5)
- *bias* in midpoint estimate of *x* = (*b-a*)/2
  (here, = -1.5)

CCS is based on creating cycles that

- contain the sensitive cells **X**
- collectively permit increase/decrease of *x* to at least (*x-r(X), x+r(X))*
- minimize information loss measured by linear cost function $\sum_k a_k z_k$

Typically (but not necessarily)

- a sensitive cell will be used as a complement for another cell whenever possible
- complementary cells are large enough to accommodate protection limits *r(X)*
- but as small as possible to minimize information loss
- an alternative is to select many small cells as complements

Multi-dimensional and linked tables are much more complex for CCS but each two-dimensional slice of such systems must comprise alternating cycles as above. Multiple cycles containing a sensitive cell are analyzed sequentially for purposes here. Thus, we continue to focus on alternating cycles.

5.2 Releasing Exact Intervals

The data releaser may choose to release exact intervals [l, u] for the suppressed cells. Even if the releaser does not do so, the sophisticated user can compute these intervals independently. So, it suffices to assume that exact intervals are available. We return to our working example. Again, we remind the reader that all situations are not as simple as this 4-element two-dimensional cycle; but, that all situations do comprise two-dimensional cycles that the intruder can analyze in precisely the same manner as we now proceed to do.

17	13	30
x+/-	**b**-/+	15
c-/+	**a**+/-	15

Assume for concreteness that $b \leq c$ and $a \leq x$. By virtue of the polyhedral geometry of linear constraint systems, the intruder can determine the following.

- l(x) $= x - a$: a of same parity as x, and l(a) = 0
- u(x) $= x + b$: b of opposite parity to x, l(b) = 0
- intruder knows the width of the exact interval $= a + b$
- if intruder can determine a or b or b-a or b/a, then x is revealed

Consequently, protection on a cycle hinges on the intruder's ability to determine a single quantity. If $(b-a)/(2x)$ is small, then the midpoint estimate is precise. Similarly, if **A**, **B** are not historically sensitive, then the intruder can examine historical data to estimate a or b or $b - a$ or b/a, and estimate x.

If **X** involves only one contributor, then

$$\text{l(x)} \leq (1\text{-p})\text{x and}$$
$$(1\text{+p})\text{x} \leq \text{u(x)}$$

Consequently,

$$\text{l(x)/(1-p)} \leq \text{x} \leq \text{u(x)/(1+p)} \tag{2}$$

In addition,

- $\text{u(x)/l(x)} \geq (1\text{+p})/(1\text{-p})$,

so that

- $(\text{u(x)-l(x)/(u(x)+l(x))} \geq \text{p}$

Often, contributor counts are released, so the intruder knows precisely the one contributor cells. If **X** involves two contributors, then the second contributor can obtain an analogous but sharper inequality from its cell equation by subtracting out its contribution. Typically, there are (many) one and two contributor cells, and the number of contributors is published, enabling identification of these cells.

By virtue of (2), exact intervals can be "shrunk" if p is known. It is often discussed as to whether the releaser should make the value of p public to enhance analyzability of the data. It would appear that the answer to that question is a resounding "NO". The next oft-asked question is whether the releaser should release the minimal safe interval (x – r(X), x + r(X)**)**. Again the answer is "NO" because in so doing, x = midpoint of (x-r(X), x+r(X)) is divulged, as are $r = r(X) = (x + r(X)) - (x)$ and $p = r/x$. Under *sliding protection*, viz., requiring only that the width of the protection interval be at least $2r(X)$, the second question is moot.

If releasing p erodes protection, how well protected is the value of this parameter? For each one or two contributor cell **X**

$$\begin{aligned} \text{p} &\leq \text{p(X)} = \text{(u(x)-l(x))/(u(x)+l(x))} \\ &= \text{((u(x)-l(x))/2)/((u(x)+l(x))/2)} \\ &= \text{(radius/midpoint) of the protection interval for } \mathbf{X} \end{aligned} \tag{3}$$

These inequalities provide (many) upper bounds for p. In the context of a national census, or a set of different censuses or censuses conducted over multiple years, many upper bounds (3) for p are available. The smallest, $p' = p(X')$, could be very precise.

A lower bound **p''** on **p** can be obtained via trial and error as follows.

- begin with any solution (e.g,, adjusted midpoint or IPF)
- choose p'' and protect **X** to within p''-percent
- if the current cycle is not selected, then $p > p''$
- do this for each one contributor cell **X**
- the largest p'' is a lower bound for p

The intruder then can obtain a tighter interval than the exact interval $l \le x \le u$:

$$l \le l/(1-p'') \le l/(1-p) \le x \le u/(1+p) \le u/(1+p'') \le u \qquad (4)$$

5.3 Vulnerability of CCS under p/q-Rule and Release of Exact Intervals

X denotes a sensitive cell under a p/q-rule, and is suppressed with complementary suppressions **A**, **B**, **C**. The releaser releases exact intervals in place of suppressions.

X $[l_X, u_X]$……+/-	**B** $[l_B, u_B]$ -/+
C $[l_C, u_C]$ -/+	**A** $[l_A, u_A]$ +/-

Assume $l_A, l_C, l_X \ge l_B$ (other cases analogous). Thus, $a, c, x \ge b$. From the polyhedral geometry of linear constraints, the intruder can deduce

- $u_X - l_X = u_B - l_B = u_A - l_A = u_C - l_C = 2q \min\{a, b, c, x\} = 2qb$
- $l_B = (1 - q)b$
- $u_B = (1 + q)b$
- $l_A = a - qb$
- $u_A = a + qb$

Should the releaser release the value of q? The answer is definitely "NO" because these equations would reveal a, b, c and x. Indeed, it makes no difference whether or not the releaser reveals q, as q is in fact knowable.

For q < 1,

- $u_B / l_B = (q+1)/(q-1)$
- $q = (u_B - l_B)/(u_B + l_B)$

Consequently,

- $b = l_B/(1 - q)$
- $a = l_A + qb$
- $c = l_C + qb$,
- $x = l_X + qb$

Thus, release of exact intervals for a p/q-rule results in COMPLETE DISCLOSURE!

6 Concluding Comments

We have shown that complementary cell suppression has negative effects on both local and global data quality. Attempts to mitigate these effects include

- release the parameter p of a p-percent rule
- release exact intervals in place of suppressions
- release the parameter q of a p/q-rule

We have shown that these alternatives may seriously threaten confidentiality. We have presented an approach by which the security of suppressed data may be compromised. The extent of these threats in practice needs to be examined, potential remedies need to be explored, and alternatives, such as [7], considered.

Disclaimer. This work solely represents the findings and opinions of the author and should not be interpreted as representing the policies or practices of the Centers for Disease Control and Prevention or any other organization or group.

References

1. Cox, L.H.: Linear sensitivity measures in statistical disclosure control. Journal of Statistical Planning and Inference 5, 153–164 (1981)
2. Cox, L.H.: Suppression methodology and statistical disclosure control. Journal of the American Statistical Association 75, 377–385 (1980)
3. Cox, L.H.: Network models for complementary cell suppression. Journal of the American Statistical Association 90, 1453–1462 (1995)
4. Fischetti, M., Salazar, J.J.: Solving the cell suppression problem on tabular data with linear constraints. Management Science 47(7), 1008–1026 (2001)
5. Fischetti, M., Salazar, J.J.: Partial cell suppression: a new methodology for statistical disclosure control. Statistics and Computing 13, 13–21 (2003)
6. Salazar, J.J.: A unified mathematical programming framework for different statistical disclosure limitation methods. Operations Research 53, 819–829 (2005)
7. Cox, L.H., Kelly, J.P., Patil, R.: Balancing quality and confidentiality for multivariate tabular data. In: Domingo-Ferrer, J., Torra, V. (eds.) PSD 2004. LNCS, vol. 3050, pp. 87–98. Springer, Heidelberg (2004)

Pre-processing Optimisation Applied to the Classical Integer Programming Model for Statistical Disclosure Control

Martin Serpell[1], Alistair Clark[1], Jim Smith[1], and Andrea Staggemeier[2]

[1] University of the West of England
Bristol, United Kingdom
Martin2.Serpell@uwe.ac.uk, Alistair.Clark@uwe.ac.uk,
James.Smith@uwe.ac.uk

[2] Head of Statistical Tools and Operational Research Team
Office for National Statistics
Newport, United Kingdom
Andrea.Staggemeier@ons.gsi.gov.uk

Abstract. A pre-processing optimisation is proposed that can be applied to the integer and mixed integer linear programming models that are used to solve the cell suppression problem in statistical disclosure control. In this paper we report our initial findings which confirm that in many situations the pre-processing optimisation can considerably reduce the resources required by the solver hence allowing either statistical tables to be protected quicker, or larger statistical tables to be protected. This pre-processing optimisation may be suitable for application to the τ-Argus Optimal Method used in protecting statistical tables.

Keywords: Statistical Disclosure Control, Cell Suppression Problem, Classical Model, Pre-processing Optimisation, External Attacker.

1 Introduction

Many statistical tables are published with some of the table cells suppressed (left blank). This is done to prevent the disclosure of individual respondents which contributed to the cell value. Cells that failed the primary rule are called primary, or sensitive, cells and must be protected by additional suppressed cells called secondary cells. Choosing which secondary cells to suppress is known, in the literature, as the cell suppression problem. The cell suppression problem involves choosing a set of secondary cells that will remove the risk of disclosing the values of the primary cells whilst also minimising the information loss from the published statistical table.

The cell suppression problem is a member of the class of NP-hard problems when solving for optimality. In fact, the problem of finding a secondary suppression pattern is easy to be achieved, for example if all cells are suppressed this is a feasible pattern but clearly not optimal. It is when solving the cell suppression problem optimally that as the size of the table to be protected grows the number

J. Domingo-Ferrer and Y. Saygın (Eds.): PSD 2008, LNCS 5262, pp. 24–36, 2008.

of possible solutions that need to be evaluated grows much quicker. For a table with n cells there are 2^n possible suppression patterns. Because of the very large number of constraints that define the cell suppression problem MIP techniques can only find the optimal solution for small and medium sized statistical tables.

It is known that removing anything that is redundant from the mathematical program can make an efficiency gain. For example redundant equations, variables and protection levels can be removed. Another pre-processing efficiency gain can be obtained by removing any table cells that have the value set to zero or whose values must be published, subject to adjusting any marginal totals necessary. This decreases the number of working variables and constraints that the solver requires to find a solution, which in turn allows larger statistical tables to be protected.

Linear programming models and local search algorithms are used on relaxed cell suppression problems to obtain near optimal solutions when integer programming models are infeasible. Some have moved away from trying to calculate the optimal solution and have instead employed heuristic techniques to find near optimal solutions quickly. Others have employed hybrid algorithms that combine linear programming and heuristic techniques [2] [9].

This paper will present further improvements which are obtained when looking at the inferences made by an external attacker to a table. Section 2 presents definitions to the problem. Section 3 puts forward a conjecture for a pre-processing optimisation. Section 4 describes how the pre-processing optimisation can be implemented. Section 5 applies the pre-processing optimisation to the classical IP model for SDC. Section 6 describes our experimental setup. Section 7 contains our results. Section 8 contains our preliminary conclusions and section 9 lists further research.

2 Definitions

The *external attacker* wishes to deduce the values of cells that have been suppressed in a published statistical table, in order to glean confidential information. The assumption made in the literature is that the external attacker has only the knowledge which is provided in the published table, i.e. he is not aware which suppressed cells are primary nor secondary but he knows that there is a number of suppressed cells in the table and their location (disclosure pattern). As each table has row and column totals, often referred to as marginals, the *external attacker* is able to calculate lower and upper bounds, feasibility range, for each of the suppressed cells by solving a set of linear constraint equations [1] [7].

A statistical table with marginal totals can be represented as a set of cells, please see details of the model in [1] and [7], $a_i, i = 1, ..., n$, satisfying m linear constraint equations such that $Ma = 0$, where M_{ij} has one of the values $\{0, +1, -1\}$.

$$\sum_{i=1}^{n} M_{ij} a_i = 0, j = 1, ..., m$$

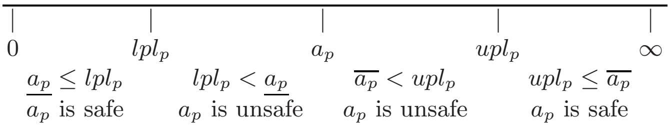
The statistical agency will define a set P of primary cells whose publication will be suppressed in order to protect the confidentiality of the contributors to those cells. The statistical agency will provide lower and upper protection levels (lpl and upl) for each cell in P such that an external attacker must not be able to calculate a_p within the range lpl_p to upl_p. For a_p to be safe

$$\underline{a_p} \leq lpl_p \quad \text{and} \quad \overline{a_p} \geq upl_p$$

where $\underline{a_p}$ is the lower bound and $\overline{a_p}$ the upper bound of the feasible range that the external attacker can calculate for a_p if only the primary cells P have been suppressed [1].

$$\begin{array}{lll} \underline{a_p} = min\ x_p & & \overline{a_p} = max\ x_p \\ \quad s.t.\ Mx = 0 & \text{and} & \quad s.t.\ Mx = 0 \\ \qquad x_i \geq 0,\ i \in P & & \qquad x_i \geq 0,\ i \in P \\ \qquad x_i = a_i, i \notin P & & \qquad x_i = a_i, i \notin P \end{array}$$

If the external attacker is able to calculate $\underline{a_p} > lpl_p$ or $\overline{a_p} < upl_p$ then a_p is unsafe (a_p can be disclosed). It should be noted that we are considering the external attacker on tables which have not yet been protected by secondary suppressed cells in order to gauge the level of disclosiveness of the tables for our pre-processing optimisation.



Noting that some primary cells may occur alone in a marginal total, whereas others (e.g. those sharing rows/columns) may effectively protect each other, we define the following partition of the set of primary cells P.

An *exposed* primary cell in a statistical table with marginal totals is one whose value can be calculated, within a given lower and upper protection limit, by an external attacker when only the primary cells P have been suppressed. That is to say, p is a member of the set E of *exposed* primary cells if $\underline{a_p} > lpl_p$ or $\overline{a_p} < upl_p$. $E \subseteq P$.

A *not exposed* primary cell in a statistical table with marginal totals is one whose value cannot be calculated, within a given lower and upper protection limit, by an external attacker when only the primary cells P have been suppressed. That is to say, p is a member of the set N of *not exposed* primary cells if $\underline{a_p} \leq lpl_p$ and $\overline{a_p} \geq upl_p$. $N \subseteq P$, $E \cup N = P$ and $E \cap N = \{\}$. The reason why there may be *not exposed* primary cells in a statistical table is due to their locations in that table. Each not exposed primary cell receives sufficient protection from other primary cells in the table to prevent an external attacker from being able to calculate a feasible range of values within the given protection level.

Proposition 1. As *not exposed* primary cells are already sufficiently protected they do not require secondary cells for their protection.

Proof. Follows by definition of N.

An *initially exposed* primary cell is a primary cell that can be exposed, by an external attacker when only the primary cells P have been suppressed, without requiring the exposure of any other primary cell. For example there may be only one primary cell in a row or column. Let L_p be the subset of linear equations M that contain the value $+1$ or -1 in the locations for a_p, $L_p \subseteq M$. This subset L_p only contains the linear equations that apply to a_p. Then we can say that p is a member of the set I of *initially exposed* primary cells if $\underline{a_p} > lpl_p$ or $\overline{a_p} < upl_p$, when,

$$\begin{array}{llcll} \underline{a_p} = min\ x_p & & & \overline{a_p} = max\ x_p & \\ & s.t.\ L_p x = 0 & \text{and} & & s.t.\ L_p x = 0 \\ & \quad x_i \geq 0,\ i \in P & & & \quad x_i \geq 0,\ i \in P \\ & \quad x_i = a_i, i \notin P & & & \quad x_i = a_i, i \notin P \end{array}$$

$I \subseteq E$.

Conversely we can say that p is not a member of I if $\underline{a_p} \leq lpl_p$ and $\overline{a_p} \geq upl_p$.

A *consequentially exposed* primary cell is an *exposed* primary cell that is not an *initially exposed* primary cell. That is to say, p is a member of the set C of *consequentially exposed* primary cells if p is a member of E but not a member of I. $C \subset E$, $C \cup I = E$ and $C \cap I = \{\}$. Hence a *consequentially exposed* primary cell is only vulnerable to an external attacker when at least one other *exposed* primary cell has been exposed. In other words, we defined I by considering only the submatrix L of consistency equations directly involving the cells in I. In contrast, *consequentially* exposed primary cells are those that, if only P is suppressed, may be exposed as a result of considering all of the consistency equations in matrix M. When an external attacker has exposed a primary cell it was for one of two reasons, the cell was either initially or consequentially exposed. If $I = \{\}$ then both $C = \{\}$ and $E = \{\}$.

3 Conjecture

In order to make a published statistical table safe from an external attacker only the *initially exposed* primary cells, I, need to be protected by suppressing secondary cells. Again, the proof follows by our definitions of I, C and N.

It is worth noting that even if it makes great improvements in some cases, this pre-processing stage does not change the NP-hard nature of the cell suppression problem, since in the worst case $I = P$, and we also now have to solve the problem of finding I.

A Corollary to this conjecture is that if $I = \{\}$ then $N = P$ and therefore the statistical table is already adequately protected.

4 Finding *Initially Exposed* Primary Cells without Using a Solver

We present here a method that provides a superset of the elements in P that contains all those in I.

Let Q be the set of cells that are not in P, these cells do not require protection. Let

$$c_j = \sum_{i \in Q} M_{ij} a_i, \quad j = 1, ..., m$$

then,

$$c_j + \sum_{i \in P} M_{ij} a_i = 0, \quad j = 1, ..., m$$

A necessary condition for us to establish that $p \in I$ is the existence of at least one constraint equation in which the amount of "uncertainty" (and hence protection) provided by the given lower and upper protection levels of the other suppressed primary cells in that constraint equation is less than the required protection limits for p. For each element $p \in P$ let J denote the set of linear constraint equations (equivalent to rows of M) in which p participates, i.e. $\forall j \in J \cdot M_{pj} \neq 0$. For each $j \in J$ let H_j be the set of primary cells in j. For each $p \in P$,

$$\underline{a'_p} = max_{j \in J}(-c_j - \sum_{i \in H_j/p} M_{ij} upl_i)$$

$$\overline{a'_p} = min_{j \in J}(-c_j - \sum_{i \in H_j/p} M_{ij} lpl_i)$$

Then we can say that p is a candidate member of the set I of *initially exposed* primary cells if $\underline{a'_p} > lpl_p$ or $\overline{a'_p} < upl_p$. The set of candidate members of I contains the set I. This is because the values of $\underline{a_p}$ and $\overline{a_p}$ that the external attacker can calculate can not be better than $\underline{a'_p}$ and $\overline{\overline{a'_p}}$, but are likely to be worse.

$$\underline{a_p} \leq \underline{a'_p} \quad \text{and} \quad \overline{a_p} \geq \overline{a'_p}$$

This is shown pictorially, for the lower protection level, in tables 1, 2 and 3.

4.1 Example

Taking a 6 by 6 statistical table with marginal totals (Table 4) as an example, the process of finding I, C and N can be shown. In our example the statistical agency has defined $P = \{8, 12, 15, 16, 19, 20, 24, 27\}$. When the test for the fully exposed primary cells is applied five primary cells are exposed, $E = \{16, 19, 20, 24, 27\}$ and therefore $N = \{8, 12, 15\}$. The values of cells 16, 20, 24 and 27 are calculated exactly and the feasibility range of cell 19 is calculated within its lower and upper protection levels which in this case is 10% of the cell's value.

By contrast applying the test for initially exposed primary cells (Table 5) we find that $I = \{16, 19, 24, 27\}$, and therefore $N \cup C = \{8, 12, 15, 20\}$. For this pre-processing optimisation to work it is not necessary (nor is it possible) to determine which cell is in C and which is in N.

Applying a SAS/OR implementation of the classical IP SDC model to the whole set of primary cells in table 4 the set of secondary cells $S = \{37, 38, 40\}$ was

Table 1. As $\underline{a_p} \leq lpl_p$, p is not a member of I. As $\underline{a'_p} \leq lpl_p$, p is not a candidate member of I.

0	$\underline{a_p}$	$\underline{a'_p}$	lpl_p	a_p	upl_p	∞

Table 2. As $\underline{a_p} \leq lpl_p$, p is not a member of I. As $lpl_p < \underline{a'_p}$, p is a candidate member of I.

0	$\underline{a_p}$	lpl_p	$\underline{a'_p}$	a_p	upl_p	∞

Table 3. As $lpl_p < \underline{a_p}$, p is a member of I. As $lpl_p < \underline{a'_p}$, p is a candidate member of I.

0	lpl_p	$\underline{a_p}$	$\underline{a'_p}$	a_p	upl_p	∞

Table 4. As $lpl_p < \underline{a_p}$, p is a member of I. As $lpl_p < \underline{a'_p}$, p is a candidate member of I.

	Total	1	2	3	4	5	6
Total	1472	193	278	203	294	233	271
A	199	$^{8}\,\mathbf{9}_{1}$	51	41	47	$^{12}\,\mathbf{3}_{1}$	48
B	164	$^{15}\,\mathbf{8}_{2}$	$^{16}\,\mathbf{1}_{1}$	54	44	$^{19}\,\mathbf{45}_{2}$	$^{20}\,\mathbf{12}_{2}$
C	245	8	70	$^{24}\,\mathbf{6}_{2}$	76	64	$^{27}\,\mathbf{21}_{2}$
D	248	33	46	45	27	37	60
E	312	87	51	18	35	49	72
F	304	48	59	39	65	35	58

Table 5. Workings to find members of the superset of I. Any primary cell that is at risk of having it's protection range violated is a member of the superset of I.

Cell	Protection range	$\underline{a'_p}$	$\overline{a'_p}$	Protected
8	8 to 10	8	10	Yes
12	2 to 4	2	4	Yes
15	7 to 9	7	9	Yes
16	0 to 2	1	1	No
19	40.5 to 49.5	44	46	No
20	10.8 to 13.2	9.9	14.1	Yes
24	5 to 7	6	6	No
27	18.9 to 23.1	19.8	22.2	No

obtained. The solver required 833 variables, 1824 constraints and 23.28 seconds of cpu time to protect table 4.

Applying a SAS/OR implementation of the modified classical IP SDC model to only the initially exposed primary cells, $I = \{16, 19, 24, 27\}$, in table 4 the set of secondary cells $S = \{37, 38, 40\}$ was also obtained. The solver required 441 variables, 916 constraints and 12.26 seconds of cpu time to protect table 4.

5 Applying the Conjecture to the Classical IP Model for SDC

The cell suppression problem is the problem faced by statistical agencies when they release statistical tables, they must balance the risk of disclosing confidential information against the loss of information from the table caused by not publishing the suppressed cells in the table [3] [4] [5] [6] [8] [9].

Here we consider the case of a single external attacker who has no other knowledge than what is in the published table. It is usually assumed that the external attacker, prior to attack, knows that the cell a_i lies within the range from lb_i to ub_i. If the external attacker has no other knowledge than that published in the table then $lb_i = 0$ and $ub_i = \infty$. Kelly et al [6], when they defined the classical model, introduced a weighing w_i for each cell a_i to represent the information loss should the cell a_i be suppressed. A variable z_i was introduced for each a_i to indicate whether or not a_i had been suppressed ($z_i = 0$ means that a_i is published and $z_i = 1$ means that a_i is suppressed). Two tables where introduced that are consistent with $a = [a_1, ..., a_n]$, these tables $f^p = [f_1^p, ..., f_n^p]$ and $g^p = [g_1^p, ..., g_n^p]$ are used to calculate the lower and upper feasible limits for $p \in P$. In the classical model the lower and upper bounds (lb_i and ub_i) are translated into LB_i and UB_i, where $LB_i = a_i - lb_i$ and $UB_i = ub_i - a_i$. Those cells that are suppressed and are members of P are called primary suppressed cells and those cells that are suppressed but are not members of P are called secondary suppressed cells.

5.1 Classical Model

$$
\begin{array}{ll}
min \ \sum_{i=1}^{n} w_i z_i & \\
subject \ to & \\
\quad z_i \in \{0,1\} & for \ i = 1, ..., n \\
and \ for \ all \ p \in P : & \\
\quad \sum_{i=1}^{n} M_{ij} f_i^p = 0 & for \ j = 1, ..., m \\
\quad a_i - LB_i z_i \leq f_i^p \leq a_i + UB_i z_i & for \ i = 1, ..., n \\
\quad \sum_{i=1}^{n} M_{ij} g_i^p = 0 & for \ j = 1, ..., m \\
\quad a_i - LB_i z_i \leq g_i^p \leq a_i + UB_i z_i & for \ i = 1, ..., n \\
\quad f_p^p \leq lpl_p & \\
\quad g_p^p \geq upl_p & \\
\quad g_p^p - f_p^p \geq spl_p &
\end{array}
$$

5.2 Modified Classical Model

Applying conjecture in this paper we derived the Classic Model from Kelly et al as follows:

$$
\begin{array}{ll}
min\ \sum_{i=1}^{n} w_i z_i & \\
subject\ to & \\
\quad z_i \in \{0,1\} & for\ i = 1, ..., n \\
\quad z_p = 1 & for\ all\ p \in P \\
and\ for\ all\ p \in I(initially\ exposed\ primary\ cells): & \\
\quad \sum_{i=1}^{n} M_{ij} f_i^p = 0 & for\ j = 1, ..., m \\
\quad a_i - LB_i z_i \le f_i^p \le a_i + UB_i z_i & for\ i = 1, ..., n \\
\quad \sum_{i=1}^{n} M_{ij} g_i^p = 0 & for\ j = 1, ..., m \\
\quad a_i - LB_i z_i \le g_i^p \le a_i + UB_i z_i & for\ i = 1, ..., n \\
\quad f_p^p \le lpl_p & \\
\quad g_p^p \ge upl_p & \\
\quad g_p^p - f_p^p \ge spl_p &
\end{array}
$$

Table 6. Range of statistical tables with marginal totals

Table	Rows	Columns	Cells	Zeros	Primary Cells	Initially Exposed	Constraint Equations	Hierarchical
1	5	5	36	3	8	6	12	No
2	5	6	42	7	8	3	13	No
3	5	7	48	5	7	4	14	No
4	5	8	54	8	17	3	15	No
5	5	9	60	5	17	4	16	No
6	7	7	64	11	14	5	16	No
7	7	8	72	7	16	5	17	No
8	7	9	80	19	13	5	18	No
9	8	8	81	13	13	7	18	No
10	8	9	90	15	17	5	19	No
11	10	10	121	19	31	4	22	No
12	10	12	143	28	40	4	24	No
13	25	5	156	5	4	4	32	No
14	25	5	156	6	11	8	32	No
15	25	5	156	7	4	4	32	No
16	25	5	156	32	7	7	32	No
17	25	5	156	35	7	4	32	No
18	25	5	156	26	9	9	32	No
19	25	5	156	7	11	10	32	No
20	50	5	300	9	25	18	56	No

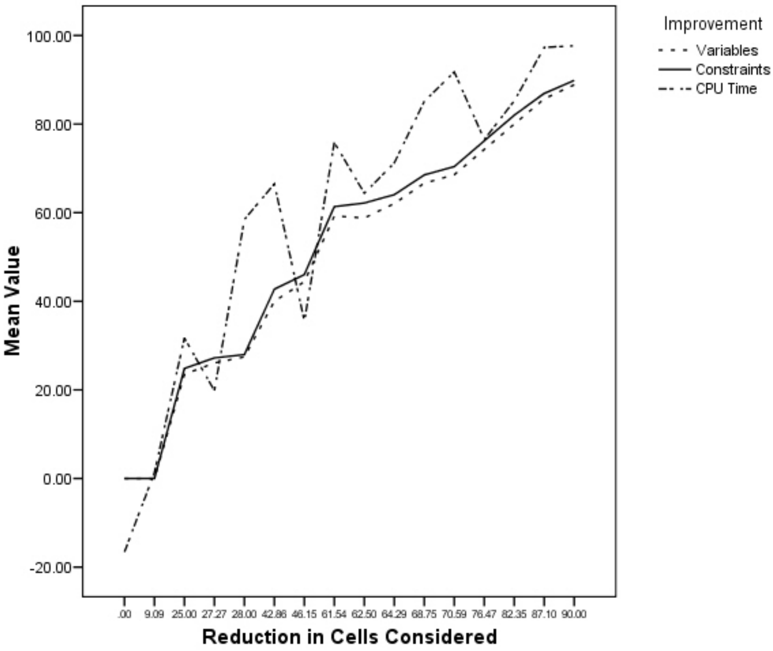


Fig. 1. Percentage Improvement in Number of Variables needed by SAS/OR, the Number of Constraints needed by SAS/OR and the CPU Time needed by SAS/OR by the Percentage Reduction in Primary Cells Considered.

6 Experimental Setup

6.1 Comparing the Classical and Modified Classical Models

A set of 20 2-dimensional non-hierarchical magnitude statistical tables with marginal totals (see Table 6) were generated for the purpose of comparing the classical and modified models [9]. These statistical tables with marginal totals were protected using a SAS/OR implementation of the classical model and a SAS/OR implementation of the modified (initially exposed primary cells only) classical model, using the same computer. These experiments were ran at ONS on a Dell Optiplex GX270 processor with 2GB RAM. The SAS version used was SAS 9 solver with SAS/OR Opt module. There are a variety of solvers in SAS and OptMILP was used. The selected secondary suppressed cells, the number of variables required, the number of constraints and the required cpu-time were recorded for comparison. For each of the statistical tables the percentage change in performance was calculated using the following formula.

$$ReductionInCellsConsidered = \frac{(SensitiveCells - InitiallyExposedCells) * 100}{SensitiveCells}$$

$$ImprovementInVariables = \frac{(ClassicalVariables - ModifiedVariables) * 100}{ClassicalVariables}$$

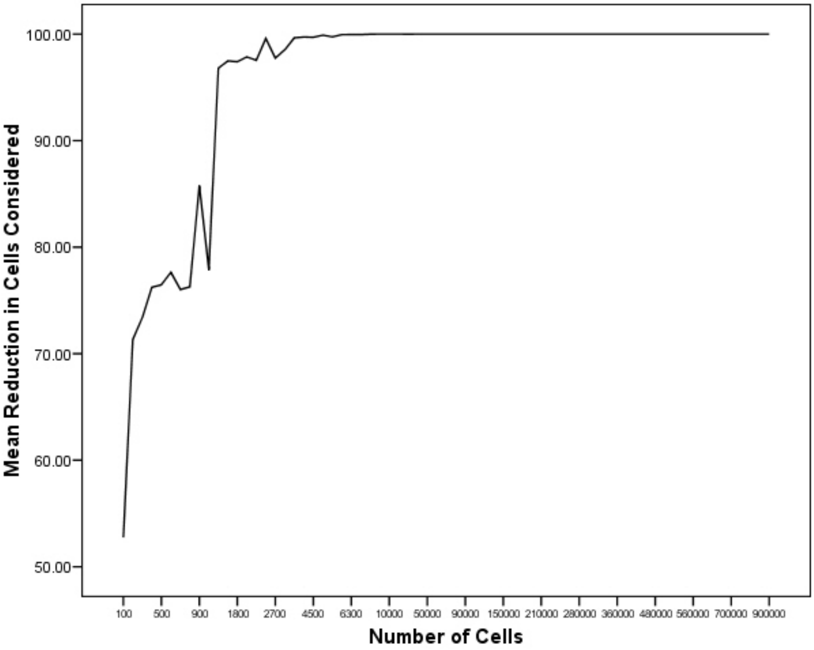


Fig. 2. The Percentage Reduction in Primary Cells Considered by the Number of Cells in the Table

$$ImprovementInConstraints = \frac{(ClassicalConstraints - ModifiedConstraints) * 100}{ClassicalConstraints}$$

$$ImprovementInCPUTime = \frac{(ClassicalCPUTime - ModifiedCPUTime) * 100}{ClassicalCPUTime}$$

For each of these statistical tables the improvement in the number of variables, constraints and cpu time was plotted against the reduction in the number of primary cells needing to be considered, see Fig. 1.

6.2 Estimating the Improvement for Different Table Sizes

A set of 3360 2-dimensional non-hierarchical statistical tables with marginal totals, sizes ranging from 100 cells to 900,000 cells, were generated with random values. For each different table size; 40 tables were generated, these tables had either 10% or 25% primary cells and either 10% or 20% of cells set to zero. For each of these tables the percentage reduction in the number of primary cells that need to be considered when using the modified classical model was plotted against the table size, see Fig. 2.

7 Results

7.1 Comparing the Classical and Modified Classical Models

Both models, classical and modified, selected the same secondary cells to suppress. The number of variables required, the number of constraints and the required cpu-time for each model is recorded in Table 7.

For every percentage reduction in the number of primary cells that need to be considered when using the modified classical model to protect a published statistical table there is an equal percentage improvement in the number of variables and constraints required to solve the associated linear programme. There is also a similar improvement in the required cpu time, however the relationship is not as smooth as it is for the number variables and constraints required, see Fig. 1. For those statistical tables where all of the primary cells are initially exposed, $P = I$, the modified classical model may require more cpu time than the classical model.

7.2 Estimating the Improvement for Different Table Sizes

The reduction in the number of primary cells that needed to be considered when using the modified classical model was affected by some of the properties of the

Table 7. Comparison of the two models

Table	Classical			Modified		
	Variables	Constraints	cpu-time (seconds)	Variables	Constraints	cpu-time (seconds)
1	612	1376	4.32	468	1034	2.95
2	714	1584	3.71	294	599	1.32
3	720	1568	8.17	432	899	2.43
4	1890	4250	4.07	378	764	0.6
5	2100	4692	4.17	540	1117	0.98
6	1856	4088	8.31	704	1469	2.39
7	2376	5216	31.07	792	1641	4.62
8	2160	4680	113.78	880	1808	27.48
9	2187	4732	38.23	1215	2554	24.65
10	3150	6834	84.81	990	2022	6.98
11	7623	16492	95.56	1089	2155	2.57
12	11583	24960	256.65	1287	2532	5.98
13	1404	2768	4.82	1404	2768	4.86
14	3588	7612	78.46	2652	5539	62.9
15	1404	2768	31.57	1404	2768	31.7
16	2340	4844	18.07	2340	4844	29.82
17	2340	4844	22.67	1404	2771	8.45
18	2964	6228	267.31	2964	6228	267.31
19	3276	6921	2.23	3276	6921	2.2
20	15300	32900	110.70	11100	23695	45.96

Table 8. Percentage Reduction in Primary Cells Considered, the Number of Variables needed by SAS/OR, the Number of Constraints needed by SAS/OR and the CPU Time needed by SAS/OR

Table	Reduction in Cells Considered	Improvement in Variables	Improvement in Constraints	Improvement in CPU Time
1	25	23.52	24.85	31.71
2	62.5	58.82	62.18	64.4
3	42.86	40	42.67	70.26
4	82.35	80	82.02	85.26
5	76.47	74.29	76.19	76.5
6	64.29	62.07	64.07	71.24
7	68.75	66.67	68.54	85.13
8	61.54	59.26	61.37	75.85
9	46.15	44.44	46.03	35.52
10	70.59	68.57	70.41	91.77
11	87.1	85.71	86.93	97.31
12	90	88.89	89.86	97.67
13	0	0	0	-0.83
14	27.27	26.09	27.23	19.83
15	0	0	0	-0.41
16	0	0	0	-65.02
17	42.86	40	42.80	62.73
18	0	0	0	0
19	9.09	0	0	1.35
20	28	27.45	27.98	58.48

statistical tables being protected. The reduction was greater for larger tables, tables that were more square than long and tables that had a higher proportion of primary cells. This is explained by each factor increasing the probability that more than one primary cell would occupy the same row or column and hence provide some protection to each other.

8 Conclusions

This pre-processing optimisation has been shown to be very effective when applied to the classical IP SDC model developed by Kelly et al [6]. This optimisation works by reducing the resources that the solver requires to protect statistical tables, hence allowing statistical tables to be protected quicker or allowing larger statistical tables to be protected. The classical IP SDC model has been implemented, as the Optimal Method, in the SDC tool, τ-Argus [5] [10]. It may be the case that this pre-processing optimisation could be applied to the τ-Argus Optimal Method to enable it to handle larger tables.

9 Further Research

Our current means of finding candidate members of I will find all the members of I, but will also include some members of C and N. Hence we need to find better techniques to identify members of I that do not require the use of a solver. As using the classical IP SDC model to protect large tables is still problematic we need to find other improvements that can be combined with this pre-processing optimisation. Our experiments involved statistical tables of various shapes and sizes; however an investigation needs to be carried out to see how the properties of the statistical tables affect the amount of improvement that this pre-processing optimisation provides. Our experiments did not include hierarchical statistical tables. How hierarchical statistical tables affect the amount of improvement that this pre-processing optimisation provides requires investigation. This pre-processing optimisation should be applied to other SDC techniques to see if similar performance improvements can be obtained.

References

1. Castro, J.: A shortest paths heuristic for statistical data protection in positive tables. INFORMS Journal on Computing 19(4), 520–533 (2007)
2. Clark, A., Smith, J.: Improvements to Cell Suppression in Statistical Disclosure Control. End-of-Project Report ONS Contract IT-06-0960A for the Office of National Statistics (ONS) (2006)
3. Fischetti, M., Salazar-González, J.J.: Solving the Cell Suppression Problem on Tabular Data with Linear Constraints. Management Science 47(7), 1008–1027 (2001)
4. Giessing, S.: Handbook on Statistical Disclosure Control, Version 1.01, ch. 4, CENEX SDC, a CENtre of EXcellence for Statistical Disclosure Control (2007)
5. Hundepool, A., van de Wetering, A., Ramaswamy, R., Wolf, P., Giessing, S., Fischetti, M., Salazar, J.J., Castro, J., Lowthian, P.: Tau-ARGUS Users Manual. CENEX-project. BPA no: 769-02-TMO (2007)
6. Kelly, J.P., Golden, B.L., Assad, A.A.: Cell suppression: Disclosure protection for sensitive tabular data. Networks 22, 28–55 (1992)
7. Salazar-González, J.J.: Extending Cell Suppression to Protect Tabular Data against Several Attackers. In: Domingo-Ferrer, J. (ed.) Inference Control in Statistical Databases. LNCS, vol. 2316, pp. 34–58. Springer, Heidelberg (2002)
8. Shlomo, N., Young, C.: Quality Measures for Statistical Disclosure Controlled Data. In: Proceedings of the European Conference on Quality in Survey Statistics (2006)
9. Staggemeier, A.T., Clark, A.R., Smith, J., Thompson, J.: Improving our knowledge of metaheuristic approaches for cell suppression problem. In: Joint UN-ECE/Eurostat work session on statistical data confidentiality, Manchester, United Kingdom, 17-19 December (2007)
10. Willenborg, L., de Waal, T.: Elements of Statistical Disclosure Control. Springer, New York (2001)

How to Make the τ-ARGUS Modular Method Applicable to Linked Tables

Peter-Paul de Wolf[1] and Sarah Giessing[2]

[1] Statistics Netherlands,
Department of Methodology and Quality,
P.O. Box 24500
2490 HA Den Haag, The Netherlands
PWOF@cbs.nl
[2] Federal Statistical Office of Germany,
65180 Wiesbaden, Germany
Sarah.Giessing@destatis.de

Abstract. The software package τ-ARGUS offers a very efficient algorithm for secondary cell suppression known as either HiTaS or the Modular approach. The method is well suited for the protection of up to 3-dimensional hierarchical tables. In practice, statistical agencies release multiple tabulations based on the same dataset. Usually these tables are linked through certain linear constraints. In such a case cell suppressions must obviously be coordinated between tables. In this paper we investigate into the possibilities for an extension of the modular approach to deal with linked tables.

1 Introduction

Some cells of the tabulations released by official statistics contain information that chiefly relates to single, or very few respondents which may often be easily identifiable. Therefore, traditionally, statistical agencies suppress part of the data, hiding some table cells from publication. Efficient algorithms for cell suppression are offered e.g., by the software package τ-ARGUS [6].

When tables are linked through simple linear constraints, cell suppressions must obviously be coordinated between these tables. A frequently occurring instance of a set of linked tables, consists of tables that share some of their marginal cells. E.g., tables specified in Eurostats SBS-regulation: a table on turnover broken down by 4-digit NACE, a second table on turnover broken down by 3-digit NACE and size class and a third table on turnover broken down by 2-digit NACE and geographic location (NUTS). These tables obviously have some marginal cells in common.

The intention of this paper is to present a collection of several alternative approaches for this coordination problem, and to give an idea of the issues that have to be considered for a decision which of the approaches (if any) should eventually get implemented in τ-ARGUS within the framework of a current joint European cooperation project.

J. Domingo-Ferrer and Y. Saygın (Eds.): PSD 2008, LNCS 5262, pp. 37–49, 2008.

Considering correctly the links between tables in the cell suppression process leads to a substantial increase in problem complexity. This tends to lead in turn to substantial increase in the amount of information that will be suppressed. One way of avoiding at least part of this increase may be to improve certain mechanisms in the current heuristics of the Modular method which cause overprotection in some situations. Section 4.1 proposes an idea for how to improve current heuristics.

2 Methodological Background

Statistical offices collect information on several properties that might be used for grouping respondents, like e.g. information about respondent economic activity (NACE) and geographic location. While with modern technologies it is no problem anymore to generate any kind of tables, or, by means of data-warehousing systems, to allow users to construct their own tabulations, solving the corresponding disclosure control problems consistently by means of secondary cell suppression can hardly be achieved in full generality because of the problem of coordinating suppressions across linked tables. [3] has described a special class of linked tables and presented an idea for an extension of the current methodologies to deal with sets of linked tables belonging to this class. This class of linked tables includes the linked tables as specified in the SBS-regulation of Eurostat.

The next subsection will first introduce some definitions and denotations which we will use throughout the paper with respect to hierarchical and linked tables structures. Subsection 2.2 briefly describes the original modular method (a.k.a. HiTaS). Finally, in subsection 2.3 we will discuss four possible approaches to deal with sets of linked tables.

2.1 Definitions and Denotations

In the terminology of tabular data statistical disclosure control, we think of an *m*-dimensional table as a tabulation of a certain continuous *response* variable by a cross combination of *m* categorical *spanning* variables.

[3] has introduced some denotation on hierarchical structures between the categories of spanning variables taken from graph theories. We follow this denotation in this paper and consider a hierarchy to be a rooted, directed tree, with the categories being the vertices of the tree. Additionally we define:

- A *relation* is a hierarchy consisting of only one father vertex and the corresponding child vertices.
- A table given as a cross combination of relations ($\mathcal{R}_1 \times \ldots \times \mathcal{R}_m$) is called a *simple table*. Note that this kind of table is often referred to as 'non-hierarchical' or 'unstructured'.
- If $\mathcal{G}$ is the covering hierarchy of a set of relations $\{\mathcal{R}_1,\ldots,\mathcal{R}_k\}$ then we say that $\{\mathcal{R}_1,\ldots,\mathcal{R}_k\}$ is a *simple breakdown* of $\mathcal{G}$.
- The *level* of a relation $\mathcal{R}_j$ in a simple breakdown of a hierarchy $\mathcal{G}$ is the level of the root of $\mathcal{R}_j$ in $\mathcal{G}$.

- Without loss of generality, we define the level of $\mathcal{R}_1$ to be 0. Note that for any $j > 1$ the root category of $\mathcal{R}_j$ is also a category of $\mathcal{R}_l$ for some $l \neq j$. We then say that $\mathcal{R}_j$ and $\mathcal{R}_l$ are *linked*.

Consider an m-dimensional table T given as ($\mathcal{G}_1 \times \ldots \times \mathcal{G}_m$). Let $\{\mathcal{R}_1^i, \ldots, \mathcal{R}_{k_i}^i\}$ denote the simple breakdown of $\mathcal{G}_i$. The *breakdown of table T into simple (sub)tables* is then given as the set $S(T)$ of up to m-dimensional simple tables $T_{j_1 j_2 \cdots j_m}$. Each of the simple tables $T_{j_1 j_2 \cdots j_m}$ is given as cross combination of relations $(\mathcal{R}_{j_1}^1 \times \mathcal{R}_{j_2}^2 \times \ldots \times \mathcal{R}_{j_m}^m)$ with $1 \leq j_i \leq k_i$. Because some of the $\mathcal{R}_{j_i}^i$ are linked, some of the tables $T_{j_1 j_2 \cdots j_m}$ in the set $S(T)$ are linked, i.e., they share identical cells.

2.2 The Original Modular Approach for Dealing with Hierarchical Tables

The disclosure risk connected to each individual cell of a table is assessed by applying certain sensitivity rules. If a cell value reveals too much information on individual respondent data, it is considered *sensitive*, and must not be published. We consider this to be the case, if the cell value could be used, in particular by any of the respondents, to derive an estimate for a respondent's value that is closer to the reported value of that unit than a pre-specified percentage p (this sensitivity rule is called the *p%* rule).

Cell suppression comprises of two steps. In a first step, sensitive cells will be suppressed (primary suppressions). In a second step, other cells (so called secondary suppressions) are selected that will also be excluded from publication in order to prevent the possibility that users of the published table would be able to recalculate primary suppressions. Naturally, this causes an additional loss of information.

By solving a set of equations implied by the additive structure of a statistical table, and some additional constraints on cell values (such as non-negativity) it is possible to obtain a *feasibility interval*, i.e., upper and lower bounds for the suppressed entries of a table, c.f. [4], for instance. A set of suppressions (the '*suppression pattern*') is called '*valid*', if the resulting bounds for the feasibility interval of any sensitive cell cannot be used to deduce bounds on an individual respondent's contribution that are too close according to the criterion employed to assess cell sensitivity. This requires that the bounds of the feasibility interval for any sensitive cell are at a 'safe' distance from its true value. Safe distance means that the distance exceeds the so called *protection level,* c.f. [7, 4.2.2].

The problem of finding an optimum set of suppressions known as the 'secondary cell suppression problem' is to find a feasible set of secondary suppressions with a minimum loss of information connected to it. The 'classical' formulation of the secondary cell suppression problem leads to a combinatorial optimization problem, which is computationally extremely hard to solve. For practical applications, the formulation of the problem must be relaxed to some degree.

The modular approach for hierarchical table cell suppression (also called HiTaS, see [2] for a detailed description) subdivides a hierarchical table T into the corresponding set $S(T)$ of simple, 'unstructured' linked (sub-)tables. The cell suppression problem is solved for each subtable separately. Within each subtable, methods based on Fischetti/Salazar Linear Optimization tools [4] are used to select secondary suppressions. For the co-ordination of secondary suppressions between linked subtables a backtracking procedure is used: the modular approach deals with the tables in $S(T)$ in a special, ordered way. During processing it notes any secondary suppression belonging also to one of the other tables. It will then suppress it in this table as well, and eventually repeat the cell suppression procedure for this table.

It must however be stressed, that a backtracking procedure is not global according to the denotation in [1]. See [1] for discussion of problems related to non-global methods for secondary cell suppression.

2.3 Extension of the Modular Approach for Dealing with Linked Tables

[3] presents an idea to extend the current methodologies to deal with a set of linked tables $\{T_1,\ldots,T_N\}$ that satisfy certain criteria. For instance, it is assumed that each table has a hierarchical structure that may differ from the hierarchical structures of the other tables. However, it is also assumed that tables that use the same spanning variables only have hierarchies that can be covered by a single hierarchy for that spanning variable. See [3] for the definition of a covering hierarchy. In essence it means that the covering hierarchy is such that all related hierarchies can be viewed as subhierarchies.

In the context of pre-planned table production processes which are typically in place in statistical agencies for the production of certain sets of pre-specified standard tabulations, it is normally no problem to satisfy these conditions. Literally speaking, the assumption is that tables in a set of linked tables may present the data in a breakdown by the same spanning variable at various amount of detail. But only under the condition that, if in one of the tables some categories of a spanning variable are grouped into a certain intermediate sum category, during SDC processing this intermediate sum category is considered in any other table presenting the data in a breakdown of the same spanning variable and at that much detail.

The idea of [3] is then as follows. For N tables $\{T_1,\ldots,T_N\}$ that need to be protected simultaneously, suppose that the specified tables contain M different spanning variables. Since the hierarchies are supposed to be coverable, an M-dimensional table exists having all the specified tables as subtables. The spanning variables will be numbered 1 up to M.

Each spanning variable can have several hierarchies in the specified tables. Denote those hierarchies for spanning variable i by $\mathcal{H}_1^i,\ldots,\mathcal{H}_{I_i}^i$ where I_i is the number of different hierarchies.

Define the M-dimensional table by the table with spanning variables according to hierarchies $\mathcal{G}_1,\ldots,\mathcal{G}_M$ such that, for each $i = 1,\ldots, M$ hierarchy $\mathcal{G}_i$ covers the set of hierarchies $\{ \mathcal{H}_j^i \}$ with $j = 1,\ldots, I_i$. This M-dimensional table will be called the cover table.

We will now describe several approaches to deal with this set of linked tables.

Complete Modular Approach

A straightforward approach would be to protect the complete cover table. HiTaS deals with all possible simple ('non-hierarchical') subtables of a hierarchical table in a specially ordered way. This would take care of all links between the tables in the set

$\{T_1,\ldots,T_N\}$, since by definition these tables are subtables of this cover table. This would result in a set of N protected tables $\{T_1^P,\ldots,T_N^P\}$. However, this approach considers a table structure which is much more complex than that of the tables which actually get published. We expect that this will tend to lead to a substantial increase in information loss compared to the other methods. Moreover, primary suppressions at low levels of a hierarchy often lead to secondary suppressions at the higher levels. Hence, unsafe cells in detailed tables that will not be published, might lead to secondary suppressions in less detailed tables that will be published. These secondary suppressions may be considered to be superfluous. This is probably not acceptable to users, given that the actual disclosure risk caused by ignoring subtables that are not foreseen for publication during disclosure control is likely to be rather low.

Adapted Modular Approach

The modular approach of HiTaS can easily be adapted. The idea is now basically to use the modular approach on the cover table T_C, but only consider those subtables that are also subtables of at least one of the specified tables $T_1,\ldots,T_N$ and disregard the other subtables. In the following we denote this subset of $S(T_C)$ (the breakdown of cover table T_C into subtables) as $S^*(T_C)$. This approach was suggested in [3] as well.

In τ-ARGUS the original modular approach is limited to hierarchical tables with up to three dimensions. This is mainly due to the fact that the Fischetti/Salazar Linear Optimization tools get too slow when dealing with higher dimensional tables. For the Adapted Modular Approach this restriction can be weakened. In theory there is no restriction on the number of dimensions of the cover table T_C, as long as each (sub)table that needs to be protected is at most three dimensional.

Linked Subtables Modular Approach

This somewhat more complex approach deals with sets of linked, simple subtables at a time. For each table T_i construct the set of linked subtables $S(T_i)$. Then consider the ordering used in HiTaS to order each set $S(T_i)$. Then deal with subtables from $S(T_i)$, …, $S(T_N)$ that are on the same order-level as linked tables using Fischetti/Salazar Linear Optimization tools. Such a set of linked subtables on the same order level is constructed in the following way: Let U and V be two simple subtables in $S^*(T_C)$. Assume U is a ν-dimensional (simple) table, where the first n spanning relations of U are not at level 0 of the corresponding covering hierarchies, i.e., U can be represented as $U := (\mathcal{R}_{i_1}^1 \times\ldots\times \mathcal{R}_{i_n}^n \times \mathcal{R}_1^{n+1} \times\ldots\times \mathcal{R}_1^{\nu})$. Then the subtable V belongs to the same set of linked subtables, if it is based on the same first n spanning relations as U, i.e., if it can be stated as $V := (\mathcal{R}_{i_1}^1 \times\ldots\times \mathcal{R}_{i_n}^n \times \mathcal{R}_1^{n+j} \times\ldots\times \mathcal{R}_1^{n+l})$ for some $M \geq l \geq j \geq 0$ where M is the dimension of the cover table T_C.

If all spanning relations of U are at level 0, i.e., $n = 0$, then the condition is that U and V share at least one level-0 spanning relation which can be expressed formally by requiring $j = 1$.

Traditional Approach

Although HiTaS cannot (yet) deal with linked tables, statistical agencies using HiTaS for secondary suppression of single tables must somehow solve the co-ordination problem. One possible approach is discussed in [7, 4.3.3]. This ‘traditional’ method is based on the idea of a backtracking procedure on the table level instead of on the subtable level.

In case of two linked tables T_1 and T_2, the approach would be as follows:

1. Protect table T_1 on its own;
2. Each cell in T_2 that is also present in T_1 will get the status (i.e., suppressed or not-suppressed) of the cell in the protected table T_1;
3. Table T_2, with the additional suppressions carried over in step 2, is protected on its own.
4. Each cell in T_1 that is also present in T_2 will get the status of the cell in the protected table T_2;
5. Repeat step 1 – 4 until no changes occur in protecting table T_1 nor in protecting T_2.

Graphically this would look like Figure 1.

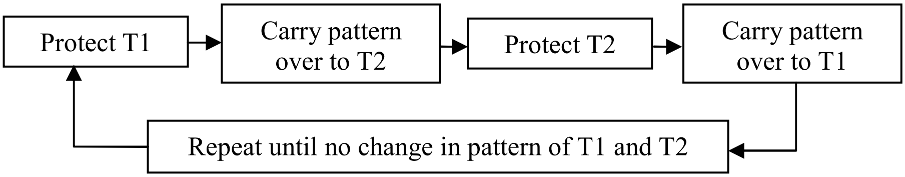


Fig. 1. Graphical representation of iteratively protecting two linked tables

Adding a third table to the set of linked tables, i.e., considering $\{T_1, T_2, T_3\}$, adds some complexity to this procedure. In that case several schemes can be thought of. E.g.,

a. Protect T_1, carry pattern over to T_2, protect T_2, carry pattern over to T_3, protect T_3, carry pattern over to T_1, repeat until no changes are added.
b. Protect T_1, carry pattern over to T_2, protect T_2, carry pattern over to T_1, protect T_1, repeat until no changes in $\{T_1, T_2\}$, carry patterns over to T_3, protect T_3, carry pattern over to T_1 and T_2, start with T_1 again. Repeat until no changes in T_1, T_2 and T_3.

The choice to be made may depend on the structure of the links between the tables T_1, T_2 and T_3. Obviously, the more linked tables need to be considered simultaneously, the more schemes can be constructed.

3 Illustrative Examples

In this section we will demonstrate the different approaches explained in the previous section using some instructive examples.

Example 1

We first consider a very simple instance of two linked tables. The specification of the two tables involves three spanning variables $\mathcal{F}$, $\mathcal{G}$ and $\mathcal{H}$, where $\mathcal{F}$ and $\mathcal{H}$ each consist of one relation only. The first table is given as $\mathcal{G}_2 \times \mathcal{F}$, the second table as $\mathcal{G}_1 \times \mathcal{F} \times \mathcal{H}$.

The simple breakdown of $\mathcal{G}_1$ consists of three relations, $\mathcal{R}_1^{\mathcal{G}}$, $\mathcal{R}_2^{\mathcal{G}}$ and $\mathcal{R}_3^{\mathcal{G}}$, where $\mathcal{R}_2^{\mathcal{G}}$ and $\mathcal{R}_3^{\mathcal{G}}$ are at level 1. The simple breakdown of $\mathcal{G}_2$ is given by ($\mathcal{R}_1^{\mathcal{G}}$, $\mathcal{R}_2^{\mathcal{G}}$, $\mathcal{R}_3^{\mathcal{G}}$, $\mathcal{R}_4^{\mathcal{G}}$, …, $\mathcal{R}_{20}^{\mathcal{G}}$) where $\mathcal{R}_4^{\mathcal{G}}$ to $\mathcal{R}_{10}^{\mathcal{G}}$ are at level 1 and $\mathcal{R}_{11}^{\mathcal{G}}$ to $\mathcal{R}_{20}^{\mathcal{G}}$ are at level 2. So $\mathcal{G}_1$ is a pure subhierarchy of $\mathcal{G}_2$, and therefore $\mathcal{G}_2$ is the covering hierarchy for variable $\mathcal{G}$.

The set $S^*(T_C)$ of subtables of the cover table T_C that are also subtables of T_1 or T_2, is then given by $\{\mathcal{R}_i^{\mathcal{G}} \times \mathcal{F} \times \mathcal{H}$, with $i = 1, \ldots, 4\} \cup \{\mathcal{R}_i^{\mathcal{G}} \times \mathcal{F}$, with $i = 5, \ldots, 20\}$.

According to the Adapted Modular Approach, we deal with these subtables successively within the usual backtracking strategy of HiTaS. The Linked Subtables Modular Approach is in this simple instance identical to the Adapted Modular Approach.

For the traditional approach we start with the first table $\mathcal{G}_2 \times \mathcal{F}$, use HiTaS for secondary suppression, carry the secondary suppressions from the area where both tables overlap, i.e., $\{\mathcal{R}_i^{\mathcal{G}} \times \mathcal{F}$, with $i = 1, \ldots, 4\}$, over to the second table $\mathcal{G}_1 \times \mathcal{F} \times \mathcal{H}$, do secondary suppression with HiTaS, and carry new secondary suppressions in $\{\mathcal{R}_i^{\mathcal{G}} \times \mathcal{F}$, with $i = 1, \ldots, 4\}$ over to the first table. In the instance the first table could then be processed successfully without selecting any new secondary suppressions in $\{\mathcal{R}_i^{\mathcal{G}} \times \mathcal{F}$, with $i = 1, \ldots, 4\}$ and thus the process finished successfully.

The results are summarized in Table 1. For a more detailed presentation of the results see Table 2 in the appendix.

The table $\mathcal{G}_2 \times \mathcal{F}$ contained 72 primary unsafe cells, 26 empty cells and 1343 cells in total. Table $\mathcal{G}_1 \times \mathcal{F} \times \mathcal{H}$ consisted of 4896 cells of which 657 cells were primary unsafe and 1055 were empty. The costs for suppressing a cell was defined to be a_i ^ 0.4 with a_i the cell value.

Table 1. Results of running three approaches to protect a set of two linked tables

Approach*	Number of secondary suppressions		Sum of costs of secondary suppressions	
	$\mathcal{G}_2 \times \mathcal{F}$	$\mathcal{G}_1 \times \mathcal{F} \times \mathcal{H}$	$\mathcal{G}_2 \times \mathcal{F}$	$\mathcal{G}_1 \times \mathcal{F} \times \mathcal{H}$
ModFull	96	709	8420	33330
ModAd	73	677	6528	31234
Trad	75	788	6440	34452

* ModFull = Complete Modular, ModAd = Adjusted Modular, Trad = Traditional.

In this (rather small) instance, with regard to the number of suppressions, the Adapted Modular (ModAd) approach outperforms the other two, i.e., the Complete Modular (ModFull) and the Traditional (Trad) approach. Table 2 (Appendix) presents the same results at some more detail by hierarchical level of the spanning variables. While the results of the traditional method on table $\mathcal{G}_2 \times \mathcal{F}$ are quite reasonable, the method performs especially bad at the second level of variable $\mathcal{F}$ in the $\mathcal{G}_1 \times \mathcal{F} \times \mathcal{H}$ table. This is perhaps a consequence of running $\mathcal{G}_2 \times \mathcal{F}$ first. Note that in this instance the disclosure risk for the suppression pattern provided by the traditional method is likely to be higher compared to patterns resulting from the other two approaches, because we decided to protect secondary suppressions carried over from the other table against exact disclosure only, assigning constant protection levels of 1 to those cells. This is probably also the reason why the sum of costs of secondary suppressions in $\mathcal{G}_2 \times \mathcal{F}$ is smaller for the traditional approach, compared to the adapted modular, even though the number of secondary suppressions is larger (73 vs. 75 cells). For more discussion on protection levels for secondary suppressions see 4.1.

Example 2

In this example, we add a third table to the two tables of example 1. This third table is given by $\mathcal{G}_2 \times \mathcal{H}$. For the cover table T_C of example 2 it holds:

$S^*(T_C) = \{\mathcal{R}_i^{\mathcal{G}} \times \mathcal{F} \times \mathcal{H}$, with $i = 1, \ldots, 4\} \cup \{\mathcal{R}_i^{\mathcal{G}} \times \mathcal{F}$, with $i = 5, \ldots, 20\} \cup \{\mathcal{R}_i^{\mathcal{G}} \times \mathcal{H}$, with $i = 5, \ldots, 20\}$.

The extended set of subtables involving some pairs (U, V) of linked subtables for the Linked Tables Modular Approach is then given by $S^{**}(T_C) = \{\mathcal{R}_i^{\mathcal{G}} \times \mathcal{F} \times \mathcal{H}$, with $i = 1, \ldots, 4\} \cup \{(\mathcal{R}_i^{\mathcal{G}} \times \mathcal{F}, \mathcal{R}_i^{\mathcal{G}} \times \mathcal{H})$, with $i = 5, \ldots, 20\}$.

For selecting secondary suppressions assigned earlier in the process in a table T_j to be carried over to a table T_i $(i \neq j)$ the areas $T_i \cap T_j$ have to be considered. Note that in each step (but only after processing the second table for the first time), we have to consider two of those areas, e.g., $T_1 \cap T_3$ and $T_2 \cap T_3$ for importing secondary suppressions to the third table. In our instance the overlap areas are $T_1 \cap T_2 = \{\mathcal{R}_i^{\mathcal{G}} \times \mathcal{F} \times \mathcal{H}$, with $i = 1, \ldots, 4\}$, $T_1 \cap T_3 = \{\mathcal{R}_i^{\mathcal{G}}$ with $i = 1, \ldots, 20\}$ and $\mathrm{T}_2 \cap \mathrm{T}_3 = \{\mathcal{R}_i^{\mathcal{G}} \times \mathcal{H}$, with $i = 1, \ldots, 4\}$.

Tables 3 and 4 in the appendix show the results of processing these 3 tables by the adapted modular and the traditional approach. The performance of the methods is similar as observed for instance 1: the adapted modular method gave superior results, especially for the 3-dimensional table.

Example 3

In this example we discuss the special case, where a table in a set of linked tables presents results for a subpopulation only, while other tables in the set present results on the full population. The instance this time consists of the two tables of example 1, $\mathcal{G}_1 \times \mathcal{F}$ and $\mathcal{G}_2 \times \mathcal{F} \times \mathcal{H}$ with a third table added, which is this time given by $\mathcal{G}_1 \times \mathcal{F} \times \mathcal{H}_5$. This third table presents data on the subpopulation falling into category 5 of hierarchy $\mathcal{H}$. We denote this special 'subhierarchy' of $\mathcal{H}$ consisting of only one category as $\mathcal{H}_5$.

Hierarchy $\mathcal{H}$ must then be extended, since we now consider two relations. The first one, $\mathcal{H}_1$, defines the partition of the full population into category 5 and a 'rest'-category consisting of all categories of $\mathcal{H}$ except category 5. The second one, $\mathcal{H}_2$, describes the partition of that 'rest' into the other categories. $\mathcal{H}^*$ denotes then the covering hierarchy of $\mathcal{H}_1$ and $\mathcal{H}_2$. We can now re-specify the second table as $\mathcal{G}_2 \times \mathcal{F} \times \mathcal{H}^*$ and the second and third table can be joined into one, e.g., $\mathcal{G}_1 \times \mathcal{F} \times \mathcal{H}_1$. In this way we avoid certain disclosure-by-differencing problems. For instance, if for a given category g^* of $\mathcal{G}$ only one respondent falls into the 'rest' category of $\mathcal{H}_1$, but the two cells specified by the two other categories of $\mathcal{H}_1$ (e.g., category 5 and the root-category) happen to be safe and remain unsuppressed. Such a problem will only be detected by a disclosure control process that explicitly considers the 'rest'.

On the other hand, in practice we sometimes deal with much more than just three linked tables. Taking into account correctly the relations between subpopulations considered for publication and subpopulations not considered for population usually adds to the complexity of the problem. In order to keep the effort for disclosure control processing within reasonable limits, in practice such subpopulation relations are often ignored. This can be justified, if the resulting disclosure risks are rather low, which is typically the case if the subpopulation of interest is comparatively small.

4 A Special Application of Partial Suppression Technology

In section 2.2 we have mentioned that on the level of individual simple sub-tables HiTaS uses methods based on Fischetti/Salazar Linear Optimization tools [4] to select secondary suppressions. In [5] the same authors propose a relaxed technique. The *complete cell suppression* method of [4] selects among all feasible suppression patterns the one with minimum information loss. This is modeled by associating a weight w_i with each cell i of the table and by requiring the minimization of the overall weight of the suppressed cells, e.g., it minimizes $\sum_{\{sup\}} w_i$ where $\{sup\}$ is the set of suppressed cells. The idea of the *partial cell suppression* methodology of [5] is, on the other hand, to compute intervals around the true cell values a_i, $[a_i - z_i^-, a_i + z_i^+]$, say. A set of intervals is considered as feasible, if the feasibility intervals for sensitive cells that could be computed taking into account the linear relations of the table and the intervals supplied by the partial suppression method cover certain pre-defined protection intervals. Based on the assumption that in a publication the true cell values would be replaced by these intervals, the loss of information associated to such a replacement is modeled as the size of the interval, i.e., $z_i^- + z_i^+$, or, in a more flexible way, as weighted linear combination of the deviation between interval bounds and true cell value $w_i^- z_i^- + w_i^+ z_i^+$. Seeking to minimize the overall information loss then means to minimize $\sum (w_i^- z_i^- + w_i^+ z_i^+)$.

Although the partial cell suppression problem is computationally much easier to solve compared to the complete cell suppression problem, and although a prototypical implementation for the partial suppression approach exists, in practice it has not yet been tested so far. Statistical agencies tend to be rather reluctant to replace traditional cell suppression by an interval publication strategy. Of course the strategy could be to suppress all cells where $z_i^- + z_i^+$ is non-zero. However, the set of these cells tends to be much larger then the set of cells suppressed as a result of the complete suppression approach.

In this paper we propose now a strategy to use partial cell suppression as complementary technique for complete cell suppression within the backtracking procedure of the modular approach. Obviously, we could also use this idea when we are carrying over suppression patterns between linked tables.

4.1 Using Partial Suppression to Compute Protection Levels

In 2.2 it was mentioned that a suppression pattern for a subtable is considered valid, only if the bounds of the feasibility interval for any sensitive cell are at a 'safe' distance from its true value, exceeding the protection level of that cell. Suitable protection levels are computed by τ-ARGUS according to [7, 4.2.2, table 4.2] and depend on the distribution of the individual contributions to a sensitive cell.

HiTaS deals with each subtable separately, carrying over secondary suppressions from overlapping subtables. A suppression pattern for a subtable T_a with secondary suppressions 'imported' from other subtables should be considered valid only, if the feasibility interval for a secondary suppression imported from, say, a subtable T_b does not jeopardize the protection provided to the sensitive cells of subtable T_b. Assume an intruder first computes the feasibility intervals for all suppressed cells of the subtable T_a. Later in his analysis, the intruder is assumed to consider these feasibility intervals as *a priori* bounds when computing feasibility intervals for suppressed cells of subtable T_b. Even with this additional *a priori* information, the resulting feasibility interval for the sensitive cells of subtable T_b should still be sufficiently wide. In the current implementation of HiTaS this issue is addressed by assigning protection levels to secondary suppressions computed by means of a simple heuristic. The following considerations aim at the development of a theoretically sound methodology to replace this heuristic.

Partial suppression provides us with a set of (smallest) intervals which can be published safely. This means that feasibility intervals for sensitive cells computed considering the partial suppression intervals as *a priori* bounds are sufficiently wide. It will therefore be enough to require for any suppression s in a subtable T_a which is an imported suppression in another subtable T_b that the feasibility interval for s computed on the basis of a suppression pattern for T_b covers the partial suppression interval of s in subtable T_a.

We therefore propose the following strategy:

(1) Compute a suppression pattern for subtable T_a using complete suppression.
(2) Compute a partial suppression pattern for subtable T_a where only cells suppressed in (1) are eligible for (partial) suppression.
(3) Assign the distances between the bounds of intervals given by the partial suppression pattern and the cell value of any suppressed cell s of T_a as protection level to s when protecting any other subtable T_b containing cell s.

Note that this strategy principally may require that protection levels of primary suppressions may have to be changed during processing.

The same strategy could also be used when dealing with the linked table settings of the current paper. E.g., in the traditional approach, the suppression pattern of the first table is carried over to the second (linked) table. If we then want to protect the second table, we will have to treat the complete suppression pattern as primary suppressions. Hence, we will need to specify safety ranges to each suppressed cell. We could use the above proposed strategy to calculate those safety ranges.

5 Summary and Final Conclusions

This paper has presented a few ideas for a backtracking algorithm that might be implemented in order to extend the τ-ARGUS Modular method, making it able to deal with sets of linked tables. We have used small illustrative examples for a first comparison of the algorithm properties. In this first comparison a method outlined in [3], here referred to as 'Adapted Modular Approach', gave promising results.

Another approach, denoted here as 'Linked Subtables Modular Approach' has certain theoretical advantages but is on the other hand more complex. Which of the two works better in practice is a question that we will be able to answer only after some testing on much larger datasets as we have used in this paper. For a decision on which algorithm to implement in a future version of τ-ARGUS the results of such a comparison should be considered.

The challenge of making the τ-ARGUS Modular method applicable to linked tables is, however, not only a matter of finding a good way for subtable construction and ordering sequences for the backtracking. To handle sets of linked tables means also to handle much larger datasets than just single tables. It means that more complex data structures are considered for disclosure control, and this tends to increase the information loss. In order to address these issues, in section 4 we have discussed partial cell suppression methodology. We have drafted a method to determine protection levels for secondary suppressions in a theoretically sound way using partial suppression methodology. This may eventually help to improve the performance of the Modular method regarding information loss.

Acknowledgements

This research was partially supported by the European Commission under grant agreement No 25200.2005.003-2007.670: 'ESSnet on statistical disclosure control'.

References

1. Cox, L.: Disclosure Risk for Tabular Economic Data. In: Doyle, L., Theeuwes, Z. (eds.) Confidentiality, Disclosure, and Data Access: Theory and Practical Applications for Statistical Agencies. North-Holland, Amsterdam (2001)
2. De Wolf, P.P.: HiTaS: A Heustic Approach to Cell Suppression in Hierarchical Tables. In: Domingo-Ferrer, J. (ed.) Inference Control in Statistical Databases. LNCS, vol. 2316. Springer, Heidelberg (2002)
3. De Wolf, P.P.: Cell suppression in a special class of linked tables. In: Joint ECE/Eurostat Worksession on Statistical Confidentiality in Manchester (December 2007), `http://epp.eurostat.ec.europa.eu/portal/page?_pageid=3154,70730193,3154_70730647&_dad=portal&_schema=PORTAL`
4. Fischetti, M., Salazar Gonzales, J.J.: Models and Algorithms for Optimizing Cell Suppression in Tabular Data with Linear Constraints. Journal of the American Statistical Association 95, 916 (2000)

5. Fischetti, M., Salazar Gonzales, J.J.: A Unified Mathematical Programming Framework for different Statistical Disclosure Limitation Mehtods. Operations Research 53(5), 819–829 (2005)
6. Hundepool, A., van de Wetering, A., Ramaswamy, R., de Wolf, P.P., Giessing, S., Fischetti, M., Salazar, J.J., Castro, J., Lowthian, P.: τ-ARGUS users's manual, version 3.1 (2006)
7. Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Lenz, R., Longhurst, J., Nordholt, E.S., Seri, G., De Wolf, P.-P.: CENEX handbook on Statistical Disclosure Control, CENEX-SDC project (2006), http://neon.vb.cbs.nl/cenex/CENEX-SDC_Handbook.pdf

Appendix

Table 2. Results for Instance 1: Number of suppressed cells for test-tables $\mathcal{G}_1\times\mathcal{F}\times\mathcal{H}$ and $\mathcal{G}_2\times\mathcal{F}$ by hierarchy levels[1]

Test-table	**Level spanning variable**		**ModAd**	**ModFull**	**Trad**
	$\mathcal{G}$	$\mathcal{F}$	# Secondary suppressions		
$\mathcal{G}_2\times\mathcal{F}$	1	1	0	0	0
	1	2	0	0	0
	2	1	0	0	0
	2	2	10	12	8
	3	1	0	0	0
	3	2	15	26	17
	4	1	0	0	0
	4	2	48	58	50
	All		73	96	75
$\mathcal{G}_1\times\mathcal{F}\times\mathcal{H}$	1	1	0	0	0
	1	2	3	3	3
	2	1	7	7	7
	2	2	270	292	357
	3	1	17	23	17
	3	2	355	346	379
	All		652	671	763

[1] For $\mathcal{G}_1\times\mathcal{F}\times\mathcal{H}$ statistics computed only for cells which are not in $\mathcal{G}_2\times\mathcal{F}$.

Table 3. Results for the set of the three linked tables of Instance 2

Approach	Number of secondary suppressions			Sum of costs of secondary suppressions		
	$\mathcal{G}_2 \times \mathcal{F}$	$\mathcal{G}_1 \times \mathcal{F} \times \mathcal{H}$	$\mathcal{G}_2 \times \mathcal{H}$	$\mathcal{G}_2 \times \mathcal{F}$	$\mathcal{G}_1 \times \mathcal{F} \times \mathcal{H}$	$\mathcal{G}_2 \times \mathcal{H}$
ModFull	96	709	101	8420	33330	8413
ModAd	73	710	77	6528	32178	6432
Trad	76	794	79	6484	34077	6158

Table 4. Results for Instance 2: Number of suppressed cells for test-tables $\mathcal{G}_2 \times \mathcal{F}$, $\mathcal{G}_1 \times \mathcal{F} \times \mathcal{H}$ and $\mathcal{G}_2 \times \mathcal{H}$ by hierarchy levels[2]

Test-	**Level spanning variable**			**ModAd**	**ModFull**	**Trad**
table	$\mathcal{G}$	$\mathcal{F}$	$\mathcal{H}$	# Secondary Suppressions		
$\mathcal{G}_2 \times \mathcal{F}$	1	1	1	0	0	0
	1	2	1	0	0	0
	2	1	1	0	0	0
	2	2	1	10	12	9
	3	1	1	0	0	0
	3	2	1	15	26	17
	4	1	1	0	0	0
	4	2	1	48	58	50
	All			73	96	76
$\mathcal{G}_1 \times \mathcal{F} \times \mathcal{H}$	1	1	2	0	0	0
	1	2	2	3	3	3
	2	1	2	7	7	8
	2	2	2	270	292	360
	3	1	2	16	23	18
	3	2	2	389	346	379
	All			685	671	768
$\mathcal{G}_2 \times \mathcal{H}$	4	1	2	54	71	53

[2] For $\mathcal{G}_1 \times \mathcal{F} \times \mathcal{H}$, statistics based only on cells which are not in $\mathcal{G}_2 \times \mathcal{F}$. For $\mathcal{G}_2 \times \mathcal{H}$, statistics based on cells neither in $\mathcal{G}_2 \times \mathcal{F}$, nor in $\mathcal{G}_1 \times \mathcal{F} \times \mathcal{H}$.

Bayesian Assessment of Rounding-Based Disclosure Control

Jon J. Forster[1] and Roger C. Gill[2]

[1] Southampton Statistical Sciences Research Institute, University of Southampton,
Southampton SO17 1BJ, UK
[2] Winton Capital Management, Oxford, UK

Abstract. In this paper, we consider how the security of a disclosure control mechanism based on randomised, but uncontrolled, rounding can be assessed by Bayesian methods. We develop a methodology, based on Markov chain Monte Carlo, for estimating the conditional (posterior) probability distribution for the original cell counts given the released rounded values. An effective rounding-based disclosure control will result in high posterior uncertainty about the true value. Conversely, a posterior distribution concentrated on a single value provides evidence of ineffective disclosure control.

1 Introduction

Armitage et al (2004) describe the disclosure risk issues associated with the UK Office for National Statistics' Neighbourhood Statistics project. This makes publicly available a number of key data sets, stratified by electoral ward. These data sets take the form of a series of margins (sometimes multiway) of a larger cross-classification. Armitage et al (2004) investigated a disclosure control mechanism based on rounding cell counts to a common base, using a stochastic mechanism. Each margin is rounded independently. Hence, the rounding is not *controlled* in the sense described by Cox (1987) where marginal subtotals are required to be mutually consistent.

One approach to assessing the resulting disclosure risk has been to compute upper and lower bounds on the true cell counts, based on the rounded counts. Details of this approach are provided by Armitage et al (2004). Where the difference between the upper and lower bounds is large, it might be concluded that significant uncertainty exists about the true cell counts and hence disclosure risk is low. Dobra et al (2003) point out that it is possible that, even where this difference is large, the data may be informative about the true cell count because most of the range between the bounds has a negligible probability of having generated the rounded data. Cox and Kim (2006) describe a variety of rounding approaches and, for a rounded individual cell, show that the (unbiased rounding)

J. Domingo-Ferrer and Y. Saygın (Eds.): PSD 2008, LNCS 5262, pp. 50–63, 2008.

approach considered in the current paper does provide differential information between cell counts between the bounds.

The aim of the current paper is to quantify more precisely the uncertainty about the true cell counts, given the rounded data, and hence to provide a more reliable assessment of disclosure risk. The approach is Bayesian, following the general framework advocated by Dobra et al (2003). Given the released rounded totals (data), we aim to provide a probability distribution for the true cell counts (parameters). Disclosure risk can then be directly assessed in terms of the probability that a given cell count can be determined to be in a sensitive range (typically zero or other small values). For an alternative approach to disclosure risk computation and control for marginal table release, see Barak et al (2007).

2 The Statistical Model

Let $\boldsymbol{x} = (x_1, \ldots, x_n)^T$ be the vector of true cell counts for a particular ward. Here $\boldsymbol{x}$ represents the **complete** cross-classification by all released variables, even if only certain margins are released. For example, if age (4 categories) and sex (2 categories) are released, either as individual margins, or as a cross-classification, or both, then $\boldsymbol{x}$ has 8 components.

We use the $p \times n$ matrix $\boldsymbol{D}$ to denote the mapping between the true cell counts and the true values of the released margins. For example, if $\boldsymbol{x}$ represents the 4×2 cross-classification by age and sex, then

$$\boldsymbol{D} = \begin{pmatrix} 0\,0\,0\,0\,1\,1\,1\,1 \\ 1\,1\,1\,1\,0\,0\,0\,0 \end{pmatrix}$$

corresponds to release of just the sex margin,

$$\boldsymbol{D} = \begin{pmatrix} 0\,0\,0\,0\,1\,1\,1\,1 \\ 1\,1\,1\,1\,0\,0\,0\,0 \\ 1\,1\,1\,1\,1\,1\,1\,1 \end{pmatrix}$$

corresponds to release of the sex margin and the overall total,

$$\boldsymbol{D} = \begin{pmatrix} 1\,0\,0\,0\,1\,0\,0\,0 \\ 0\,1\,0\,0\,0\,1\,0\,0 \\ 0\,0\,1\,0\,0\,0\,1\,0 \\ 0\,0\,0\,1\,0\,0\,0\,1 \\ 0\,0\,0\,0\,1\,1\,1\,1 \\ 1\,1\,1\,1\,0\,0\,0\,0 \\ 1\,1\,1\,1\,1\,1\,1\,1 \end{pmatrix}$$

corresponds to release of both margins and the overall total, and

$$
\boldsymbol{D} = \begin{pmatrix}
1&0&0&0&0&0&0&0\\
0&1&0&0&0&0&0&0\\
0&0&1&0&0&0&0&0\\
0&0&0&1&0&0&0&0\\
0&0&0&0&1&0&0&0\\
0&0&0&0&0&1&0&0\\
0&0&0&0&0&0&1&0\\
0&0&0&0&0&0&0&1\\
1&0&0&0&1&0&0&0\\
0&1&0&0&0&1&0&0\\
0&0&1&0&0&0&1&0\\
0&0&0&1&0&0&0&1\\
0&0&0&0&1&1&1&1\\
1&1&1&1&0&0&0&0\\
1&1&1&1&1&1&1&1
\end{pmatrix}
$$

corresponds to release of the 2 × 4 cross classification, both margins and the overall total. Hence p can be greater than or less than n.

The disclosure control mechanism takes the true value $\boldsymbol{Dx}$ of the margins to be released, and applies a random perturbation to obtain the rounded margins $\boldsymbol{y}$, for release. The mechanism is stochastic and is designed so that $E(\boldsymbol{y}) = \boldsymbol{Dx}$, where expectation is with respect to the perturbation mechanism, so no 'bias' is introduced.

We have examined the following perturbation mechanism, proposed by Nargundkar and Saveland (1972). Let b be a rounding base (assumed to be a small integer; in the examples below, we use $b = 5$), and let $\lfloor x \rfloor$ indicate the largest multiple of b which is less than or equal to x and $\lceil x \rceil$ indicate the smallest multiple of b which is greater than or equal to x. Then, the stochastic rounding mechanism has the following form

$$
y_i = \begin{cases} \lfloor (\boldsymbol{Dx})_i \rfloor \text{ with probability } 1 - \frac{1}{b}[(\boldsymbol{Dx})_i \bmod b] \\ \lceil (\boldsymbol{Dx})_i \rceil \text{ with probability } \frac{1}{b}[(\boldsymbol{Dx})_i \bmod b] \end{cases} \tag{1}
$$

where $a \bmod b = a - \lfloor a \rfloor$ and $y_1, \ldots, y_n$ are generated independently.

An alternative, but equivalent formulation for this rounding mechanism is

$$
y_i = \lfloor (\boldsymbol{Dx})_i + z_i \rfloor \tag{2}
$$

where z_i is an integer, uniformly distributed on $\{0, 1, \ldots, b-1\}$.

The likelihood for model (1) is given by

$$
f(\boldsymbol{y}|\boldsymbol{x}) = \prod_{i=1}^{p} \left[1 - \frac{1}{b}[(\boldsymbol{Dx})_i \bmod b]\right]^{I(y_i = \lfloor (\boldsymbol{Dx})_i \rfloor)} \left[\frac{1}{b}[(\boldsymbol{Dx})_i \bmod b]\right]^{I(y_i = \lceil (\boldsymbol{Dx})_i \rceil)} \times
$$

$$
I\left(y_i \in \{\lceil (\boldsymbol{Dx})_i \rceil, \lfloor (\boldsymbol{Dx})_i \rfloor\}\right) \tag{3}
$$

where the indicator function $I(\cdot)$ is equal to 1 if $\cdot$ is true and 0 otherwise. The term $I\left(y_i \in \{\lceil(\boldsymbol{Dx})_i\rceil, \lfloor(\boldsymbol{Dx})_i\rfloor\}\right)$ in each component of the product in (3) defines the bounds on which the method of Armitage et al (2004) is based.

Bayesian inference encapsulates the uncertainty about the unknown true cell counts $\boldsymbol{x}$, given the perturbed margins $\boldsymbol{y}$ by a posterior distribution $f(\boldsymbol{x}|\boldsymbol{y})$, given by Bayes' theorem as

$$f(\boldsymbol{x}|\boldsymbol{y}) \propto f(\boldsymbol{y}|\boldsymbol{x})f(\boldsymbol{x})$$

where $f(\boldsymbol{y}|\boldsymbol{x})$ is given by (3), and $f(\boldsymbol{x})$ is a *prior distribution* representing the uncertainty about $\boldsymbol{x}$ prior to obtaining the data $\boldsymbol{y}$.

Here, we choose a vague prior distribution for $\boldsymbol{x}$, representing a high level of uncertainty. We assume that, in the absence of observed data,

$$f(\boldsymbol{x}) = \frac{1}{k^n}\prod_{i=1}^{n} I(x_i \in \{1, \ldots, k\}. \tag{4}$$

In other words, we assume that the cell counts x_i are independently uniformly distributed between 1 and k where k is chosen to be large. Provided that k is larger than any bound likely to arise as a result of the rounding process, then the constraint that $x_i \leq k$ is irrelevant for practical purposes. Later, we describe a Bayesian approach where information available at higher geographical levels of aggregation may be incorporated into a more informative prior distribution for $\boldsymbol{x}$.

With this prior,

$$f(\boldsymbol{x}|\boldsymbol{y}) \propto \prod_{i=1}^{p}\left[1 - \frac{1}{b}[(\boldsymbol{Dx})_i \bmod b]\right]^{I(y_i = \lfloor(\boldsymbol{Dx})_i\rfloor)}\left[\frac{1}{b}[(\boldsymbol{Dx})_i \bmod b]\right]^{I(y_i = \lceil(\boldsymbol{Dx})_i\rceil)} \times \\ I\left(y_i \in \{\lceil(\boldsymbol{Dx})_i\rceil, \lfloor(\boldsymbol{Dx})_i\rfloor\}\right) \tag{5}$$

The posterior distribution (5) summarises uncertainty about the true cell counts $\boldsymbol{x}$, in light of rounded data $\boldsymbol{y}$. In particular, uncertainty about an individual cell count is summarised by its marginal distribution, for example

$$f(x_1|\boldsymbol{y}) = \sum_{x_2=1}^{k} \cdots \sum_{x_n=1}^{k} f(\boldsymbol{x}|\boldsymbol{y}). \tag{6}$$

Therefore, Bayesian disclosure risk assessment involves computing unnormalised joint (5) or marginal (6) probabilities for true cell counts, and then normalising. Cox and Kim (2006) give expressions for exact posterior probabilities for the case where all cells are independently rounded ($\boldsymbol{D}$ is an identity matrix). In principle, for arbitrary $\boldsymbol{D}$, (5) or (6) can be calculated for every $\boldsymbol{x}$ which satisfies the bounds

$$\prod_{i=1}^{p} I\left(y_i \in \{\lceil(\boldsymbol{Dx})_i\rceil, \lfloor(\boldsymbol{Dx})_i\rfloor\}\right) = 1 \tag{7}$$

which can be calculated using the method decribed by Armitage et al (2004). However, the number of such $\boldsymbol{x}$ can be very large. Even for the simple 2 × 4 example described below, there are 233 280 possible $\boldsymbol{x}$ and the number rapidly becomes infeasible for even moderate-sized examples.

3 Simulation Using the Metropolis-Hastings Method

An alternative approach is to generate a sample from $f(\boldsymbol{x}|\boldsymbol{y})$ and use sample proportions to estimate probabilities. It is possible to sample (approximately) from $f(\boldsymbol{x}|\boldsymbol{y})$ using only the unnormalised expression (5), by using the Metropolis-Hastings method. This method generates dependent observations from $f(\boldsymbol{x}|\boldsymbol{y})$ by simulating a Markov chain with equilibrium distribution $f(\boldsymbol{x}|\boldsymbol{y})$. For details of this and other Markov chain Monte Carlo (MCMC) methods, see Gamerman (1997).

Starting with an arbitrary $\boldsymbol{x}^0$ with $f(\boldsymbol{x}|\boldsymbol{y}) > 0$, we represent the generated sample by $\{\boldsymbol{x}^0, \boldsymbol{x}^1 \boldsymbol{x}^2, \ldots\}$ where $\boldsymbol{x}^{t+1}$ is generated from $\boldsymbol{x}^t$ by first proposing a value $\boldsymbol{x}^\star$ from an arbitrary proposal distribution. Then, the proposal is accepted $(\boldsymbol{x}^{t+1} = \boldsymbol{x}^\star)$ with probability

$$\alpha(\boldsymbol{x}^\star|\boldsymbol{x}^t) = \min\left\{1, \frac{f(\boldsymbol{x}^\star|\boldsymbol{y})q(\boldsymbol{x}^t|\boldsymbol{x}^\star)}{f(\boldsymbol{x}^t|\boldsymbol{y})q(\boldsymbol{x}^\star|\boldsymbol{x}^t)}\right\} \tag{8}$$

and rejected $(\boldsymbol{x}^{t+1} = \boldsymbol{x}^t)$ otherwise. Note that, as $f(\boldsymbol{x}|\boldsymbol{y})$ appears in both the numerator and denominator of (8), the normalising constant is not required and (5) can be used.

Subject to some technical conditions governing convergence (the most critical of which is irreducibility – roughly speaking whether any $\boldsymbol{x}$ satisfying (7) be reached from any other using the process) the distribution of $\boldsymbol{x}$ tends to $f(\boldsymbol{x}|\boldsymbol{y})$ and hence after discarding any initial unrepresentative burn-in iterations, the resulting sample can be used to summarise the posterior distribution (5) or (6).

All that is required to implement this approach is a proposal distribution which ensures irreducibility. We suggest a distribution which proposes modest perturbations to $\boldsymbol{x}$, through

$$\boldsymbol{x}^\star = \boldsymbol{x}^t + \boldsymbol{\epsilon} \tag{9}$$

where $\boldsymbol{\epsilon}$ has a discrete distribution. We chose the sample space for $\boldsymbol{\epsilon}$ to include some combination of

1. All vectors of the form $\boldsymbol{\epsilon}$=$(0, \ldots, 0, 1, 0, \ldots, 0)^T$ and $\boldsymbol{\epsilon}$=$(0, \ldots, 0, -1, 0, \ldots, 0)^T$. These moves correspond to adding or subtracting 1 from a cell count, leaving all other cell counts unchanged.
2. All vectors of the form $\boldsymbol{\epsilon} = (0, \ldots, 0, 1, 0, \ldots, 0, -1, 0, \ldots, 0)^T$. These moves correspond to moving an individual from one cell to another, leaving all other cell counts unchanged.
3. The subset of those vectors in 2 above where for at least one of the classifying variables the category is the same for the two cells whose counts are changed. Hence at least one one-way margin of the cross-classification is unchanged.

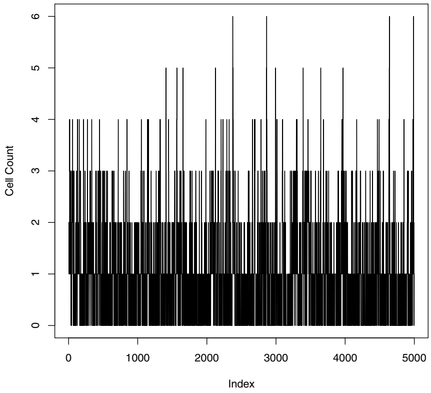


Fig. 1. Typical time series plot for Metropolis-Hastings approach

4. The sum of 2 vectors of the form of 2 or 3 above, including the restriction to cases which preserve at least one two-way margin of the cross-classification.

Provided that proposals are generated uniformly within each of the classes (1,2,3,4) above, then $q(\boldsymbol{x}^\star|\boldsymbol{x}^t) = q(\boldsymbol{x}^t|\boldsymbol{x}^\star)$ and (8) simplifies to

$$\alpha(\boldsymbol{x}^\star|\boldsymbol{x}^t) = \min\left\{1, \frac{f(\boldsymbol{x}^\star|\boldsymbol{y})}{f(\boldsymbol{x}^t|\boldsymbol{y})}\right\}. \tag{10}$$

The only potential difficulties are, in finding a starting value $\boldsymbol{x}^0$ with $f(\boldsymbol{x}^0|\boldsymbol{y})$ > 0 and in ensuring that the resulting algorithm is irreducible. A starting value can either be identified by directly evaluating the bounds using the method of Armitage et al (2004) or, if that is infeasible, by a stochastic search (applying successive proposal steps until a $\boldsymbol{x}^0$ with $f(\boldsymbol{x}|\boldsymbol{y}) > 0$ is identified). Irreducibility is a tricky issue, but we hope that a sufficiently rich class of proposals has been identified to ensure this, although we cannot guarantee this. We note that for tight bounds (unrounded margins), the proposals in 4 include those defined by Dobra (2003) as *primitive moves* and proved by him to constitute an irreducible (Markov) basis, when the released margins correspond to the sufficient statistics for the parameters of a decomposable graphical model. The rounded margins in the data releases we examined were all non-overlapping, and hence decomposable. It remains to show that the extra proposal types (1,2,3) are sufficient to allow transition between all possible alternative marginal configurations consistent with the released rounded totals.

In practice, therefore, our approach first chooses at random which of the classes of proposal 1-4 above to generate, then generates a proposed $\boldsymbol{\epsilon}$ uniformly from within that class, accepting the proposal with probability given by (10). This Metropolis-Hastings algorithm was applied to several example datasets (see Sections 5 and 6), and proved to be effective, and reasonably efficient at generating the required posterior distributions. Figure 1 is a typical plot of the time series of a cell count (x_i; here for Example 2; the plot has been 'thinned' so that only every 50th observation is plotted), illustrating good mixing with no signs of poor convergence.

4 Gibbs Sampler

An alternative MCMC approach for generating from $f(\boldsymbol{x}|\boldsymbol{y})$ is based on the alternative formulation (2) for the rounding process. Here, we consider the (unknown) perturbations $\boldsymbol{z}$ as part of our analysis and attempt to generate from the joint posterior distribution $f(\boldsymbol{z}, \boldsymbol{x}|\boldsymbol{y})$. To achieve this, we note that the conditional distributions $f(\boldsymbol{z}|\boldsymbol{x}, \boldsymbol{y})$ and $f(\boldsymbol{x}|\boldsymbol{z}, \boldsymbol{y})$ are straightforward to generate from and hence a Gibbs sampler (an MCMC approach which involves generating from conditional distributions) is immediately available. Starting from $\boldsymbol{x}^0$, we generate $\boldsymbol{z}^1$ from $f(\boldsymbol{z}|\boldsymbol{x}^0, \boldsymbol{y})$ and then $\boldsymbol{x}^1$ from $f(\boldsymbol{x}|\boldsymbol{z}^1, \boldsymbol{y})$. The method then proceeds by iterative updating $\boldsymbol{z}$ and $\boldsymbol{x}$ in this fashion. In fact the $\boldsymbol{x}$ are updated component by component with each cell count x_i being generated conditionally given the current values of the other cell counts.

Given $\boldsymbol{x}$, z_i is distributed uniformly on $\{\max\{0, y_i - (\boldsymbol{D}\boldsymbol{x})_i\}, \ldots, \min\{b-1, y_i - (\boldsymbol{D}\boldsymbol{x})_i + b - 1\}\}$. The conditional distribution of x_i given $\boldsymbol{z}$ and $x_j, j \neq i$ is uniform over a constrained region where the constraints are determined by examining those rows of $\boldsymbol{D}$ where the value in the ith column is greater than zero. For such a row, denoted $\boldsymbol{D}_j$, the corresponding constraint on x_i is derived from

$$y_j - z_j \leq \boldsymbol{D}_j\boldsymbol{x} \leq y_j - z_j + b - 1.$$

However, it is easy to construct an example where the Gibbs sampler is not irreducible. Suppose that just the two margins of a 2 × 2 table are released, both rounded to base 2, and that they are $(0, 0)$ and $(2, 2)$. The only possible tables which could have generated these margins are $(1, 0, 0, 1)$ and $(0, 1, 1, 0)$. However, transition between these two states is impossible using the Gibbs sampler as described above which only allows transitions which change a single x_i at a time. For this reason, we focus on the Metropolis-Hastings algorithm from now on.

5 Example 1: An Artificial Example

We focus on two examples. The first is a 2×4 table with rounded data available on the individual cell counts, the margins and the total. This is an artificial example used to illustrate the bounding method described by Armitage et al (2004). The rounded data, including the rounded margins and total are displayed in Table 1.

Table 1. Rounded 2 × 4 table, margins and total

5	10	0	5	25
5	20	5	0	30
0	25	15	5	50

Table 2. Bounds for Table 1

$[1, 3]$	$[6, 13]$	$[2, 4]$	$[1, 9]$
$[1, 3]$	$[16, 23]$	$[7, 9]$	$[0, 4]$

Table 3. >95% predictive probability intervals for Table 1

$[1, 3]$	$[8, 12]$	$[3, 4]$	$[4, 8]$
$[1, 3]$	$[16, 20]$	$[7, 9]$	$[0, 2]$

For this table, the bounds are displayed in Table 2 and predictive probability intervals of at least 95% are given in Table 3. For some cells, the width of the 95% intervals is around half that of the 100% probability intervals defined by the bounds.

Finally, the complete posterior distribution for all eight cells is displayed in Figure 2. The posterior distribution clearly provides considerably more information than is available from the bounds alone. Two distributions are displayed, obtained by complete enumeration and by MCMC. The distributions are almost identical, indicating that the MCMC approach is working correctly.

6 Example 2: Realistic Structure

The Neighbourhood Statistics project releases data on, amongst many other things, Income Support claimants. At the time this analysis was developed, the format of the released data could be summarised as in Table 4, which represents an imaginary ward.

Table 4. Income Support data for an imaginary ward – structure similar to Neighbourhood Statistics release

Total
10

Age					
< 20	20-29	30-39	40-49	50-59	≥ 60
0	0	5	0	0	10

Gender	
Male	Female
5	5

		Family		
		Single	Couple	
Age	< 60	5	0	—
	≥ 60	5	5	10
		10	0	10

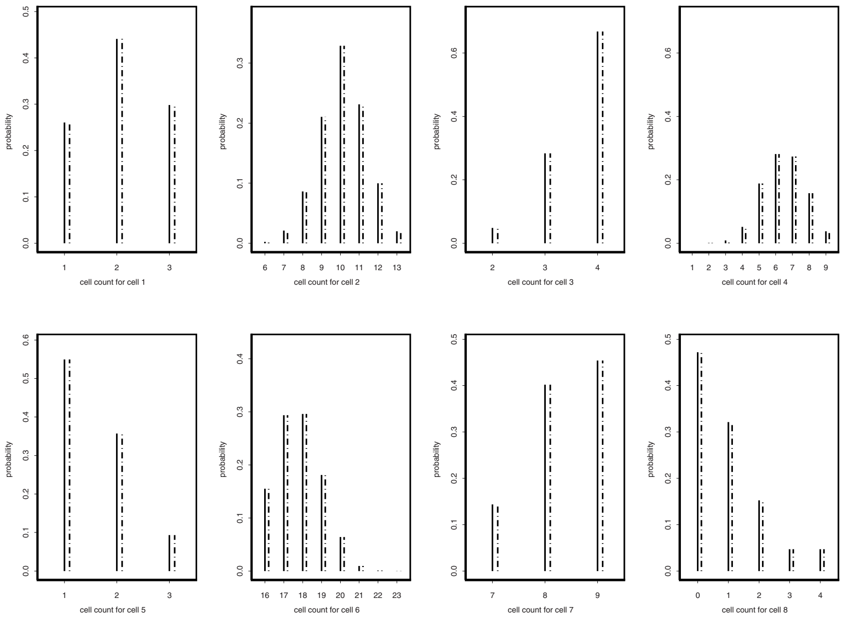


Fig. 2. Posterior distributions for the cells of Table 1, obtained by complete enumeration (solid line) and MCMC (dashed line)

The posterior distributions for the first 8 cells of the complete three-way cross-classification by age, gender and family make-up are displayed in Figure 3. The remaining cells are presented in Figure 4 in Appendix A. Note that each row of distributions in the figures corresponds to a particular agegroup, starting with Under 20 (cells 1–4) and ending with 60 and over (cells 21–24). Within each row, the cells are ordered Male/Single, Male/Couple Female/Single, Female/Couple, so, for example, cell 1 corresponds to under 20/Male/Single.

It can be seen immediately that for many cells there is much more information available from the posterior distributions than would be provided by the bounds alone. In particular there are eight cells for which the probability of a zero is greater than 0.9. Although the bounds indicate that the cell count in these cells could be as high as 4, the probability that it is greater than 1 is negligible (less than 0.5%). For these cells, the bounds give a potentially misleading impression of the disclosure protection provided by the rounding.

In this example, this behaviour can partly be attributed to the table being sparse. Hence, if *only* the rounded total (10) was released, the marginal posterior probability of any cell count being zero, based on the same prior, can be calculated exactly, to be 0.640. Hence, there is significant concentration of posterior probability at zero, due only to the fact that we know there are relatively few individuals distributed throughout a larger number (24) of cells.

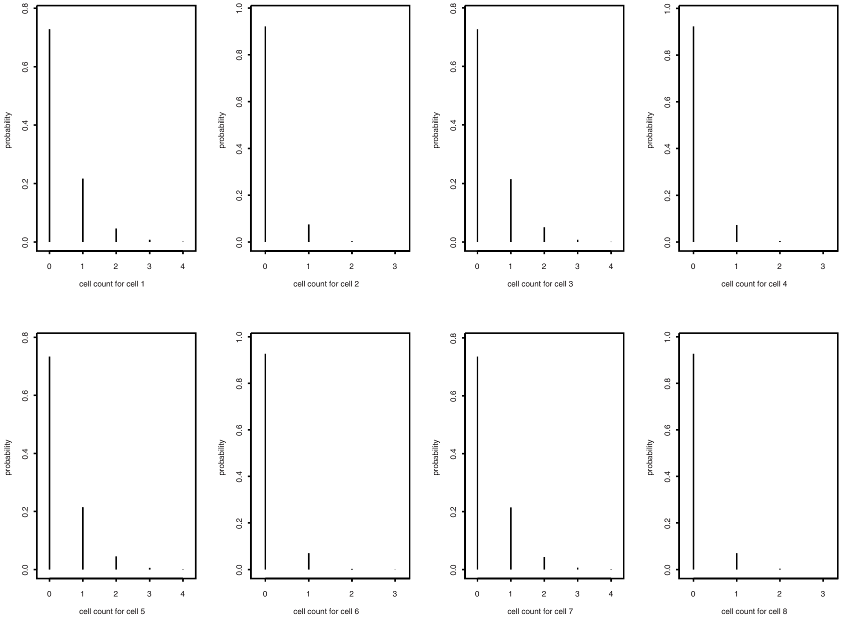


Fig. 3. Posterior distributions for Example 2 (cells 1 to 8)

Table 5. Maximum cell modal probability, by ward, for several wards

Ward	A	B	C	D	E	F	G	H	I	J
Probability	0.679	0.774	0.585	0.925	0.739	0.462	0.446	0.736	0.754	0.852
Ward	K	L	M	N	O	P	Q	R	S	
Probability	0.858	0.571	0.719	0.804	0.713	0.927	0.818	0.790	0.861	

We performed the same analysis for a number of actual wards. As a measure of how concentrated the posterior distribution can become, over a given cell, we calculated the modal (maximum) probability for each cell, and then extracted the most concentrated (highest mode) cell for each ward. The figures are displayed in Table 5. In each case the posterior distributions demonstrate greater concentration than would be implied by the bounds alone. An alternative measure of concentration of the posterior distribution in a cell, suggested by a referee, would be the ratio of the width of a 95% probability interval for a cell count, divided by the width of a 100% probability interval (difference between the bounds). For Table 1, the values of this measure range from 1 (no concentration) to 0.5 (an approximate halving of the plausible range). In practical examples, this reduction can be even greater.

7 Discussion and Extensions

Research on the basic approach described above was extended in two directions.

Total
1195

Age					
< 20	20-29	30-39	40-49	50-59	≥ 60
20	145	165	140	115	605

Gender	
Male	Female
825	365

		Family		
		Single	Couple	
Age	< 60	485	105	—
	≥ 60	525	80	605
		1010	185	1195

Table 6. Income support data for a district

7.1 Informative Prior

In Section 2, we used a vague, noninformative prior distribution for the true underlying cell counts $\boldsymbol{x}$, which essentially assumed that any pattern of counts was equally likely *a priori*. However, there is prior information available which we can use to construct a more realistic prior distribution. That information is the rounded data provided at higher geographical levels of aggregation. For example, for the data in Table 4, we could use the information provided in Table 6 at district level, to help determine which $\boldsymbol{x}$ are more or less likely. Although Table 6 has been rounded, the larger cell counts in the table mean that the effect of this rounding is negligible when considering relative cell proportions.

In fact, incorporating this prior information did not seem to greatly affect posterior inferences, so we only give brief details here. The prior distribution for $\boldsymbol{x}$, was based on the district data in the following hierarchical way. The x_i were assumed to have independent Poisson distributions with mean μp_i at the first stage. At the second stage μ was given an improper uniform distribution, and $\boldsymbol{p} = (p_1, \ldots, p_n)^T$ was given a prior density

$$f(\boldsymbol{p}) \propto \prod_{i=1}^{p} (\boldsymbol{D}\boldsymbol{p})_i^{\alpha_i}$$

where $\alpha_1, \ldots, \alpha_p$ reflect prior belief concerning the relative sizes of the released margins, obtained form the district level data. The prior density mimics a multinomial likelihood, with the α_i parameters representing 'prior counts' in the released margins. The overall magnitude of the α_i parameters reflects strength of prior belief. As we do not expect a ward to exactly reflect the district, the values of the α_i parameters are generally set to be smaller than the released district-level counts, but with the relative values preserved, at least approximately. If the α_i are given integer values, with consistent sums over overlapping margins, then computation with this prior is particularly straightforward. It can be set up as a missing data problem where the α_i are thought of as aggregated prior cell counts, with the actual prior cell counts included in a MCMC sampling scheme.

In all the examples we investigated, the more informative priors have little impact on the posterior inferences. We suspect that this is due to the fact that information is only available concerning margins, and that information about the interior of the table would need to be available for the prior to have a large impact.

7.2 Rounding to a Non-adjacent Base Multiple

In examples where bounds are too tight for adequate disclosure protection, allowing rounding to a non-adjacent multiple of base b, with some small probability, will result in wider separation of bounds. However, this will have little effect on the posterior distribution for the cell counts, as the following example illustrates.

We supposed that the rounding had been done according to the following modification of the scheme described in (1), where, with probability α the rounding is to a non-adjacent multiple of b, $\lceil(\boldsymbol{Dx})_i\rceil+b$ or $\lfloor(\boldsymbol{Dx})_i\rfloor-b$ (unless $0 \le (\boldsymbol{Dx})_i < b$ in which case rounding to a negative multiple is prohibited). In each case, the rounding is unbiased, as described in Section 2. Hence, a cell count of zero always remains at zero, even with this modification. The likelihood (3), posterior distribution and computational algorithm are modified accordingly.

We applied this method to the Example 2, using the same released data, but under the assumption that the modified rounding had been applied. We assumed values of $\alpha = 0$ (original approach), $\alpha = 0.05$ and $\alpha = 0.1$. Although the bounds are widened, as indicated by more cell counts having positive posterior probability, the total probability within this increased range is negligible for both $\alpha = 0.05$ and $\alpha = 0.1$.

7.3 Summary

We have shown that providing full posterior distributions for cell counts gives a more informative summary of disclosure protection than simply calculating bounds. The data are in the form of contingency tables, but relatively minor modification (allowing $\boldsymbol{\epsilon}$ to take non-integer values) would be required in order to apply the methodology described here to non-integer valued tables.

Our main analysis has been with a reference prior, but the general approach can easily be extended to incorporate other forms of prior information, for example concerning structural zeros, or as in Section 7.1. As the prior is supposed to incorporate a potential intruder's state of information, then assessing sensitivity to realistic intruder prior scenarios should be considered.

An obvious, and important, future extension would be to extend this methodology to controlled random rounding.

Acknowledgment

This work was carried out as part of the ONS-funded project ‘Neighbourhood Statistics: Statistical Disclosure Control for Rounded Frequency Tables’. The authors are grateful to Chris Skinner for helpful discussion and advice, and to two anonymous reviewers for suggestions which helped improve the paper.

References

1. Armitage, P., Merrett, K., Lyons, A., Tame, E.: Neighbourhood statistics in England and Wales: disclosure control problems and solutions. In: Monographs of official statistics Work session on statistical data confidentiality, Part 2, Luxembourg, April 7-9, 2003, pp. 239–249 (2004)
2. Barak, B., Chaudhuri, K., Dwork, C., Kale, S., McSherry, F., Talwar, K.: Privacy, accuracy, and consistency too: a holistic solution to contingency table release. In: Proceedings of the 27th ACM SIGMOD International Conference on Management of Data / Principles of Database Systems (2007)
3. Cox, L.H.: A constructive procedure for unbiased controlled rounding. Journal of the American Statistical Association 82, 520–524 (1987)
4. Cox, L.H., Kim, J.J.: Effects of rounding on the quality and confidentiality of Statistical data. In: Domingo-Ferrer, J., Franconi, L. (eds.) PSD 2006. LNCS, vol. 4302, pp. 48–56. Springer, Heidelberg (2006)
5. Dobra, A.: Markov bases for decomposable graphical models. Bernoulli 9, 1093–1108 (2003)
6. Dobra, A., Fienberg, S.E., Trottini, M.: Assessing the risk of disclosure of confidential categorical data. Bayesian Statistics 7, 125–144 (2003)
7. Gamerman, D.: Markov Chain Monte Carlo. Chapman and Hall, London
8. Nargundkar, M.S., Saveland, W.: Random rounding to prevent statistical disclosures. In: Proceedings of the Social Statistics Section, American Statistical Association, pp. 382–385 (1972)

Appendix A: Posterior Distributions for Cell Counts in Example 2 (cells 9 to 24)

Figure 4 presents the posterior distributions for the cell counts for cells 9-24 of Example 2.

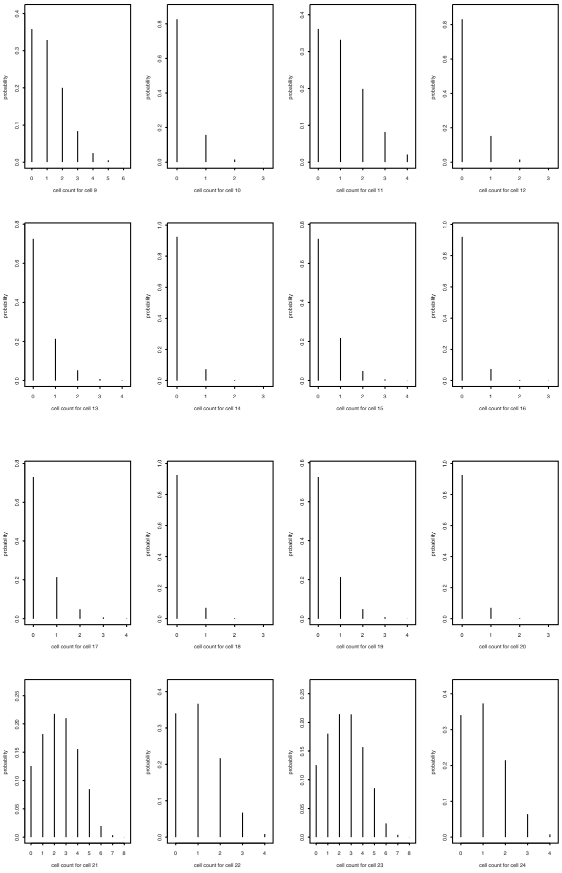


Fig. 4. Posterior distributions for Example 2 (cells 9 to 24)

Cell Bounds in Two-Way Contingency Tables Based on Conditional Frequencies

Byran Smucker and Aleksandra B. Slavković

Department of Statistics, Pennsylvania State University,
University Park, PA 16802, U.S.A.
bjs379@psu.edu, sesa@stat.psu.edu

Abstract. Statistical methods for disclosure limitation (or control) have seen coupling of tools from statistical methodologies and operations research. For the summary and release of data in the form of a contingency table some methods have focused on evaluation of bounds on cell entries in k-way tables given the sets of marginal totals, with less focus on evaluation of disclosure risk given other summaries such as conditional probabilities, that is, tables of rates derived from the observed contingency tables. Narrow intervals - especially for cells with low counts - could pose a privacy risk. In this paper we derive the closed-form solutions for the linear relaxation bounds on cell counts of a two-way contingency table given observed conditional probabilities. We also compute the corresponding sharp integer bounds via integer programming and show that there can be large differences in the width of these bounds, suggesting that using the linear relaxation is often an unacceptable shortcut to estimating the sharp bounds and the disclosure risk.

Keywords: Confidentiality; Contingency tables; Integer programming; Linear programming; Statistical disclosure control; Tabular data.

1 Introduction

Social or government agencies often collect data with intent to release a sufficient amount as public information that can be used for statistical inference, the results of which could affect policy decisions or further research. However, if too much information is released, confidentiality of individuals or organizations that has likely been guaranteed upon collection of the data could be compromised. Thus, there must be a trade-off between releasing the useful data and maintaining privacy.

Statistical disclosure limitation (SDL) deals with developments of methods and tools for evaluating trade-offs between disclosure risk and data usefulness. Many of the SDL methods developed in recent years lie at the interface of operations research and statistical methods; see a detailed review in [20]. There are many ways in which data confidentiality can be violated, as well as many ways to determine whether a violation has occurred. In this paper we are concerned with tabular data releases (e.g., marginal totals and conditional probabilities

J. Domingo-Ferrer and Y. Saygın (Eds.): PSD 2008, LNCS 5262, pp. 64–76, 2008.

with sample sizes) and the *feasibility interval* [25], that is, the bounds on a cell entry in a contingency table that can be induced by given released information. If these feasibility intervals are too narrow - or if the table is uniquely identified because the lower and upper bounds are the same - the risk of a disclosure could be high, particularly in cells with small counts.

We are particularly interested in the cell bounds that can be calculated when we are given observed conditional probabilities, that is, tables of rates derived from the observed table of counts. This is an important question because although many categorical data summaries are in the form of marginal tables, agencies often release rates or percentages representing proportions of individuals who fall in a certain category given some other characteristics (see [21], p. 7 for an example). Furthermore, the conditionals preserve association measures such as odds and odds-ratios relevant for data utility (e.g., see [21], [16]). We explore the question of what information about original cell counts can be extracted from knowing these conditional probabilities along with the sample size, and thus what is the effect on disclosure risk. This paper widens the statistical disclosure limitation literature by considering the effect of releasing a summary statistic - conditional probabilities - that heretofore has received little attention.

In this paper we calculate cell bounds given conditional probability information and sample size, using an integer/linear programming formulation. We improve upon the formulation proposed by Slavković and Fienberg [22] by requiring the marginals upon which we condition to be nonzero while allowing individual cells to be zero, and derive the closed-form solutions for the linear relaxation bounds on cell counts thus significantly reducing necessary computing time. This formulation actually produces somewhat wider bounds than those in [22], but is more realistic since it accommodates sampling zeros. In Section 2, we give some technical background on the optimization formulation and a brief review of the current results on calculation of cell bounds in contingency tables. In sections 3 and 4, we describe the linear and integer programming formulation for two-way tables and derive closed-form solutions for the linear relaxation, demonstrating them with two simple examples. We differentiate between two formulations depending if the calculation is done by a data owner or an intruder. We also compute the corresponding sharp integer bounds via integer programming and show that there can be large differences in the width of these bounds, suggesting that using the linear relaxation is often an unacceptable shortcut to estimating the sharp bounds and the disclosure risk for contingency tables given observed conditional frequencies.

2 Optimization Methods and Cell Bound Calculation for Contingency Tables

In this paper, we solve linear and integer programs in order to calculate cell bounds for two-way contingency tables given certain information. A linear program consists of a linear objective function, optimized subject to linear constraints. It can be represented in standard form as:

$$\begin{aligned} &\textit{Minimize} && \mathbf{cx} \\ &\textit{subject to} && \mathbf{Ax} = \mathbf{b} \\ & && \mathbf{x} \geq \mathbf{0} \end{aligned} \tag{1}$$

where there are n decision variables and m constraints, $\mathbf{c}$ is a row vector of length n, $\mathbf{x}$ is a column vector of length n, $\mathbf{A}$ is a $m \times n$ matrix, and $\mathbf{b}$ is a column vector of length m. An integer program can be formulated as (1) with the additional constraints that all decision variables be integer. Note that decision variables in this context are variables within an optimization program whose values are to be optimized; a random variable in the larger statistical context are variables whose values are determined by some random process (or, more technically, random variables are functions from a given sample space to the real numbers).

In the context of this paper, we use integer programming (IP) to calculate exact integer upper and lower bounds on entries in contingency tables. Integer programs are solved using methods such as Branch-and-Bound and Branch-and-Cut algorithms (see [18]), as in CPLEX, the commercial software [15] used in this work.

Calculating cell bounds for the entries of contingency tables given marginal totals has a long history, and goes back to Bonferroni [1], Fréchet [12], and Hoeffding [13] in their work on bounds for cumulative distribution functions given univariate marginals ([9], [10]). Given an $I \times J$ table with total sample size (n_{++}) and marginal totals (n_{i+} and n_{+j}), the *Fréchet bounds* have the following form for the ij^{th} cell:

$$min\{n_{i+}, n_{+j}\} \geq n_{ij} \geq max\{0, n_{i+} + n_{+j} - n_{++}\}.$$

Work has been done on generalizations of these bounds, i.e. bounds for k-way contingency tables. Given marginal totals for k-way tables, Dobra and Fienberg [6] give explicit formulas for the bounds when the table can be represented as a decomposable graph, a construct in which the expected counts in the cells of the table can be written as functions of the marginals. They extended these results to the case in which the graph is reducible, though when the table cannot be represented as a graph (which is often the case), other methods such as linear programming must be employed. The same authors in [7] further extended this idea to general k-way tables by generalizing the "shuttle algorithm" originally developed by Buzzigoli and Gusti [2] for three-way tables. This algorithm exploits hierarchical relationships within the table, and sequentially updates the bounds for cells until they cannot be further improved. A number of similar problems and approaches have been addressed in the context of statistical disclosure control but these have generally either successfully focused on two-way tables (e.g. [17], [2]) or broken down in higher-dimensional contexts (see [3]), though Cox [4] demonstrates that so-called network tables can overcome some of these problems.

Significantly less work has been done on examining bounds induced by given observed conditional probabilities. Researchers have begun to examine the cell bounds induced by conditional probabilities in conjunction with given marginals, as well as conditionals alone, using both mathematical programming (linear and integer) and tools from algebraic statistics such as Markov bases (e.g., see [21], [22], [11], and [5]). In this paper we only focus on the calculation of linear and integer bounds given conditionals alone, but offer an improved formulation with closed form bounds thus reducing the potential computational burden; one of the biggest criticisms of using Markov bases for these problems is in part computational inefficiency.

As mentioned above, a natural way to obtain sharp bounds given marginals and/or conditionals is via IP. Solutions to IP's can be difficult and computationally expensive, which may lead to the desire to use the linear relaxation bounds as an approximation to the sharp IP bounds. Given the marginals, the maximal gap between an IP and its linear relaxation has been studied and theoretically has been shown to be exponentially large ([24], [14]). These results imply that it could be misleading to assess disclosure risk by using the linear relaxation as an approximation to the sharp integer bounds. Onn [19] also showed that there could be arbitrary gaps in the bounds on cell entries given the margins, which could further increase the disclosure risk. In this paper, we show empirically that the same is true in the case of given conditional probabilities.

3 Bounds for Cells in Two-Way Tables Given Conditional Probabilities

In this section we consider $I \times J$ tables, using a simple 2×2 example to demonstrate the formulation of the integer and linear programming problems with result on cell bounds. Then, using this formulation we prove a theorem about the linear relaxation bounds for this situation. We assume a single, unweighted tabular data release.

3.1 Setting and Notation

Let X and Y be two random variables and $O = \{o_{ij}\}$ be the $I \times J$ table (matrix) of observed counts with sample size N. The joint probability distribution of these two random variables can be represented as $P = \{p_{ij}\}, i = 1, ..., I, j = 1, ..., J$, where $p_{ij} = P(X = i, Y = j)$ and $\sum_i \sum_j p_{ij} = 1$. Further, the marginal probability distributions for X and Y are $p_{i\cdot} = \sum_{j=1}^{J} p_{ij} = P(X = i)$ and $p_{\cdot j} = \sum_{i=1}^{I} p_{ij} = P(Y = j)$ respectively, and conditional probability distributions are $C = \{c_{ij}\}$ and $D = \{d_{ij}\}$ where $c_{ij} = \frac{p_{ij}}{p_{\cdot j}} = P(X = i|Y = j)$ and $d_{ij} = \frac{p_{ij}}{p_{i\cdot}} = P(Y = j|X = i)$ for $i = 1, ..., I, j = 1, ..., J$. Additionally, $\sum_i c_{ij} = 1$ and $\sum_j d_{ij} = 1$.

Note that these probability distributions involve true parameters, and under the assumption of multinomial sampling the observed counts are just estimators of those parameters. We are in particular interested in the estimated (observed) conditional probabilities and will represent them as $\hat{C} = \{\hat{c}_{ij}\}$ and $\hat{D} = \{\hat{d}_{ij}\}$ with $\hat{c}_{ij} = \frac{o_{ij}}{o_{\cdot j}}$ and $\hat{d}_{ij} = \frac{o_{ij}}{o_{i\cdot}}$.

As we have stated, the observed counts for the ij^{th} cell are represented by o_{ij} while in the following integer and linear programs, the decision variables (those variables which can be varied subject to constraints) used to define cell bounds are represented by n_{ij}. One can think of the observed counts as fixed (as they are a realization from the joint probability distribution P), while the n_{ij}'s can vary with relation to the optimization programs. In what follows, we focus on the case of given row conditionals, $\hat{D}$ and sample size, but similar statements can be derived for the column conditionals, $\hat{C}$.

3.2 Formulation of Optimization Problem for a 2 × 2 Table

We use a simple fictitious example (see [22]) to demonstrate the optimization setup. Suppose we have a sample of 25 male students and 25 female students and we ask them whether they have ever illegally downloaded mp3's on the internet. Thus X=gender and Y=illegally downloaded? with $i = 1, 2$ (male, female), $j = 1, 2$ (yes, no). These data are summarized in Table 1.

Table 1. Counts for 2-way Table

	Download Yes	Download No
Male	15	10
Female	5	20

From Table 1, using $\hat{d}_{ij} = \frac{o_{ij}}{o_{i1}+o_{i2}}$, we can calculate the following 2×2 matrix of row conditional probabilities; that is, the percentage of students downloading activity given gender:

$$\hat{D} = \begin{bmatrix} 0.6 & 0.4 \\ 0.2 & 0.8 \end{bmatrix}$$

Similarly, we can calculate $P(Gender|Download) = \hat{c}_{ij} = \frac{o_{ij}}{o_{1j}+o_{2j}}$:

$$\hat{C} = \begin{bmatrix} \frac{3}{4} & \frac{1}{3} \\ \frac{1}{4} & \frac{2}{3} \end{bmatrix}$$

Since in this case we have conditional probabilities that are repeating decimals, rounding becomes an issue which can affect the calculated bounds. We explore this issue in Section 3.3.

In most statistical models for contingency table data, each population parameter, p_{ij}, is assumed to be greater than zero (i.e. no structural zeros). However, for a given sample we can certainly observe a sampling zero. Because of this, instead of placing a lower bound of 1 on each cell (as in [22]), we make the lower bound zero and instead require that each margin have a count of at least one. This is necessary to satisfy the definition of conditional probability.

With this in mind, to calculate linear relaxation lower bounds on the ij^{th} cell counts in the original table based on the row conditionals (matrix $\hat{D}$), the following linear program is constructed:

$$Min\ n_{ij} \tag{2}$$
$$s.t.\ n_{11} + n_{12} + n_{21} + n_{22} = 50 \tag{3}$$
$$-\hat{d}_{12}n_{11} + \hat{d}_{11}n_{12} = 0 \tag{4}$$
$$-\hat{d}_{22}n_{21} + \hat{d}_{21}n_{22} = 0 \tag{5}$$
$$n_{11} + n_{12} \geq 1 \tag{6}$$
$$n_{21} + n_{22} \geq 1 \tag{7}$$
$$n_{ij} \geq 0, \quad \forall i, j \tag{8}$$

In the above linear program, the $\hat{d}_{ij}$'s are assumed known and calculated from the observed data O. The corresponding integer program to calculate exact bounds is formulated by simply including integer constraints on all decision variables. This formulation corresponds to *Example 2* in [22], except for the constraints given by equations (6) and (7) above. In fact, the same linear/integer program can be written with the observed cell counts, o_{ij}'s instead of $\hat{d}_{ij}$'s, by replacing equations (4) and (5) by

$$-o_{12}n_{11} + o_{11}n_{12} = 0 \tag{9}$$
$$-o_{22}n_{21} + o_{21}n_{22} = 0. \tag{10}$$

This formulation with the original counts has not been considered before, but it is important for providing feasible IP solutions and a more precise assessment of the disclosure risk by the data owner. We discuss this further below and in Section 4.1.

To calculate lower bounds for each cell, we solve four optimization problems, each one having a different cell in the objective function. To calculate upper bounds on each cell, we solve the same four optimization problem but maximize the objective function instead of minimize. The results of the integer program are listed in Table 2. These bounds using the improved formulation are actually the same as calculated in [22], although this may not be the case in general.

Similarly, we can calculate the sharp integer bounds for the $\hat{C}$ conditionals (see Table 3). Because the $\hat{C}$ conditionals require rounding, the IP often gives infeasible solution, and these integer bounds can only be calculated if the original data are given. Thus the agency can calculate the exact IP bounds using o_{ij}'s while an intruder, given no other external information, can only calculate the LP-relaxation bounds (see Table 3). In the next section, we present the closed form solution for the LP-relaxation bounds, and their implications for disclosure.

3.3 Exact Formulas for Linear Relaxation Bounds Given Conditional Probabilities

For the linear relaxation as we have formulated it in (2)-(8), notice that the lower bounds for each cell in Table 2 are equal to the conditional probability for that cell. We prove this along with the closed form solution for the upper bounds for $I \times J$ tables, and display the results of a simple calculation for the *mp3* data. In [22], the LP lower bounds were some integer or real-valued number greater than or equal to 1, but there were no closed form solutions. These new results can be extended to k-way contingency tables, as we show in [23], since the linear program associated with it has the same form.

Table 2. IP and LP Results for 2-way Table for $\hat{D}$ conditionals

	Download Yes	Download No
Male	[3,27], [0.6,29.4]	[2,18], [0.4,19.6]
Female	[1,9], [0.2,9.8]	[4,36], [0.8,39.2]

Recall that I is the number of categories in the first variable, J is the number of categories in the second, and N is the total sample size. We prove *Theorem* 1 in the case of our 2×2 example. Any other size of contingency table would have a linear program with the same structure, and could be proved similarly.

Theorem 1. *Assume we have an $I \times J$ contingency table, and none of the rows in the contingency table sum to zero. Based on the conditional probabilities $P(Y = j|X = i)$ and the sample size N, we can construct a linear program of the form* (2)-(8)*. This linear program is minimized when $n_{ij} = \hat{d}_{ij}$, and maximized for the ij^{th} cell at $(N - (I - 1))\hat{d}_{ij}$.*

Proof. For the lower bound, note that the lower bound for n_{ij} cannot be zero (unless $\hat{d}_{ij} = 0$, for which the result holds), because if it were, the other cell which defines its conditional distribution would be forced to zero by (4) or (5). This cannot happen because of constraints (6) and (7). Constraints (4) and (5) are derived from the the conditional probability relationship $\hat{d}_{ij} = \frac{n_{ij}}{\sum_j n_{ij}}$. Since (6) and (7) hold, and n_{ij} is minimized when its marginal is as small as possible, n_{ij} will be minimized when its marginal is 1, which forces n_{ij} to be precisely equal to its conditional probability, $\hat{d}_{ij}$.

For the upper bound, we maximize the objective function defined in (2). Since we are maximizing n_{ij}, the marginal total for each of the rows (beside the i^{th} row) in the contingency table will be as small as possible, namely 1, as required by constraints (6) and (7). This is possible because each of the cells can have a value equal to their conditional probability. Thus, for all but the i^{th} row, the marginal total is 1. So now there are $N - (I - 1)$ counts to distribute among the J cells in row i. Because constraints (4) and (5) are derived from the given conditional probabilities (i.e. $\hat{d}_{ij} = \frac{n_{ij}}{\sum_j n_{ij}}$), n_{ij} can be no larger than the value which satisfies $\frac{n_{ij}}{N-(I-1)} = \hat{d}_{ij}$ which means n_{ij} is maximized at $n_{ij} = (N - (I - 1))\hat{d}_{ij}$. □

To demonstrate the calculation for the first cell, just note that the lower bound is simply the associated $\hat{d}$ conditional probability, 0.6, and the upper bound is $(50-(2-1))*0.6 = 29.4$. The LP-relaxation bounds given are slightly wider than the IP bounds (Table 2); in this case, they seem to be a reasonable approximation. Although, we would argue that a more mathematically precise definition of "reasonable approximation" is needed.

Similarly, we can derive the result for given the observed column conditional, $\hat{C}$ and sample size. Now the bound would be: $\hat{c}_{ij} \leq n_{ij} \leq (N-(J-1))\hat{c}_{ij}$. For the $\hat{C}$ conditionals, we show in Table 3 the sharp integer bounds as well as the linear relaxations given conditional probabilities rounded to one and two decimal places. Notice the effect that rounding can have on the LP bounds.

Table 3. IP and LP results (rounded to one and two decimal places) for two-way table for $\hat{C}$ conditionals

	Download Yes	Download No
Male	[6,33], [0.8,39.2], [.75,36.75]	[2,14], [0.3,14.7], [0.33,16.17]
Female	[2,11], [0.2,9.8], [0.25, 12.25]	[4,28], [0.7,34.3], [0.67,32.83]

While the IP bounds calculated in this way give a more precise assessment of disclosure risk than their LP counterparts, it has been pointed out that the gaps exist even within the bounds; e.g., by using algebraic tools Slavković and Fienberg [22] showed that there are only four possible tables of counts satisfying these constraints. Agencies can use the rounding to release less precise values. This leads to some loss of utility but also to a gain in protection as the bounds become wider. The effect of such rounding on data utility and the number of possible tables is currently being explored by Lee and Slavković [16].

4 Example: Delinquent Children Data

In this section we consider a 4×4 table of counts originally used in [8] to demonstrate various statistical disclosure techniques for tabular data. Slavković and Fienberg [22] used this example to demonstrate the effect of released conditional frequencies in comparison to release of marginal totals, and utilized tools from computational algebra and Markov bases for the calculation of bounds and the number of tables. Table 4 shows the number of juvenile delinquents broken down by county and education level. Titles, row and column headings are fictitious.

Consider the case in which we are given the sample size, $N = 135$, as well as an estimate of $P(Education\ \ Level|County)$, that is $\hat{D}$. We can calculate the linear relaxation bounds immediately using Theorem 1. Slavković and Fienberg [22] calculated sharp integer bounds using Markov bases, and showed, at that time a surprising result, that there is only one table of counts that satisfies these released conditionals. We show here that the data owner does not need to use the algebraic tools but can get the same bounds and thus the same result by solving the integer program described below by using the observed counts.

Table 4. 4 × 4 Table. Delinquent Children Data and Integer Programing Bounds.

	Low	Medium	High	Very High
Alpha	15	1	3	1
Beta	20	10	10	15
Gamma	3	10	10	2
Delta	12	14	7	2

4.1 Formulation of Optimization Problems for 4 × 4 Example

Similar to Section 3.2, an integer program can be constructed as follows:

$$\begin{aligned} & Min\ n_{ij} && (11)\\ & s.t. \sum_i \sum_j n_{ij} = N \\ & \hat{d}_{ij} \sum_{k \neq j} n_{ik} + (\hat{d}_{ij} - 1) n_{ij} = 0,\ \forall i,\ j = 1,2,3 && (12)\\ & \sum_j n_{ij} \geq 1\ \forall i \\ & n_{ij} \geq 0\ \forall i, j \\ & n_{ij}\ integer\ \forall i, j \end{aligned}$$

where $\hat{d}_{ij}$ are elements of $\hat{D}$, and equation (12) is derived from the following:

$$\begin{aligned} \hat{d}_{ij} \sum_k n_{ik} - n_{ij} &= \hat{d}_{ij} n_{ij} - n_{ij} + \hat{d}_{ij} \sum_{k \neq j} n_{ik} \\ &= \hat{d}_{ij} \sum_{k \neq j} n_{ik} + (\hat{d}_{ij} - 1) n_{ij} = 0 \end{aligned}$$

If we know all but one of the conditional probabilities the last one is determined, and this eliminates four constraints of the original form: $\hat{d}_{ij} = \frac{n_{ij}}{\sum_k n_{ik}}\ \forall i, j$.

Because the decimal representations of the numbers in $\hat{D}$ must be rounded, this integer program is infeasible. However, if we consider the conditional probability in terms of the original data, we can construct an integer program that is feasible. Let $\hat{d}_{ij} = \frac{o_{ij}}{\sum_k o_{ik}} = \frac{n_{ij}}{\sum_k n_{ik}}$. Linearizing the second equality leads to:

$$\begin{aligned} 0 = o_{ij} \sum_k n_{ik} - \sum_k o_{ik} n_{ij} &= o_{ij} n_{ij} - \sum_k o_{ik} n_{ij} + o_{ij} \sum_{k \neq j} n_{ik} \\ &= o_{ij} \sum_{k \neq j} n_{ik} + (o_{ij} - \sum_k o_{ik}) n_{ij} \end{aligned}$$

$\forall i, j = 1, 2, 3$, where the o_{ij}'s are the observed cell counts in Table 4. By replacing constraints in equation (12) by the one above, we can calculate the sharp integer bounds on the cell counts. Note that the simplification of coefficients in the matrix assumes knowledge of the marginal distribution, and thus under the assumptions of this section would not be available to an intruder. Again, these bounds could only be calculated by the agency releasing the data.

Notice that in this example, the sharp integer bounds uniquely identify the original table (that is, the lower bound is equal to the upper bound). Slavković and Fienberg [22] showed the same result (for IP) but by using tools from algebraic geometry and Markov bases. This is also an extreme case of entry uniqueness problem which is related to the entry uniqueness given the margins (see [19]). Table 5 shows the linear relaxation results calculated using Theorem 1, rounding to one, two, and three decimal places, respectively. Note that the bounds given in [22] are uniformly narrower than the bounds presented here. However, this is the result of an unrealistic formulation which forces each cell to have a count of at least 1. When this constraint is relaxed, the wider bounds in Table 5 result.

It is evident, with this example as well, that rounding can have a significant effect on the bounds providing a "false" sense of disclosure risk since the bounds are much wider. However, notice that the cell with small counts do have short LP-relaxation bounds given the rounding at two or three decimal places, e.g. $o_{12} = [0.05, 6.6]$. The data owner would most likely decide in this case that these bounds are too tight and not release the conditional frequencies with the sample size, even without running the above described IP and knowing that there is only one possible table.

Table 5. Linear Relaxation Results for 4 × 4 Table (rounding to 1, 2, and 3 decimal places)

	Education Level			
County	Low	Medium	High	Very High
Alpha	[0.7,92.4], [0.75,99], [0.75,99]	[0.1,13.2], [0.05,6.6], [0.05,6.6]	[0.1,13.2], [0.15,19.8], [0.15,19.8]	[0.1,13.2], [0.05,6.6], [0.05,6.6]
Beta	[0.3,39.6], [0.37,48.84], [0.363,47.916]	[0.2,26.4], [0.18,23.76], [0.182,24.024]	[0.2,26.4], [0.18,23.76], [0.182,24.024]	[0.3,39.6], [0.27,35.64], [0.273,36.036]
Gamma	[0.1,13.2], [0.12,15.84], [0.12,15.84]	[0.4,52.8], [0.4,52.8], [0.4,52.8]	[0.4,52.8], [0.4,52.8], [0.4,52.8]	[0.1,13.2], [0.08,10.56], [0.08,10.56]
Delta	[0.3,39.6], [0.34,44.88], [0.343,45.276]	[0.4,52.8], [0.4,52.8], [0.4,52.8]	[0.2,26.4], [0.2,26.4], [0.2,26.4]	[0.1,13.2], [0.06,7.92], [0.057,7.524]

5 Conclusions

To date statistical disclosure limitation methodologies for tables of counts have been heavily focused on the release of unaltered marginal totals from such tables, and in part on inferences that are possible by an intruder from such releases. Many statistical agencies also release other forms of summary data from tables, such as tables of observed conditional frequencies. These are predominantly released as two-way and three-way tables, with conditioning on a single variable.

In this paper, we improved on the LP/IP formulation initially proposed in [22] by not restricting the counts in individual cells to be greater than one. While a zero marginal would result in a division by zero when calculating a conditional probability, there need not be any such restriction upon individual cells. The result is wider - though more realistic - bounds as well as closed-form solutions for the linear relaxation bounds thus reducing typically necessary optimization computing time. The proposed bounds hold even if there are observed zero cell counts. These zeros, however, may reveal extra information about their complementary cells and this requires some further careful investigation in particular for k-way tables. These new results can be extended to k-way contingency tables, as we show in [23], since the linear program associated with it has the same form.

Our improved formulation also circumvents the feasibility problem with calculation of sharp IP bounds given observed conditionals and sample size by calculating them using the observed counts directly, a fact relevant for data owners. The simple examples also show that IP may produce significantly narrower bounds than the linear relaxation of the same optimization problem. These large discrepancies can be seen especially in large and sparse tables k-way tables which we further explore in [23]. Because of these discrepancies and potential gaps within the IP bounds similar to the gaps described by [19] in the case of margins, the LP bounds often do not seem to be good approximation to the IP bounds. Thus these LP bounds may not be often reasonable for detecting whether there is a "true" potential disclosure, except perhaps as a crude approximation in the event of time-prohibitive sharp IP calculations. More precise mathematical definition of "reasonable" approximation is needed.

Note that in the 4×4 table (in addition to some other example we considered but not reported here), the sharp integer bounds given full conditional probabilities uniquely identify the counts in the original table. This occurred more often with smaller tables, but actually the most elementary example of all (the 2×2 table) did not yield a unique specification. At this point, we do not fully understand the underlying characteristics of a table that would produce a unique specification. There is some kind of tradeoff between the sample size and the number of cells, though in our examples the ratio between these two quantities certainly does not suggest anything obvious.

To further examine the relationship between the sample size and bounds given conditionals, and their effect on risk and utility, we are currently running simple simulations. For example, if we multiply each entry in Table 4 by 10, this has the effect of changing the sample size from 135 to 1350, increasing the width of IP and LP-relaxation bounds, and increasing the number of possible tables while maintaining the same conditional probabilities. Having the same conditionals is important for the utility aspect of SDL as they preserve certain associations within cell counts in the table. Also, in most of the datasets we analyzed (excluding the small example in Section 3.2) the IP based on the released conditionals proved infeasible because of rounding issues. Therefore, it is likely, without external information, that in practice releasing conditional probabilities would not allow intruders to calculate sharp integer bounds, but would give sufficient information for statistical inference.

Acknowledgments

The research reported here was supported in part by NSF Grant SES-0532407 to the Department of Statistics, Pennsylvania State University, and NSF Grant DMS-0439734 to the Institute for Mathematics and Its Application at the University of Minnesota.

References

1. Bonferroni, C.E.: Teoria statistica delle classi e calcolo delle probabilitá. Publicazioni del R. Instituto Superiore di Scienze Economiche e Commerciali di Firenze, 8 (1936)
2. Buzzigoli, L., Gusti, A.: An algorithm to calculate the upper and lower bounds of the elements of an array given its marginals. In: Statistical Data Protection (SDP 1998) Proceedings, pp. 131–147. Eurostat, Luxembourg (1998)
3. Cox, L.: Bounds on entries in 3-dimensional contingency tables. In: Domingo-Ferrer, J. (ed.) Inference Control in Statistical Databases. LNCS, vol. 2316, pp. 21–33. Springer, Heidelberg (2002)
4. Cox, L.: Contingency tables of network type: Models, markov basis and applications. Statistica Sinica 17, 1371–1393 (2007)
5. Dobra, A., Fienberg, S., Rinaldo, A., Slavković, A., Zhou, Y.: Algebraic statistics and contingency table problems: Log-linear models, likelihood estimation and disclosure limitation. In: Putinar, M., Sullivant, S. (eds.) IMA Volumes in Mathematics and its Applications: Emerging Applications of Algebraic Geometry, vol. 149, pp. 63–88. Springer, Heidelberg (2008)
6. Dobra, A., Fienberg, S.E.: Bounds for cell entries in contingency tables given marginal totals and decomposable graphs. Statistical Journal of the United Nations Economic Commission for Europe 18(4), 363–371 (2001)
7. Dobra, A., Fienberg, S.E.: Bounds for cell entries in contingency tables induced by fixed marginal totals. Statistical Journal of the United Nations ECE 18, 363–371 (2003)
8. Federal Committe on Statistical Methodology, Statistical Policy Working Paper 22 (Version Two). Report on Statistical Disclosure Limitation Methodology (2005)
9. Fienberg, S.E.: Fréchet and Bonferroni bounds for multi-way tables of counts with applications to disclosure limitation. In: Statistical Data Protection: Proceedings of the Conference, pp. 115–129. Eurostat, Luxembourg (1999)
10. Fienberg, S.E.: Contingency tables and log-linear models: Basic results and new developments. Journal of the American Statistical Association 95(450), 643–647 (2000)

11. Fienberg, S.E., Slavkovic, A.B.: Preserving the confidentiality of categorical statistical data bases when releasing information for association rules. Data Mining and Knowledge Discovery 11, 155–180 (2005)
12. Fréchet, M.: Les Probabilitiés Associées a un Système dÉvénments Compatibles et Dépendants, Vol. Premiere Partie. Hermann & Cie, Paris (1940)
13. Hoeffding, W.: Scale-invariant correlation theory. Schriften des Mathematischen Instituts und des Instituts fur Angewandte Mathematik der Universit at Berlin 5(3), 181–233 (1940)
14. Hosten, S., Sturmfels, B.: Computing the integer programming gap (2003), http://www.citebase.org/abstract?id=oai:arXiv.org:math/0301266
15. ILOG CPLEX, ILOG CPLEX 10.1 User's Manual. ILOG (2006)
16. Lee, J., Slavković, A.: Synthetic tabular data preserving the observed conditional probabilities. In: PSD 2008 (submitted, 2008)
17. Lu, H., Li, Y., Wu, X.: Disclosure analysis for two-way contingency tables. In: Domingo-Ferrer, J., Franconi, L. (eds.) PSD 2006. LNCS, vol. 4302, pp. 57–67. Springer, Heidelberg (2006)
18. Nemhauser, G.L., Wolsey, L.A.: Integer and Combinatorial Optimization. Wiley-Interscience (1988)
19. Onn, S.: Entry uniqueness in margined tables. In: Domingo-Ferrer, J., Franconi, L. (eds.) PSD 2006. LNCS, vol. 4302, pp. 94–101. Springer, Heidelberg (2006)
20. Salazar-Gonzalez, J.-J.: Statistical confidentiality: Optimization techniques to protect tables. Computers and Operations Research 35, 1638–1651 (2008)
21. Slavković, A.B.: Statistical Disclosure Limitation Beyond the Margins: Characterization of Joint Distributions for Contingency Tables. PhD thesis, Carnegie Mellon University (2004)
22. Slavković, A.B., Fienberg, S.E.: Bounds for cell entries in two-way tables given conditional relative frequencies. In: Domingo-Ferrer, J., Torra, V. (eds.) PSD 2004. LNCS, vol. 3050, pp. 30–43. Springer, Heidelberg (2004)
23. Smucker, B., Slavković, A.: Cell bounds in K-way tables given conditional frequencies. Journal of Official Statistics (to be submitted, 2008)
24. Sullivant, S.: Small contingency tables with large gaps. Siam J. Discrete Math. 18(4), 787–793 (2005)
25. Willenborg, L., de Waal, T.: Statistical Disclosure Control in Practice. Lecture Notes in Statistics III. Springer, New York (1996)

Invariant Post-tabular Protection of Census Frequency Counts

Natalie Shlomo[1] and Caroline Young[2]

[1] Univerity of Southampton, Highfield, Southampton, UK SO17 1BJ
N.Shlomo@soton.ac.uk
[2] Office for National Statistics, 1 Myddleton Street, London, UK EC1R 1UW
Caroline.Young@ons.gsi.gov.uk

Abstract. Some countries use forms of random rounding as a post-tabular method to protect Census frequency counts disseminated in tables. These methods typically result in inconsistencies between aggregated internal cells to marginal totals and across same cells in different tables. A post-tabular method for perturbing frequency counts is proposed which preserves totals and corrects to a large extent inconsistencies. The perturbation is based on invariant probability transition matrices and the use of microdata keys. This method will be compared to common pre and post-tabular methods for protecting Census frequency counts.

Keywords: Invariant probability transition matrix, Microdata keys, Additivity, Consistency.

1 Introduction

Protecting Census frequency tables containing whole population counts is a difficult problem for many Statistical Agencies who are under legal, ethical and moral obligations to minimize disclosure risk while meeting user demands for maximizing data utility. The main concern for disclosure risk in a Census context arises from small counts in tables, i.e. ones and twos, since these can lead to identity disclosure and potential attribute disclosure depending on the number and placement of zero cells in the rows/columns of the table. Statistical Disclosure Limitation (SDL) methods for Census tabular data should not only protect small cells in the tables but also introduce ambiguity and uncertainty into the zero cells.

We assume that Census tables, whether produced by the Statistical Agency or generated in flexible table building software, have undergone pre-determined SDL rules such as: minimal population and average cell sizes above thresholds, collapsing and fixing categories of variables spanning the tables, etc. In spite of these efforts to reduce disclosure risk, further and more invasive SDL methods are usually necessary. Common SDL methods that are typically implemented at Statistical Agencies for Census frequency counts include pre-tabular methods, post-tabular methods and combinations of both.

Pre-tabular methods are implemented on the microdata prior to the tabulation of the tables. The most commonly used method is record swapping between a pair of

J. Domingo-Ferrer and Y. Saygın (Eds.): PSD 2008, LNCS 5262, pp. 77–89, 2008.

households matching on some control variables (Dalenius and Reis, 1982, Willenborg and de Waal, 2001, Fienberg and McIntyre, 2004). This method has been used for protecting Census tables at the United States Bureau of the Census and the Office for National Statistics (ONS) in the United Kingdom. Record swapping can be seen as a special case of a more general pre-tabular method based on PRAM (Gouweleeuw, Kooiman, Willenborg and De Wolf, 1998, Willenborg and De Waal, 2001). The method adds "noise" to categorical variables by changing values of categories for a small number of records according to a prescribed probability matrix and a stochastic process based on the outcome of a random multinomial draw. PRAM can also be carried out in such a way as to ensure marginal distributions. In addition, since PRAM is a stochastic perturbation, users can make use of the probability transition matrix to correct for the measurement error purposely introduced into the data. In practice, Statistical Agencies prefer record swapping since the method is easy to implement and marginal distributions are preserved exactly on higher aggregations of the data

Post-tabular methods are implemented on the entries of the tables after they are computed and typically take the form of random rounding, either on the small cells of the tables or on all entries of the tables. The method of small cell adjustments (rounding) has been carried out on the Census tables at the Australian Bureau of Statistics (ABS) and the UK ONS, and full random rounding has been carried out at Statistics Canada and Statistics New Zealand. Within the framework of developing the SDL software package, Tau Argus, a fully controlled rounding option has been added (Hundepool, 2002, Salazar-Gonzalez, Bycroft, and Staggemeier, 2005). The procedure uses linear programming techniques to round entries up or down and in addition ensures that all rounded entries add up to rounded totals. Other post-tabular methods include cell suppression or some form of random perturbation on the cells of the Census tables (for example, the method used in the 1991 UK Census was to add $-1,0,1$ to each cell count in a table according to prescribed probabilities). Cell suppression is not typically used in a Census context because of the large number of tables that need to be consistently suppressed. Cell perturbation based on a stochastic mechanism produces similar results to record swapping but with inconsistencies in cells across tables and marginal totals.

Another method of SDL prominent in the literature for minimizing disclosure risk in frequency tables is to report only the marginal minimal sufficient statistics (see Dobra and Fienberg, 2003, Fienberg and Slavkovic, 2005 and references therein). While these methods may be good solutions when generating completely flexible outputs via web-based advanced query systems, they generally do not apply for standard frequency tables that are disseminated by Statistical Agencies. Indeed, Statistical Agencies will release key statistics at higher aggregations without any perturbation applied at all. Disclosure risk is typically managed by fixing the categories of geographies and other variables in order to avoid the disclosure risk that occurs from differencing and linking tables, introducing ambiguity in zero cells and masking small cells.

Since more invasive SDL methods are needed to protect against disclosure risk in a Census context, this has a negative impact on the utility of the data. It is well known that Census data have errors due to data processing, coverage adjustments, non-response and edit and imputation procedures, although much effort is devoted to minimizing these errors. When assessing disclosure risk, it is essential to take into account measurement errors and the protection that is already inherent in the data. For example, a quantitative measure of disclosure risk should take into account the amount of imputation and adjust parameters of the SDL methods accordingly to be inversely proportional to the imputation rate. This ensures that the data is not overly protected causing unnecessary loss of information. It should also be noted that once Census results are disseminated, they are generally perceived and used by the user community as accurate counts.

In this paper, we propose a new SDL method for protecting Census frequency counts which combines some of the characteristics and good qualities of the pre and post-tabular methods mentioned above. Section 2 describes the method. Section 3 compares the method to other common SDL methods with respect to disclosure risk and data utility. The analysis will be carried out on a real table from the UK 2001 Census. We conclude with a brief discussion in Section 4.

2 Invariant Post-tabular SDL Method

In developing a method for protecting Census frequency counts we need to ensure the following properties:

1. Additivity – all internal cells add up to marginal totals.
2. Consistency – internal cells and totals appearing across different tables are the same.
3. Reduce the risk of being able to deduce cell values through linking and differencing tables.
4. Preserve stochastic properties – cells are perturbed using a stochastic process, statistical properties are preserved and information about the perturbation can be disseminated to users to take into account in their analysis.

We consider a combination of pre and post-tabular methods to develop a new method that will follow the above guiding principles.

2.1 Invariant Probability Matrices

The perturbation of internal cells of the frequency table will be carried out using an invariant probability transition matrix, similar to the method that is used in PRAM. Let $\mathbf{P}$ be a $L \times L$ transition matrix containing conditional probabilities:

$$p_{ij} = p(\text{perturbed cell value is } j \mid \text{original cell value is } i)$$

for cell values from 0 to L (a cap is put on the cell values and any cell value above the cap would have the same perturbation probabilities). Let $\mathbf{t}$ be the vector of frequencies of the cell values where the last component would contain the number of cells above cap L and $\mathbf{v}$ the vector of relative frequencies: $\mathbf{v} = \mathbf{t}/n$, where n is the number of cells in the table. In each cell of the table, the cell value is changed or not changed according to the prescribed transition probabilities in the matrix $\mathbf{P}$ and the result of a draw of a random multinomial variate u with parameters q_{ij} $(j = 1,2,...L)$. If the *j*-th value is selected, value *i* is moved to value *j*. When $i = j$, no change occurs.

Let $\mathbf{t}^*$ be the vector of the perturbed frequencies. $\mathbf{t}^*$ is a random variable and $\mathrm{E}(\mathbf{t}^* \mid \mathbf{t}) = \mathbf{tP}$. Assuming that the probability transition matrix $\mathbf{P}$ has an inverse $\mathbf{P}^{-1}$, this can be used to obtain an unbiased moment estimator of the original table: $\hat{\mathbf{t}} = \mathbf{t}^*\mathbf{P}^{-1}$. Statistical analysis can be carried out on $\hat{\mathbf{t}}$. In order to ensure that the probability transition matrix has an inverse and to control the amount of perturbation, the matrix $\mathbf{P}$ typically is dominant on the main diagonal, i.e. each entry on the main diagonal is over 0.5.

Place the condition of invariance on the transition matrix $\mathbf{P}$, i.e. $\mathbf{tP} = \mathbf{t}$. This releases the users of the perturbed table of the extra effort to obtain unbiased moment estimates of the original data, since $\mathbf{t}^*$ itself will be an unbiased estimate of $\mathbf{t}$. The property of invariance means that the expected values of the marginal distribution of the cell values being perturbed are maintained and we ensure no bias in the total.

To obtain an invariant probability transition matrix, the following two stage algorithm is given in Willenborg and De Waal (2001): Let $\mathbf{P}$ be any probability transition matrix: $p_{ik} = p(c^* = k \mid c = i)$ where c represents the original cell value and c^* represents the perturbed cell value. Now calculate the matrix $\mathbf{Q}$ using Bayes formula by $Q_{kj} = p(c = j \mid c^* = k) = \dfrac{p_{jk}\, p(c = j)}{\sum_l p_{lk}\, p(c = l)}$. We estimate the entries Q_{kj} of this matrix by $\dfrac{p_{jk} v_j}{\sum_l p_{lk} v_l}$, where v_j is the relative frequency of cell value *j*. For $\mathbf{R} = \mathbf{PQ}$ we obtain an invariant matrix where $\mathbf{vR} = \mathbf{vPQ} = \mathbf{v}$ since $r_{ij} = \sum_k \dfrac{v_j p_{ik} p_{jk}}{\sum_l p_{lk} v_l}$ and $\sum_i v_i r_{ij} = \sum_k v_j p_{ik} = v_j$. The vector of the original frequencies $\mathbf{v}$ is an eigenvector of $\mathbf{R}$. In practice, $\mathbf{Q}$ can be calculated by transposing matrix $\mathbf{P}$, multiplying each column j by v_j and then normalizing its rows so that the sum of each row equals one. Since the property of invariance distorts the desired probabilities on the diagonal (the probabilities of not changing a cell value), we propose defining a parameter α and calculating $\mathbf{R}^* = \alpha\mathbf{R} + (1-\alpha)\mathbf{I}$ where $\mathbf{I}$ is the identity matrix of the appropriate size. $\mathbf{R}^*$ is also invariant and the amount of perturbation is controlled by the parameter α.

The perturbation process is typically carried out using a "with replacement" strategy where each cell is independently perturbed based on a random multinomial draw and changing or not changing the cell value according to the outcome of the draw. In order to obtain the exact marginal distribution (and an exact total), we can carry out a "without" replacement strategy for selecting cell values to change. In the first step, the expected number of cell values that need to be changed are calculated based on the probabilities in the transition matrix. In the second step, the expected number of cells are selected randomly and the change made. For example, assume 20 cells with a value of one and the following probabilities: 0.80 of them will remain a one, 0.10 of them will change to a zero and 0.10 will change to a two. The expected number of changes are two cells to a zero, two cells to a two and the remainder with no change. Select randomly two cells from among the 20 cells and change the value to zero. Select randomly another two cells from the remaining 18 cells and change the value to two. This selection method not only ensures an exact total but also reduces the additional variance that is induced by the perturbation. This method was used to perturb the Sample of Anonymized Records (SARs) of the 2001 UK Census (Gross, Guiblin and Merrett, 2004).

To define the matrix $\mathbf{P}$, the Statistical Agency needs to decide on the amount of perturbation required for each cell value, at what probabilities the cell values will change to a different value, and the spread of the perturbation. The invariant probability matrix ensures a bias of zero in the resulting table and the values and spread of the probabilities on the off-diagonals determine the perturbation variance. One constraint on the matrix $\mathbf{P}$ might be maximizing the entropy of the perturbation by defining an equal probability of moving to all neighboring cells for each cell value. This constraint will maximize the perturbation variance for a given value on the diagonal.

An example of the impact of the spread of the perturbation on the variance is as follows: assume a 9×9 probability transition matrix $\mathbf{P}$ for cell values 0,1,…8 and a probability of 0.7 on the diagonal (i.e., no change occurs on the cell value with a probability of 0.7), and all the off-diagonals have equal probabilities (maximum entropy) of 0.0375 (=0.3/8). Assume also an equal number of cells having values 0,1,…,8. The invariant matrix $\mathbf{R}^*$ with $\alpha = 0.5$ will have 0.751 on the diagonal and 0.031 on the off-diagonals. The expectation of the perturbation process based on the invariant matrix is 0 and the average perturbation variance across cells 3.72. Assume on the other hand the probability transition matrix $\mathbf{P}$ presented on the left side of (1) and the invariant matrix $\mathbf{R}^*$ on the right side of (1). For an equal number of cells for each value, the expectation of the perturbation process based on the invariant matrix is 0 and the average perturbation variance across cells 0.37.

$$\begin{pmatrix} 0.70 & 0.15 & 0.15 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\ 0.15 & 0.70 & 0.15 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.15 & 0.70 & 0.15 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.15 & 0.70 & 0.15 & 0.00 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.00 & 0.15 & 0.70 & 0.15 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.00 & 0.00 & 0.15 & 0.70 & 0.15 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.15 & 0.70 & 0.15 & 0.00 \\ 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.15 & 0.70 & 0.15 \\ 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.15 & 0.15 & 0.70 \end{pmatrix} \quad \begin{pmatrix} 0.81 & 0.12 & 0.06 & 0.01 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\ 0.12 & 0.77 & 0.10 & 0.01 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\ 0.05 & 0.10 & 0.74 & 0.10 & 0.01 & 0.00 & 0.00 & 0.00 & 0.00 \\ 0.01 & 0.01 & 0.11 & 0.76 & 0.11 & 0.01 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.01 & 0.11 & 0.76 & 0.11 & 0.01 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.00 & 0.01 & 0.11 & 0.76 & 0.10 & 0.01 & 0.01 \\ 0.00 & 0.00 & 0.00 & 0.00 & 0.01 & 0.10 & 0.74 & 0.10 & 0.05 \\ 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.01 & 0.10 & 0.77 & 0.12 \\ 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.01 & 0.06 & 0.12 & 0.81 \end{pmatrix} \quad (1)$$

As mentioned, using a "without" replacement selection strategy for changing cell values will ensure exact totals and eliminate the perturbation variance. However, we need to develop a different selection process that will preserve the consistency of cell values for same cells disseminated across different tables. This method is described in Section 2.2.

2.2 Microdata Keys

Motivated by the method developed by the Australian Bureau of Statistics (ABS) (Fraser and Wooton, 2006), a random number (a key) is defined for each record in the microdata. When records are grouped together to form a cell in a table, their keys are also aggregated and their sum defines the seed for the perturbation. The aim is that all records aggregated into same cells will have a consistent seed and therefore a consistent perturbation across all tables. The key can be defined for example by generating a random uniform number u_i from 0 to some large number for each record i in the microdata. The aggregation of the records to form a cell c of a table produces an aggregation of the random uniform numbers: $U^c = \sum_{i \in c} u_i$. Taking the modulo of 100 of U^c, i.e. $Q^c = \text{mod}(U^c, 100)$ and a large enough table, a uniform distribution from 0 to 99 is obtained which can be used to carry out a consistent perturbation across tables.

The proposed method of a permanent seed generated from the microdata imposes a "with" replacement selection strategy where cells are independently perturbed and therefore this selection method will only preserve the expectations of the totals as described in Section 2.1. In order to carry out a "without" replacement selection strategy to preserve exact totals with the intent to also preserve consistency to a large extent, we might consider drawing the cell values to change systematically using the following algorithm:

1. For each cell value c_i, determine the expectations of the outcome of the perturbations by multiplying the number of cells n_{c_i} by the transition probabilities $r^*_{c_i c_j}$: $E_{c_i c_i} = n_{c_i} \times r^*_{c_i c_i}$ and $E_{c_i c_j} = n_{c_i} \times r^*_{c_i c_j}$
2. Sort the original cells having a value of c_i by their aggregated key Q^{c_i}
3. Select the first $E_{c_i c_i}$ values and maintain their value, select the next $E_{c_i c_j}$ and change the value from c_i to c_j, repeat for all c_j
4. Place the perturbed cells back in their original order

Since the sorting of the cells is based on a constant Q^{c_i}, which is a random uniform number between 0 to 99, we can expect that some of the consistency will be preserved using a "without" replacement selection strategy.

For cell values of zero, no microdata keys are defined. If the perturbation is zero-restricted (cell values of zero are not perturbed), then consistency will not be a problem for the non-zero cells. In general, Statistical Agencies prefer not to perturb zero cells in order to ensure that no positive values appear for structural zeros in a table.

Another feature in the perturbation process is that there may be key statistics in a table where the Statistical Agency would not want any perturbation applied. The records that are involved in producing these statistics can have their Microdata Keys determined a priori so that U^c is set to a specific chosen value that would indicate that no perturbation should be applied.

2.3 Preserving Additivity

In order to preserve consistency all non-zero cells should undergo the perturbation process including margins. Since the expected value of the totals are to be preserved using the invariant probability transition matrix and microdata keys, only minor discrepancies will occur between aggregated internal cells to the perturbed margins. To obtain exact additivity to the perturbed margins we need to carry out linear programming. Note that preserving the additivity will likely impact on the consistency of cells across tables. However with this proposed method, since only minor adjustments will be needed to correct the additivity, the impact on consistency will be minimal.

One simple way to preserve additivity is through an IPF algorithm as explained below for a two-dimensional table defined by cell counts m_{rc} with rows $r = 1,..,R$ and columns $c = 1,..,C$. Assume marginal row values by M_r and marginal columns values by M_c.

1. Calculate row totals for each $r = 1,..,R$: $t_r = \sum_{c=1}^{C} m_{rc}$
2. Calculate the ratio: $\alpha_r = M_r / t_r$ for each $r = 1,..,R$
3. Multiply internal cells in each $r = 1,..,R$ by α_r
4. Calculate column totals for each $c = 1,..,C$: $t_c = \sum_{r=1}^{R} \alpha_r m_{rc}$
5. Calculate the ratio: $\beta_c = M_c / t_c$ for each $c = 1,..,C$
6. Multiply internal cells in each $c = 1,..,C$ by β_c
7. Repeat steps 1 though 6 until the algorithm converges and all internal cells add up to their marginal totals

At this stage, internal cells add up exactly to perturbed margins but they are non-integer values. Because of the perturbation method and the use of an invariant probability transition matrix, the non-integer values are similar to the perturbed integer values. The non-integer values can be rounded to the nearest integer which will offset the additivity slightly due to rounding errors but is easy to implement. To preserve exact additivity and obtain integer values after IPF, the controlled rounding procedure in Tau–Argus to base 1 (Hundepool, 2002, Salazar-Gonzalez, et al., 2005) can be implemented or other software for controlled rounding or "reshuffling" algorithms (Boudreau, Filep and Liu, 2004).

3 Application

We examine a typical table extracted from one estimation area (EA) of the unperturbed 2001 UK Census data. The table is disseminated by Output Areas (OA) which are the smallest Census tracts that are published for the UK Census. The number of OAs in the EA is 1,487 and includes on average about 125 households. For each OA, the table is defined as follows (the number of categories is given in parenthesis): Economic Activity (9) × Sex (2) × Long-Term Illness (2), i.e. a total of 36 categories. The table includes 317,064 individuals between the ages of 16 and 74 in 53,532 internal cells. The average cells size is 5.92 although the table is skewed with very large and very small columns. Table 1 presents the distribution of cell values in the table.

Table 1. Distribution of cells according to cell value

Cell values	Number of cells	Percent
0	17,915	33.5%
1	8,813	16.5%
2	5,913	11.0%
3	4,253	7.9%
4	2,992	5.6%
5	2,210	4.1%
6	1,610	3.0%
7	1,131	2.1%
8	906	1.7%
9	675	1.3%
10	570	1.1%
11+	6,544	12.2%
Total	53,532	100.0%

We implement the proposed method described in Section 2 using the initial probability transition matrix presented in Table 2. All values under 10 in the matrix represent the cell values themselves. For example, a cell value of 6 can move to cell values of 4,5,7 or 8 with equal probability. For cell values over 11, we calculate their residual from base 3, i.e. the value of 11 has a residual of (2), the value of 12 has a residual of (0) and the value of 13 has a residual of (1). The changes to these cell values are based on perturbations, i.e. adding a -2,-1 or +1 to the original cell value. For example, from Table 2, a large cell above 11 with a residual of 0 or 2 can be perturbed

Table 2. Initial probability transition matrix for the Census table

Cell Value	Perturbed Cell Values												
	0	1	2	3	4	5	6	7	8	9	10	11+*	
												(2)	(0)	(1)
0	1.0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	.10	.70	.10	.10	0	0	0	0	0	0	0	0	0	0
2	0	.15	.70	.15	0	0	0	0	0	0	0	0	0	0
3	0	.05	.05	.80	.05	.05	0	0	0	0	0	0	0	0
4	0	0	.05	.05	.80	.05	.05	0	0	0	0	0	0	0
5	0	0	0	.05	.05	.80	.05	.05	0	0	0	0	0	0
6	0	0	0	0	.05	.05	.80	.05	.05	0	0	0	0	0
7	0	0	0	0	0	.05	.05	.80	.05	.05	0	0	0	0
8	0	0	0	0	0	0	.05	.05	.80	.05	.05	0	0	0
9	0	0	0	0	0	0	0	.05	.05	.80	.05	.05	0	0
10	0	0	0	0	0	0	0	0	.05	.05	.80	.05	.05	0
11+ (2)	0	0	0	0	0	0	0	0	0	0	.10	.80	.10	0
11+ (0)	0	0	0	0	0	0	0	0	0	0	0	.10	.80	.10
11+ (1)	0	0	0	0	0	0	0	0	0	0	0	.10	.10	.80

* Cell values over 11 relate to perturbations of -2,-1 or +1.

either with -1 or +1 with equal probability while a residual of 1 can be perturbed either with -2 or -1 with equal probability.

Based on the distribution of cell values in Table 1 and the initial probability transition matrix in Table 2, we calculate an invariant probability transition matrix shown in Table 3 where the same explanation holds for larger cell values over 11. The mean of the perturbation process is 0 and the average perturbation variance across cells 0.41. Note that there is a slightly larger spread in the invariant probability transition matrix with very small probabilities of perturbation at the edges.

Table 3. Final invariant probability transition matrix for the Census table

Cell Value	Perturbed Cell Values													
	0	1	2	3	4	5	6	7	8	9	10	11+*		
												(2)	(0)	(1)
0	.977	.023	0	0	0	0	0	0	0	0	0	0	0	0
1	.048	.816	.089	.044	.003	.001	0	0	0	0	0	0	0	0
2	0	.133	.791	.063	.012	.002	0	0	0	0	0	0	0	0
3	0	.090	.088	757	.035	.029	.002	.001	0	0	0	0	0	0
4	0	.008	.023	.049	.848	.037	.032	.001	.001	0	0	0	0	0
5	0	.004	.004	.055	.051	.817	.036	.030	.002	.001	0	0	0	0
6	0	0	0	.004	.059	.050	.818	.035	.032	.002	.001	0	0	0
7	0	0	0	.002	.004	.059	.050	.813	.037	.033	.002	0	0	0
8	0	0	0	0	.002	.004	.056	.047	.819	.037	.028	.008	0	0
9	0	0	0	0	0	.002	.004	.055	.049	.825	.033	.028	.002	.002
10	0	0	0	0	0	0	.002	.004	.045	.039	.738	.146	.022	.005
11+(2)	0	0	0	0	0	0	0	0	.003	.008	.035	.841	.074	.040
11+(0)	0	0	0	0	0	0	0	0	0	.001	.006	.084	.821	.088
11+(1)	0	0	0	0	0	0	0	0	0	.001	.001	.046	.088	.865

* Cell values over 11 relate to perturbations of -2,-1 or +1.

For this application, we perturb only internal cells in order to examine the impact on the original totals and use a "with" replacement selection strategy for selecting cell values to change based on aggregated microdata keys Q^c and the invariant probability transition matrix in Table 3. This ensures that the expected values of the marginal totals are preserved, i.e. marginal totals after perturbation should be similar to the original marginal totals. We implement the IPF to obtain exact marginal totals, round each internal cell value to its nearest integer and again aggregate to obtain new marginal totals.

The results for the marginal totals before perturbation, after perturbation and the IPF with rounding to the nearest integer are presented in Table 4 for the column variables: Sex, Long- Term Illness, Economic Activity and the Overall Total. These totals are obtained by aggregating across the 1,487 geographical areas and the other categories of the column variables. Using the invariant probability transition matrix, the totals after perturbation are similar to the original totals. After applying IPF and rounding to the nearest integer, there is an improvement in the marginal totals compared to the original totals.

Table 4. Marginal totals for Overall Total, Sex, Long Term Illness and Economic Activity before perturbation, after perturbation, and after IPF and rounding

Variable	Before Perturbation	After Perturbation		After IPF and nearest integer rounding	
	Total	Total	Percent Difference	Total	Percent Difference
Overall Total	317,064	316,699	0.12	316,959	0.03
		Sex			
Males	159,777	159,655	0.08	159,748	0.02
Females	157,287	157,044	0.15	157,211	0.05
		Long-Term Illness			
Long-Term Illness	46,060	46,094	-0.07	46,106	-0.10
No Long-Term Illness	271,004	270,605	0.15	270,853	0.06
		Economic Activity			
Part Time Employed	45,853	45,783	0.15	45,831	0.05
Full Time Employed	158,342	158,171	0.11	158,258	0.05
Unemployed	7,162	7,299	-1.91	7,296	-1.87
Full Time Student	10,828	10,822	0.06	10,825	0.03
Retired	37,910	37,626	0.75	37,669	0.64
Student	18,041	17,977	0.35	18,034	0.04
Looking After Home	19,244	19,253	-0.05	19,275	-0.16
Permanently Disabled	12,205	12,168	0.30	12,171	0.28
Other	7,479	7,600	-1.62	7,600	-1.62

As a final step, we implement the controlled rounding option to Base 1 of Tau-Argus to obtain integer value internal cells and exact margins that equal the totals before the perturbation as presented in the second column of Table 4.

We compare the proposed method to several standard SDL methods for Census frequency tables with respect to disclosure risk and data utility in Table 5. The comparison is made to other methods that have relatively the same amount of perturbation compared to the probability transition matrix that was used in Table 3. Descriptions of the SDL methods, disclosure risk and data utility measures used in this analysis as presented in Table 5 can be found in Shlomo, 2007.

With respect to disclosure risk from small cells, post-tabular random and controlled rounding procedures eliminate small cells in tables and hence perform better than any of the other methods. Compared to record swapping techniques, the small cells in the table using the proposed invariant post-tabular method were slightly less protected. It should be noted however that in this particular application, we allowed the dissemination of small cells in the Census table. The Statistical Agency can define the probability transition matrix in Table 2 to have greater perturbation on small cells depending on its disclosure risk thresholds. Indeed, post-tabular methods provide more control on the perturbation of small cells compared to pre-tabular methods. Therefore, disclosure risk based on small cells can substantially be reduced using the proposed invariant post-tabular method although this may impact negatively on data utility.

Table 5. Disclosure risk and data utility measures for SDL methods on Census table (values for the original table in parenthesis)

Method	Disclosure Risk	Data Utility			
	Prop. Small Cells not Perturbed	Average Hellingerís Distance	Cramer's V (0.121)	Variance Average Cell Size (188.3)	Between Variance* (0.000233)
Invariant Post-tabular Method (after IPF and Controlled Rounding)	0.73	0.71	1.79	0.31	2.02
Random Round (RR) Base 3	0	2.03	11.58	0.52	11.42
Semi-controlled RR Base 3	0	2.04	11.88	0.54	13.14
Controlled Round Base 3	0	1.95	9.97	0.39	12.91
Cell Suppression	0	0.42	0.22	-0.04	-0.64
Random Swap** 10%	0.65	1.39	-3.65	-1.31	-4.82
Targeted Swap** 10%	0.49	1.58	-1.93	-0.59	-3.49

* The variable is the proportion of full-time male students with no long-term illness.

** 10% of records are selected from the microdata and are paired with another 10% of records and geography variables swapped.

Since the invariant post-tabular method can perturb non-zero cell values to zero cells, we gain ambiguity in the original zero cells of the table and thus reduce the risk of attribute disclosure. Disclosure risk is further reduced by minimizing the possibilities of linking and differencing tables for the following reasons:

1. Perturbation distributions have longer tails with varying perturbation variances across cells and are harder to decipher,
2. Most of the consistency across cells in same tables is preserved making it harder to identify and link cells that have undergone the perturbation.

It should be noted that cell suppression also eliminates small cells, but all other values are exactly preserved and therefore tables can be differenced to produce small cells and are prone to very high disclosure risk. Also, cell suppression is not generally used in a Census context because of the difficulty in suppressing same cells across different tables.

With respect to data utility, besides the method of cell suppression, the utility in the data is better preserved under the invariant post-tabular method compared to the other SDL methods. Since more zero cells are introduced into the perturbed table, the impact on statistical analysis as a result of "sharpened" distributions is similar to the rounding procedures but with less damage. This effect is the opposite to the attenuation seen in record swapping reflected in the negative values. For example, the

Cramer's V statistics has only a 1.8% difference from the value obtained from the original table compared to 10.0% to 11.9% for the rounding procedures to base 3 and the between variance for the proportion of full-time male students with no long-term illness across geographical areas has a 2.0% difference compared to 11.4% to 13.1%. Some of the difference between these methods can be explained by the fact that we allow small cells in the table for the invariant post-tabular method and therefore higher disclosure risk on small cells.

4 Discussion

Oganian and Karr, 2006 and Shlomo and De Waal, 2008 discuss the advantages of combining different SDL methods in order to improve data utility with respect to preserving variances, ensuring unbiasedness and preserving consistency and statistical inference. The analysis of the invariant post-tabular method combining pre-tabular and post-tabular SDL methods demonstrates that disclosure risk can be managed while raising data utility. The proposed method depends on linear program techniques to ensure exact additivity constraints which complicates implementation in a production line. This constraint however reduces consistency of cells across tables. By perturbing with an invariant probability transition matrix and weakening the additivity constraint to some extent by ensuring unbiased expectations of the totals (i.e., the totals are similar to the original totals as shown in Table 4), the method is very easy to implement and consistency exactly preserved across tables. Such a method could be used in a flexible table generating web-based package.

References

1. Boudreau, J.R., Filep, K., Liu, L.: Iterative Rounding for Large Frequency Tables. In: Proceedings of the Government Statistics Section, American Statistical Association (2004)
2. Dalenius, T., Reiss, S.P.: A Technique for Disclosure Control. Journal of Statistical Planning and Inference 7, 73–85 (1982)
3. Dobra, A., Fienberg, S.E.: Bounding Entries in Multi-way Contingency Tables Given a Set of Marginal Totals. In: Haitovsky, Y., Lerche, H.R., Ritov, Y. (eds.) Foundations of Statistical Inference: Proceedings of the Shoresh Conference 2000, pp. 3–16. Phisica-Verlag (2003)
4. Fienberg, S.E., McIntyre, J.: Data Swapping: Variations on a Theme by Dalenius and Reiss. Journal of Official Statistics 9, 383–406 (2005)
5. Fienberg, S.E., Slavkovic, A.B.: Preserving the Confidentiality of Categorical Statistical Data Bases when Releasing Information for Association Rules. Data Mining and Knowledge Discovery 11, 155–180 (2005)
6. Fraser, B., Wooton, J.: A Proposed Method for Confidentializing Tabular Output to Protect Against Differencing. Internal Report, Data Access and Confidentiality Methodology Unit, Australian Bureau of Statistics (2006)
7. Gouweleeuw, J., Kooiman, P., Willenborg, L.C.R.J., De Wolf, P.P.: Post Randomisation for Statistical Disclosure Control: Theory and Implementation. Journal of Official Statistics 14, 463–478 (1998)

8. Gross, B., Guiblin, P., Merrett, K.: Implementing the Post-Randomisation Method to the Individual Sample of Anonymised Records from the 2001 Census (2004), `http://www.ccsr.ac.uk/sars/events/2004-09-30/gross.pdf`
9. Hundepool, A.: The CASC Project. In: Domingo-Ferrer, J. (ed.) Inference Control in Statistical Databases. LNCS, vol. 2316. Springer, Heidelberg (2002)
10. Oganian, A., Karr, A.: Combinations of SDC Methods for Micro-data Protection. In: Domingo-Ferrer, J., Franconi, L. (eds.) PSD 2006. LNCS, vol. 4302, pp. 102–113. Springer, Heidelberg (2006)
11. Salazar-Gonzalez, J.J., Bycroft, C., Staggemeier, A.T.: Controlled Rounding Implementation. In: Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality, Geneva (2005)
12. Shlomo, N.: Assessing the Impact of SDC Methods on Census Frequency Tables. In; Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality, Manchester (2007), `http://www.unece.org/stats/documents/2007/12/confidentiality/wp.18.e.pdf`
13. Shlomo, N., De Waal, T.: Protection of Micro-data Subject to Edit Constraints Against Statistical Disclosure. Journal of Official Statistics 24(2), 1–26 (2008)
14. Willenborg, L.C.R.J., De Waal, T.: Elements of Statistical Disclosure Control in Practice. Lecture Notes in Statistics, vol. 155. Springer, Heidelberg (2001)

A Practical Approach to Balancing Data Confidentiality and Research Needs: The NHIS Linked Mortality Files

Kimberly Lochner, Stephanie Bartee, Gloria Wheatcroft, and Christine Cox

National Center for Health Statistics, Hyattsville, MD 20782
KLochner@cdc.gov

Abstract. Demand is increasing for statistical agencies to integrate information from different sources and to make such data files publicly available. The National Center for Health Statistics (NCHS) conducts record linkage activities for its surveys, with a major focus being the linkage to the National Death Index. In 2004, NCHS updated the mortality follow-up for the National Health Interview Survey (NHIS) which, because of confidentiality protections, was made available only through the NCHS Research Data Center. In 2007, NCHS released a public-use version of the NHIS Linked Mortality Files. The release of a public-use file was based upon an approach to maximize the amount of mortality information included and minimize the amount of perturbed data, while maintaining the confidentiality protections of survey participants. Comparative analyses between the public-use and restricted-use linked mortality files show that the two data files yield very similar results for both all-cause and cause-specific mortality.

Keywords: confidentiality; health surveys; record linkage; mortality.

1 Introduction

There is increasing demand for federally sponsored health surveys to integrate information from additional data sources in order to enhance the availability and quality of information on exposures and outcomes and to make such data files publicly available. The National Center for Health Statistics (NCHS) enhances several of its population-based surveys through record linkage to administrative files. However, the linking of records from different data sources can increase the chance that an individual's identifiable information may be at risk of being disclosed and NCHS must balance the desire to provide publicly available, high quality, and timely data, with maintaining appropriate safeguards for the confidentiality of individual responses.

A major focus of NCHS record linkage activities is the mortality follow-up of its surveys through record linkage to the National Death Index (NDI), which maintains a national file of death record information [1, 2]. In 2004, NCHS completed a mortality follow-up study for the 1986 to 2000 National Health Interview Survey (NHIS) years through a probabilistic record linkage with the NDI, with data available only through the NCHS Research Data Center (RDC) to ensure that identifiable information of

J. Domingo-Ferrer and Y. Saygın (Eds.): PSD 2008, LNCS 5262, pp. 90–99, 2008.

NHIS participants was not released. In 2007, in order to maximize access to these previously highly utilized data sources for public health research, NCHS developed a public-use version of the files.

This paper describes the several steps NCHS undertook to release a public-use version of the NHIS Linked Mortality Files in an effort to balance the data needs of the research community, while protecting the confidentiality of survey participants. The process of releasing public-use linked mortality microdata files included assessing re-identification risk, applying statistical disclosure methods to protect the confidentiality of data, and evaluating the analytic properties of the public-use files (which include some perturbed data) as compared to the original linked data files. This paper does not address the various disclosure avoidance techniques used in the release of other NCHS public-use data files.

2 Re-identification Risk Simulation and Data Perturbation Plan

The NHIS is a cross-sectional household interview survey of the civilian noninstitutionalized population of the United States. The NHIS collects data on a broad range of health topics and socio-demographic information. Descriptions of the NHIS design have been published elsewhere [3, 4]. The NHIS years 1986 to 2000 have mortality follow-up through December 31, 2002 through probabilistic record linkage to the NDI. A complete description of the methodology used to link NHIS records to the NDI can be found at www.cdc.gov/nchs/data/datalinkage/matching_methodology _nhis_final.pdf (accessed December 11, 2007) [5]. Throughout this paper, the original NHIS Linked Mortality Files available in the NCHS RDC are referred to as "restricted-use" files to distinguish them from the modified public-use files.

NCHS has strong confidentiality protections for its data products and a Disclosure Review Board (DRB) that reviews the disclosure potential of proposed public-use data releases. NCHS proposed a data release plan for public-use NHIS Linked Mortality Files that would reduce the re-identification risk to survey participants and maximize the amount of mortality information included, while limiting the amount of perturbed data introduced. Standard approaches to protecting confidentiality include masking techniques, e.g., creating categorical variables from continuous data, swapping values, and imputation [6, 7, 8].

The first step in the plan was to determine the mortality information (e.g. timing of death, cause of death) for inclusion in the public-use data files. We proposed to replace exact follow-up time with approximate follow-up time and limit information on exact cause of death to a grouped recode. Information on age and interview date is readily available from the public-use NHIS survey data files. The restricted-use files provide more detail on the NHIS interview date and NHIS participant age than what is available on the public-use NHIS survey data files. For example, NCHS has made available on the restricted-use NHIS Linked Mortality files the exact date of NHIS interview (month, day, year) as well as detailed information on age at interview in years (not top-coded), date of birth (month, day, year), and age at death. Such detail on interview date, age, and timing of death facilitate the creation of more detailed and specific follow-up times for mortality analyses. Table 1 lists key variables included on the restricted-use files as well as the reduced number of variables available on the public-use NHIS Linked Mortality files.

Table 1. Selected variables available for the NHIS* linked mortality files

	Restricted-use	Public-use
Final mortality status	Yes	Yes
Death date	Yes (month, day, year)	Yes (quarter, year)
Underlying cause-of-death	Yes	Yes (grouped recode)
Contributing cause-of-death	Yes	Yes†
Age at interview	Yes	Yes (top coded at 85+)‡
Age at death	Yes	No
Age last presumed alive	Yes	No
Date of birth	Yes (month, day, year)	Yes (month, year) ‡
Interview date	Yes (month, day, year)	Yes (quarter, year) ‡

* National Health Interview Survey.

† Flags only for diabetes, hypertension, or hip fracture.

‡ Available on the public-use NHIS survey data files.

We assessed re-identification risk by identifying existing publicly available data sources that have variables in common with the proposed mortality information on the public-use NHIS Linked Mortality Files. Using key NHIS public-use socio-demographic variables and mortality information for NHIS decedents, we identified unique records and then matched them to the external data sources. This exercise demonstrated that in most cases "unique" records, although matching based upon the unique combination of factors, were, in fact, not being correctly matched to the same individual. This is most likely due to differences in reporting between the sources of data, i.e., self-reported survey data and administrative death records. We considered all NHIS decedent "unique" records, which were correctly matched to these public data sources, to be at risk for being re-identified.

After identifying the cases at risk for re-identification, we constructed a plan to modify the mortality data for those NHIS decedents at risk and allow for the release of a NHIS linked mortality public-use file. All cases considered "re-identifiable" were subject to data perturbation and were randomly assigned to have either date of death or underlying cause-of-death perturbed. Information regarding vital status was not perturbed. Cases requiring date of death perturbation had either the quarter or year randomly perturbed, and in some cases both fields were perturbed. For those cases requiring underlying cause-of-death perturbation, we implemented a hot-deck method and imputed the 113 grouped underlying-cause-of-death recode by replacing the original value with a value from a decedent with similar characteristics [9]. To further reduce re-identification risk, an additional random sample of decedents was subjected to perturbation. A final attempt to match back the perturbed data to the external data files yielded no unique correct matches. The perturbed cases are not identified on the public-use files.

3 Comparative Analysis

We replicated analyses conducted on the restricted-use files to those conducted on the modified public-use files to demonstrate the analytic comparability between the two versions of the linked mortality files. We restricted all analyses to those eligible for mortality follow-up, who were at least 25 years of age at the time of the NHIS interview, were non-Hispanic white, non-Hispanic black, or Hispanic, and with no missing values for education level, marital status, and cause of death. We compared mean follow-up times and distributions for select causes of death and used Cox proportional hazards models to compare the relative hazards for mortality risk among a standard set of socio-demographic characteristics, which were observed at the time of NHIS interview. All analyses take into account the complex survey design of the NHIS [10, 11].

We examined mortality in the public-use and restricted-use NHIS Linked Mortality Files using time from NHIS interview until death; respondents who were not identified as dying by the end of the follow-up period were assumed to be alive. For the public-use files, duration of follow-up was constructed using NHIS interview year and year of death. For respondents assumed alive, follow-up time was calculated by assigning ½-year of follow-up during their NHIS interview year and a full year of follow-up for each year thereafter until the end of 2002. For the restricted-use files, duration of follow-up was calculated using complete information on the month, day, and year of the NHIS interview and the month, day, and year of death or, for respondents assumed alive, until the end of the follow-up period, December 31, 2002.

In addition to all-cause mortality, we examined 14 causes of death that are among the ten leading causes of death in the United States and/or contribute to the most years of potential life lost [12]: heart disease, ischemic heart disease, cancer (all sites), lung cancer, colorectal cancer, breast cancer (estimated for women only), prostate cancer, cerebrovascular diseases, diabetes, pneumonia and influenza, chronic liver diseases and cirrhosis, unintentional injuries, suicide, and homicide. The cause-specific death categories are based upon the underlying causes of death from the ICD-10 113-group recode. The cause-specific analyses presented in this paper do not control for the transition in coding rules between ICD-9 and ICD-10 because that transition does not affect the comparisons of interest in this paper [13]. Due to an insufficient number of deaths in certain population subgroups, we restricted the cause-specific mortality analyses to non-Hispanic whites and non-Hispanic blacks.

3.1 Results of the Comparative Analysis

The final sample for the comparative analyses included 897,232 records and 114,264 deaths. The weighted distribution for covariates included in the models is the same for both sets of analyses using the public-use and restricted-use linked mortality files. The average age of this sample is 47.9 years and fewer than two percent of respondents are aged 85 or above. Females outnumber males (52.6 to 47.4 percent, respectively), and non-Hispanic whites make up just over 80 percent of the sample while non-Hispanic blacks (10.9 percent) and Hispanics (8.2 percent) account for considerably smaller proportions. A vast majority of the sample is married at the time of NHIS interview (69.0 percent) and the modal educational category is a high school degree or GED (36.0 percent), with 20.4 percent having less than a high school education, 21.4 percent some college, and 22.1 percent at least a college degree. Over 35 percent of the sample resides in the South, while nearly 25 percent resides in the Midwest and 19 percent in the West.

Table 2 shows the comparative descriptive statistics for mortality outcome variables among the public-use and restricted-use files, respectively. The total number and percentage of persons who were identified as dying in each of the two files (n = 114,264; Percent = 11.8) is identical. As mentioned above, this illustrates that the vital status of individuals was not changed for anyone as a result of the perturbation process for the public-use file. However, there are some modest differences in the cause of death distributions when comparing the public-use and restricted-use files. For example, the number of deaths attributed to some of the more common causes of death, such as heart disease (n = 37,272) and lung cancer (n = 8,838) in the public-use file is greater than the number of deaths attributed to those causes in the restricted-use file (n = 36,689 and n = 8,395, respectively). Similarly, there are modest differences for some of the less common causes, such as unintentional and intentional injuries.

Table 2. Mortality characteristics of 897,232 adults aged 25 years and older in the 1986-2000 NHIS* linked mortality files

	Public-use		Restricted-use	
	Unweighted n	Weighted (%)	Unweighted n	Weighted (%)
Follow-up, mean yrs	9.1	8.7	9.1	8.6
Cause-specific deaths†				
Diseases of the heart	37,272	32.5	36,689	32.0
Ischemic heart disease	11,434	10.0	11,290	9.8
Cancer, all sites	30,220	26.6	30,197	26.5
Lung cancer	8,838	7.8	8,395	7.4
Colorectal cancer	3,044	2.6	3,094	2.7
Breast cancer‡	2,421	4.3	2,372	4.2
Prostate cancer	1,762	3.0	1,786	3.0
Cerebrovascular diseases	7,802	6.8	7,855	6.8
Diabetes	3,361	2.9	3,384	2.9
Pneumonia/Influenza	3,306	2.9	3,342	2.9
Chronic liver disease/ Cirrhosis	1,238	1.1	1,268	1.1
Unintentional injuries	3,242	2.9	3,294	2.9
Suicide	1,097	1.0	1,117	1.1
Homicide	410	0.3	425	0.4

* National Health Interview Survey. Mortality follow-up through December 31, 2002.

† Underlying cause-of-death codes are based upon the ICD-10 113-group recode. As only select leading causes of death presented, total number of cause specific deaths does not sum to 114,264. Weighted percentages for cause-specific deaths are based upon the sample of decedents.

‡ Women only.

Table 3 presents results from Cox proportional hazards models of all-cause mortality: one estimated from the public-use file and one estimated from the restricted-use file. The results of both models are consistent with expectations, given the results from similar models that used an earlier version of this data set [14]. Moreover, hazard ratios (HR) and 95 percent confidence intervals (CI) are essentially identical when comparing the results from the public-use and restricted-use files. For example, in the restricted-use file, mortality from all-causes was higher for men compared to women (HR = 1.69, 95 percent CI: 1.67, 1.71), and this result was replicated in the public-use data. For the other covariates, similar results were obtained using the two data files. We also estimated models stratified by sex and race/ethnicity (data not shown). The sex-specific models yield results that are consistent with previous research and again the public-use and restricted-use files obtain nearly identical hazard ratios and 95 percent confidence intervals. Similarly for non-Hispanic whites, non-Hispanic blacks and Hispanics, covariates exhibit relationships with all-cause mortality that are consistent with what one would expect from the literature [14]. Given differences in the way that the duration of follow-up variable was calculated for the restricted-use and public-use versions of the NHIS Linked Mortality Files, the slight differences in model results for all-cause mortality can be accounted for by differences in the duration of follow-up variables.

Table 3. All-cause mortality by socio-demographic characteristics for adults aged 25 years and older in the 1986-2000 NHIS* linked mortality files, n = 897,232, deaths = 114,264

	Public-use		Restricted-use	
	HR†	95% CI†	HR†	95% CI†
Age in years	1.09	1.09, 1.09	1.09	1.09, 1.09
Sex				
Women	1.00		1.00	
Men	1.69	1.67, 1.71	1.69	1.67, 1.71
Race/Ethnicity				
non-Hispanic white	1.00		1.00	
non-Hispanic black	1.15	1.13, 1.18	1.15	1.13, 1.18
Hispanic	0.89	0.86, 0.92	0.89	0.87, 0.92
Marital status				
Married	1.00		1.00	
Widowed	1.23	1.21, 1.25	1.23	1.21, 1.25
Divorced/separated	1.40	1.36, 1.43	1.40	1.36, 1.43
Never married	1.48	1.44, 1.53	1.48	1.44, 1.53
Education level				
Less than high school	1.68	1.64, 1.72	1.68	1.64, 1.72
High school/GED	1.41	1.37, 1.44	1.41	1.37, 1.44
Some college	1.28	1.25, 1.31	1.28	1.25, 1.31
College degree or more	1.00		1.00	
Region				
Northeast	0.97	0.95, 1.00	0.98	0.95, 1.00
Midwest	0.99	0.96, 1.01	0.99	0.96, 1.01
South	1.05	1.03, 1.08	1.05	1.03, 1.08
West	1.00		1.00	

* National Health Interview Survey. Mortality follow-up through December 31, 2002.

† HR, hazard ratio; CI, confidence interval. Estimated from a Cox proportional hazards model.

Each cause-specific mortality table compares the model results from the public-use version and the restricted-use version of the NHIS Linked Mortality Files. Due to space constraints, we present results for two of the 14 cause specific mortality analyses. Table 4 presents cause-specific results for cancer (all sites) mortality. Mortality risk increases just over seven percent for each additional year of age in both the public-use data model and the restricted-use data model. Men experience higher cancer mortality risk than women over the course of the follow-up period (public-use data HR = 1.57, 95 percent CI: 1.53, 1.62 and restricted-use data HR = 1.59, 95 percent CI: 1.55, 1.63). In the restricted-use data, compared to those who attained more than a high school education, those with less than a high school education have a HR = 1.37 (95 percent CI: 1.32, 1.42), and in the public-use data the estimates are essentially the same (HR = 1.36, 95 percent CI: 1.31, 1.41).

Table 4. Mortality from cancer by socio-demographic characteristics for non-Hispanic white and black adults aged 25 years and older in the 1986-2000 NHIS* linked mortality files, n = 802,387

	Public-use (deaths = 28,709)		Restricted-use (deaths = 28,679)	
	HR†	95% CI†	HR†	95% CI†
Age in years	1.07	1.07, 1.07	1.08	1.07, 1.08
Sex				
Women	1.00		1.00	
Men	1.57	1.53, 1.62	1.59	1.55, 1.63
Race/Ethnicity				
non-Hispanic white	1.00		1.00	
non-Hispanic black	1.17	1.13, 1.22	1.18	1.13, 1.22
Marital status				
Married	1.00		1.00	
Widowed	0.86	0.83, 0.90	0.87	0.84, 0.91
Divorced/separated	1.29	1.24, 1.35	1.29	1.23, 1.35
Never married	0.92	0.87, 0.98	0.94	0.89, 0.99
Education level				
Less than high school	1.36	1.31, 1.41	1.37	1.32, 1.42
High school/GED	1.24	1.20, 1.29	1.24	1.20, 1.29
More than high school	1.00		1.00	
Region				
Northeast	1.08	1.04, 1.13	1.08	1.04, 1.13
Midwest	1.04	1.00, 1.09	1.05	1.00, 1.09
South	1.09	1.05, 1.14	1.09	1.05, 1.14
West	1.00		1.00	

* National Health Interview Survey. Mortality follow-up through December 31, 2002.

† HR, hazard ratio; CI, confidence interval. Estimated from a Cox proportional hazards model.

In our analytic samples, homicide accounts for only 0.3 percent of deaths. Both the public-use and restricted-use files show similar results, but there is more variation in point estimates and their associated standard errors than for all-cause or the more common cause-specific mortality outcomes (table 5). In the restricted-use files, homicide mortality is 2.7 times as likely for men as women, 3.9 times as likely for non-Hispanic blacks compared to non-Hispanic whites, and 2.3 as likely for those with less than high school education compared to those with more than a high school degree. The hazard ratios in the public-use files are HR = 2.7, HR = 4.0, and HR = 2.4 for men, non-Hispanic blacks, and those with less than a high school education, respectively.

Table 5. Mortality from homicide by socio-demographic characteristics for non-Hispanic white and black adults aged 25 years and older in the 1986-2000 NHIS linked mortality files, n = 802,387

	Public-use (deaths = 320)		Restricted-use (deaths = 331)	
	HR†	95% CI†	HR†	95% CI†
Age in years	0.98	0.97, 0.99	0.99	0.98, 1.00
Sex				
Women	1.00		1.00	
Men	2.70	2.13, 3.42	2.70	2.14, 3.40
Race/Ethnicity				
non-Hispanic white	1.00		1.00	
non-Hispanic black	4.01	3.01, 5.33	3.90	2.92, 5.20
Marital status				
Married	1.00		1.00	
Widowed	1.26	0.70, 2.29	1.50	0.88, 2.57
Divorced/separated	1.60	1.15, 2.21	1.62	1.15, 2.27
Never married	1.88	1.32, 2.68	1.89	1.33, 2.69
Education level				
Less than high school	2.44	1.71, 3.50	2.31	1.63, 3.26
High school/GED	1.65	1.22, 2.23	1.55	1.16, 2.07
More than high school	1.00		1.00	
Region				
Northeast	0.46	0.30, 0.70	0.46	0.30, 0.71
Midwest	0.82	0.55, 1.20	0.80	0.54, 1.18
South	1.07	0.76, 1.52	1.03	0.72, 1.47
West	1.00		1.00	

* National Health Interview Survey. Mortality follow-up through December 31, 2002.

† HR, hazard ratio; CI, confidence interval. Estimated from a Cox proportional hazards model.

A comparison of the results for the public-use and restricted-use files for each of the 14 causes yields no substantive differences in conclusions, and hazard ratios and confidence intervals that are very similar. However, there tends to be less agreement in the estimates for the less common causes of death when comparing results from the public-use data and restricted-use data models. Results for all 14 causes of death as well as sex and race/ethnic specific analyses can be found at www.cdc.gov/nchs/data/datalinkage/nhis_mort_compare_2007_final.pdf (accessed May 5, 2008).

4 Discussion

With the release of public-use linked mortality files, NCHS has intended to balance the data needs of the research community while protecting the confidentiality of survey participants. The availability of nationally representative longitudinal mortality follow-up data that has high quality information on risk factors and socio-demographic characteristics is critical for health research. The updated mortality follow-up for the NHIS creates a prospective component to this cross-sectional data and the 2007 public-use release of the NHIS Linked Mortality Files expands access to this rich data source. The modifications made to the public-use file to allow its release include both limiting mortality information as compared to the restricted-use file and perturbing data for a select number of records.

The comparative analysis shows that the two data files yield similar descriptive and model results. Because the perturbation process in the public-use files did not affect the vital status of any individuals in the file, the only differences in results between the two files when examining all-cause mortality arose due to less specificity in the time to death information. The comparative analysis of cause-specific mortality across the public-use and restricted-use versions of the NHIS Linked Mortality Files also yielded only slight differences in model results, even for causes of death such as chronic liver disease and cirrhosis, homicide, unintentional injuries, and suicide. Also, since the release of the public-use linked mortality files, researchers concerned about whether their results might be affected by the data perturbation have contacted us to check their findings on the original data. We have made every attempt to comply with such requests, and have found, even with much smaller sample sizes, that findings are robust with no substantive changes in effect size or statistical significance. Yet, caution in using the public-use files is urged for researchers requiring more detail on timing of death or age or when examining the mortality patterns of small subgroups of the population, such as numerically small racial/ethnic minority groups, very old individuals, or young adults, or rare causes of death.

The new public-use version of the NHIS Linked Mortality Files provides the public health, social science, demographic, and medical communities with a data set that is readily available, very large, nationally representative, and rich in detail for both baseline covariates and specificity in outcomes. This paper makes a pragmatic, but unique, contribution to both the providers and users of data by discussing the issues related to confidentiality protection, demonstrating that the masking procedures implemented to reduce re-identification risk resulted in a public-use file with many of the variables that data users will need, e.g., information for the calculation of follow-up time and cause-of-death information, and providing a comparative analysis of the restricted-use and public-use versions of the data.

More information on NCHS's Data Linkage Activities and access to the public-use linked mortality data files for the National Health Interview Survey, the Third National Health and Nutrition Examination Survey and the Second Longitudinal Study of Aging can be found at the NCHS Data Linkage website, www.cdc.gov/nchs/ &d/nchs_datalinkage/data_linkage_activities.htm.

References

1. National Center for Health Statistics. Office of Analysis and Epidemiology. NCHS – Linked Mortality Files. Hyattsville, MD, `http://www.cdc.gov/nchrs/r&d/nchs_datalinkage/data_linkage_mortality.htm`
2. National Center for Health Statistics. Division of Vital Statistics. The National Death Index. Hyattsville, MD, `http://www.cdc.gov/nchs/ndi.htm`
3. Massey, J.T., Moore, T.F., Parsons, V.L., et al.: Design and Estimation for the National Health Interview Survey, 1985-1994. Vital and Health Statistics Report, Series 2 (1989)
4. Botman, S.L., Moore, T.F., Moriarity, C.L., et al.: Design and Estimation for the National Health Interview Survey, 1995–2004. Vital and Health Statistics Report, Series 2 (2000)

5. National Center for Health Statistics. Office of Analysis and Epidemiology. The 1986-2000 National Health Interview Survey Linked Mortality Files: Matching Methodology (2000), http://www.cdc.gov/nchs/data/datalinkage/matching_methodology_nhis_final.pdf
6. Fienberg, S.E., Willenborg, L.C.R.J.: Introduction to the Special Issue: Disclosure Limitation Methods for Protecting the Confidentiality of Statistical Data. JOS. 14, 337-345 (1998)
7. Domingo-Ferrer, J., Torra, V.: Disclosure Protection Methods and Information Loss for Microdata. In: Doyle, P., Lane, J., Theeuwes, J., et al. (eds.) Confidentiality, Disclosure and Data Access. North-Holland, Amsterdam (2001)
8. Winkler, W.E.: Masking and Re-identification Methods for Public-use Microdata: Overview and Research Problems. In: Domingo-Ferrer, J., Torra, V. (eds.) PSD 2004. LNCS, vol. 3050. Springer, Heidelberg (2004)
9. Lessler, J.T., Kalsbeek, W.D.: Nonsampling Errors in Surveys. John Wiley & Sons, Inc., New York (1992)
10. Cox, D.R.: Regression Models and Life Tables. Journal of the Royal Statistical Society 34, 187–220 (1972)
11. SUDAAN: Software for the Statistical Analysis of Correlated Data, 9.01. RTI International
12. National Center for Health Statistics. Health, United States, 2006. National Center for Health Statistics, Hyattsville, MD (2006)
13. Anderson, R.N., Minino, A.M., Hoyert, D.L., et al.: Comparability of Cause of Death between ICD-9 and ICD-10: Preliminary Estimates. National Vital Statistics Reports 49 (2001)
14. Rogers, R., Hummer, R.A., Nam, C.B.: Living and Dying in the U.S.A. Academic Press, San Diego (2000)

From *t*-Closeness to PRAM and Noise Addition Via Information Theory

David Rebollo-Monedero[1], Jordi Forné[1], and Josep Domingo-Ferrer[2]

[1] Telematics Engineering Dept., Technical University of Catalonia
C. Jordi Girona 1-3, E-08034 Barcelona, Catalonia
[2] UNESCO Chair in Data Privacy, Dept. of Computer Engineering and Maths,
Rovira i Virgili University
Av. Països Catalans 26, E-43007 Tarragona, Catalonia

Abstract. *t*-Closeness is a privacy model recently defined for data anonymization. A data set is said to satisfy *t*-closeness if, for each group of records sharing a combination of key attributes, the distance between the distribution of a confidential attribute in the group and the distribution of the attribute in the data is no more than a threshold *t*. We state here the *t*-closeness property in terms of information theory and then use the tools of that theory to show that *t*-closeness can be achieved by the PRAM masking method in the discrete case and by a form of noise addition in the general case.

Keywords: *t*-closeness, Microdata anonymization, Information theory, Rate distortion theory, PRAM, Noise addition.

1 Introduction

A microdata set is a data set whose records carry information on invidual respondents, like people or enterprises. The attributes in a microdata set can be classified as follows:

- *Identifiers.* These are attributes that *unambiguously* identify the respondent. Examples are passport number, social security number, full name, etc. Since our objective is to prevent confidential information from being linked to specific respondents, we will assume in what follows that, in a pre-processing step, identifiers have been removed/encrypted.
- *Key attributes.* Borrowing the definition from [1,2], key attributes are those that, in combination, can be linked with external information to re-identify (some of) the respondents to whom (some of) the records in the microdata set refer. Examples are job, address, age, gender, etc. Unlike identifiers, key attributes cannot be removed, because any attribute is potentially a key attribute.
- *Confidential outcome attributes.* These are attributes which contain sensitive information on the respondent. Examples are salary, religion, political affiliation, health condition, etc.

J. Domingo-Ferrer and Y. Saygın (Eds.): PSD 2008, LNCS 5262, pp. 100–112, 2008.

There are several privacy models to anonymize microdata sets. k-Anonymity [3,2] is probably the best known. However, it presents several shortcomings which have motivated the appearance of enhanced privacy models reviewed below. t-Closeness [4] is one of those recent proposals. Despite its conceptual appeal, t-closeness lacks computational procedures which allow to reach it with minimum data utility loss.

1.1 Contribution and Plan of This Paper

We state here t-closeness as an information-theoretic problem, in such a way that the knowledge body of information theory can be used to find a solution to it. The resulting solution turns out to be the PRAM masking method [5,6] in the discrete case and a form of noise addition in the general case.

Sec. 2 reviews the state of the art in k-anonymity-based privacy models. Sec. 3 gives an information-theoretic formulation of t-closeness. Sec. 4 is a theoretical analysis of the solution to t-closeness. Empirical results are reported in Sec. 5. Conclusions are drawn in Sec. 6.

2 Background and Motivation

k-Anonymity requires that each combination of key attribute values should be shared by at least k records in the data set. To enforce k-anonymity, at least there are two computational procedures: the original approach based on generalization and recoding of the key attributes and a microaggregation-based approach described in [7] and illustrated in Fig 1. While k-anonymity prevents identity disclosure (re-identification is infeasible within a group sharing the same key attribute values), it may fail to protect against identity disclosure: such is the case if the k records sharing a combination of key attribute values also share the value of a confidential attribute. Several enhancements of k-anonymity have been proposed to address the above and other shortcomings. Some of them are mentioned in what follows.

In [8], an evolution of k-anonymity called p-sensitive k-anonymity was presented. Its purpose is to protect against attribute disclosure by requiring that there be at least p different values for each confidential attribute within the records sharing a combination of key attributes. p-Sensitive k-anonymity has the limitation of implicitly assuming that each confidential attribute takes values uniformly over its domain, that is, that the frequencies of the various values of a confidential attribute are similar. When this is not the case, achieving p-sensitive k-anonymity may cause a huge data utility loss.

Like p-sensitive k-anonymity, l-diversity [9] was defined with the aim of solving the attribute disclosure problem that can arise with k-anonymity. A data set is said to satisfy l-diversity if, for each group of records sharing a combination of key attributes, there are at least l "well-represented" values for each confidential attribute. Depending on the definition of "well-represented", l-diversity can reduce to p-sensitive k-anonymity or be a bit more complex. However, it shares with the latter the problem of huge data utility loss. Also, it is insufficient to prevent attribute disclosure, because at least the following two attacks are conceivable:

- *Skewness attack.* If, within a group of records sharing a combination of key attributes, the distribution of the confidential attribute is very different from its distribution in the overall data set, then an intruder linking a specific respondent to that group may learn confidential information (*e.g.* imagine that the proportion of respondents with AIDS within the group is much higher than in the overall data set).
- *Similarity attack.* If values of a confidential attribute within a group are l-diverse but semantically similar (*e.g.* similar diseases or similar salaries), attribute disclosure also takes place.

t-Closeness [4] tries to overcome the above attacks. A microdata set is said to satisfy t-closeness if, for each combination of key attributes, the distance between the distribution of the confidential attribute in the group and the distribution of the attribute in the whole data set is no more than a threshold t. t-Closeness can be argued to protect against skewness and similarity (see [10] for a more detailed analysis):

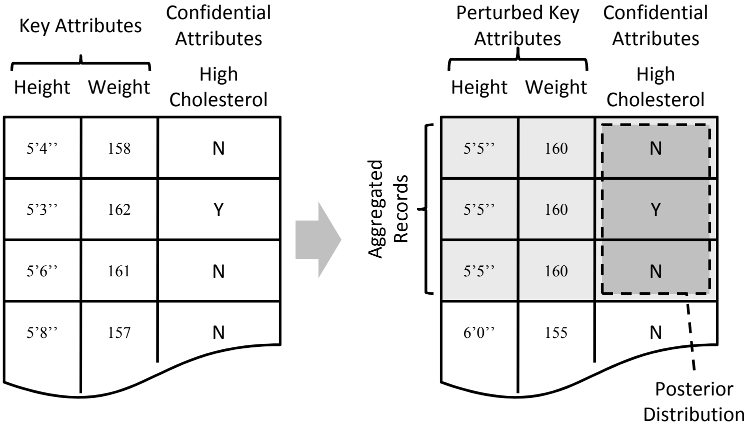
- To the extent to which the within-group distribution of confidential attributes resembles the distribution of those attributes for the entire dataset, skewness attacks will be thwarted.
- Again, since the within-group distribution of confidential attributes mimics the distribution of those attributes over the entire dataset, no semantic similarity can occur within a group that does not occur in the entire dataset. (Of course, within-group similarity cannot be avoided if all patients in a data set have similar diseases.)

The main limitation of the original t-closeness paper is that no computational procedure to reach t-closeness was specified. This is what we address in the remainder of this paper by leaning on the framework of information theory.

3 Information-Theoretic Formulation of t-Closeness

3.1 Conventions

Throughout the paper, the measurable space in which a random variable (r.v.) takes on values will be called an *alphabet*. All alphabets are assumed to be Polish spaces to ensure the existence of regular conditional probabilities, for example, any discrete space or the k-dimensional Euclidean space $\mathbb{R}^k$. We shall follow the convention of using uppercase letters for r.v.'s, and lowercase letters for particular values they take on. Probability density functions (PDFs) and probability mass functions (PMFs) are denoted by p, subindexed by the corresponding r.v. in case of ambiguity risk. For example, both $p_X(x)$ and $p(x)$ denote the value of the function p_X at x. The notation for information-theoretic quantities follows [11].



Height	Weight	High Cholesterol
5'4''	158	N
5'3''	162	Y
5'6''	161	N
5'8''	157	N

Height	Weight	High Cholesterol
5'5''	160	N
5'5''	160	Y
5'5''	160	N
6'0''	155	N

Fig. 1. Perturbation of key attributes to attain k-anonymity, t-closeness and similar privacy properties

3.2 Problem Statement

Let W and X be jointly distributed r.v.'s in arbitrary alphabets, possibly discrete, continuous, or mixed Cartesian products. In the problem of database t-closeness described above and depicted in Fig. 1, X represents (the tuple of) key attributes to be perturbed, which could otherwise be used to identify an individual. In the same application, confidential attributes containing sensitive information are denoted by W. Assume that the joint distribution of X and W is known, for instance, an empirical distribution directly drawn from a table, or a parametric statistical model inferred from a subset of records.

A *distortion measure* $d(x, \hat{x})$ is any measurable, nonnegative, real-valued function representing the distortion between the original data X and a perturbed version $\hat{X}$, the latter also a r.v., commonly but not necessarily in the same alphabet of X. The associated expected distortion $\mathcal{D} = \mathrm{E}\, d(X, \hat{X})$ provides a measure of utility of the perturbed data, in the intuitive sense that low distortion approximately preserves the values of the original data, and their joint statistical properties with respect to any other data of interest, in particular W. For example, if $d(x, \hat{x}) = \|x - \hat{x}\|^2$, then $\mathcal{D}$ is the mean-square error (MSE).

Consider now, on the one hand, the distribution p_W of the confidential information W, and on the other, the conditional distribution $p_{W|\hat{X}}$ given the observation of the perturbed attributes $\hat{X}$. In the database k-anonymization problem, whenever the posterior distribution $p_{W|\hat{X}}$ differs from the prior distribution p_W, we have actually gained some information about individuals statistically linked to the perturbed key attributes $\hat{X}$, in contrast to the statistics of the general population. Concordantly, define the *privacy risk* $\mathcal{R}$ as the Kullback-Leibler (KL) divergence D between the posterior and the prior distributions, that is, $\mathcal{R} = \mathrm{D}(p_{W|\hat{X}} \| p_W)$, which is one of the measures proposed in the original t-closeness paper [4].

Simple information-theoretic manipulations show that the privacy risk thus defined coincides with the mutual information [11] $\mathcal{R} = \mathrm{I}(W; \hat{X})$, and that both the KL divergence and the mutual information may be equivalently defined exchanging the roles of W and $\hat{X}$. Recall that the KL divergence vanishes (that is, one has 0-closeness) if, and only if, the distributions match (almost surely), which in turn is equivalent to requiring that W and $\hat{X}$ be statistically independent. Of course, in this extreme case, the utility of the published data, represented by the distribution $p_{W\hat{X}}$, usually by means of the corresponding table, is severely compromised. In the other extreme, leaving the original data undistorted, i.e., $\hat{X} = X$, compromises privacy, because in general $p_{W|X}$ and p_W differ.

We would like to remark at this point that the use of an information-theoretic quantity for database privacy assessment is by no means new. In addition to the t-closeness work already cited, [12, 13, 14] already used Shannon entropy as a measure of information loss, pointing out limitations affecting specific applications. We would like to stress out that we use a KL divergence as a measure of *information disclosure* (rather than loss), consistently with the equivalence between the case when $p_{W|\hat{X}} = p_W$ and the complete absence of privacy risk. On the other hand, the flexibility in our definition of distortion measure as a measure of *information loss* may enable us to preserve the statistical properties of the perturbed data to an arbitrary degree, possibly with respect to any other data of interest. Of course, the choice of distortion measure should ultimately rely on each particular application.

Consequently, we are interested in the tradeoff between two contrasting quantities, privacy and distortion, by means of perturbation of the original data. More precisely, consider *randomized perturbation rules* on the original data X, determined by the conditional distribution $p_{\hat{X}|X}$ of the perturbed data $\hat{X}$ given X. In the special case when the alphabets involved are finite, $p_{\hat{X}|X}$ may be regarded as a transition probability matrix, such as the one that appears in the PRAM masking method [5, 6]. The Markov chain $W \leftrightarrow X \leftrightarrow \hat{X}$, stating the conditional independence of $\hat{X}$ and W given X, emphasizes that this randomized rule has only X as input, but not W. Two remarks are in order. First, we consider randomized rules because deterministic quantizers are a particular case, and at this point we may not discard the possibility that more general rules attain a better tradeoff. Secondly, we consider rules that affect and depend on X only, but not W, for simplicity. Specifically, implementing and estimating convenient conditional distributions $p_{\hat{X}|WX}$ rather than $p_{\hat{X}|X}$ will usually be more complex, and require large quantities of data to prevent overfitting issues.

To sum up, we are interested in a randomized perturbation minimizing the privacy risk given a distortion constraint (or viceversa). In mathematical terms, we consistently define the *privacy-distortion function* as

$$\mathcal{R}(\mathcal{D}) = \inf_{\substack{p_{\hat{X}|X} \\ \mathrm{E}\, d(X,\hat{X}) \leqslant \mathcal{D}}} \mathrm{I}(W; \hat{X}). \tag{1}$$

For conceptual convenience, we provide an equivalent definition introducing an auxiliary r.v. Q, playing the role of randomized quantization index, a randomized quantizer $p_{Q|X}$, and a reconstruction function $\hat{x}(q)$:

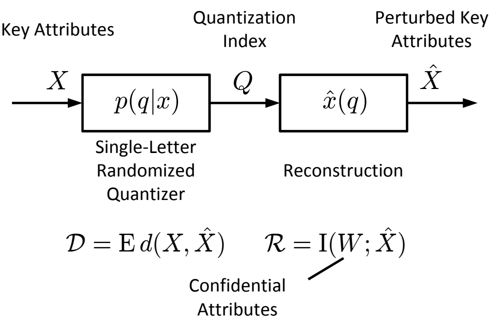


Fig. 2. Information-theoretic formulation of the privacy-distortion problem

$$\mathcal{R}(\mathcal{D}) = \inf_{\substack{p_{Q|X},\, \hat{x}(q) \\ \mathrm{E}\, d(X,\hat{X}) \leqslant \mathcal{D}}} \mathrm{I}(W;Q).$$

It can be shown [15] that there is no loss of generality in assuming that Q and $\hat{X}$ are related bijectively, thus $\mathrm{I}(W;Q) = \mathrm{I}(W;\hat{X})$, and that both definitions indeed lead to the same function. The elements involved in the definition of the privacy-distortion function are depicted in Fig. 2.

Even though the motivating application for this work is the problem of database t-closeness, it is important to notice that our formulation in principle addresses any applications where perturbative methods for privacy are of interest. Another illustrative application is privacy for location-based services (LBS). In this scenario, private information such as the user's location (or a sequence thereof) may be modeled by the r.v. X, to be perturbed, and W may represent a user ID. The posterior distribution $p_{\hat{X}|W}$ now becomes the distribution of the user's perturbed location, and the prior distribution $p_{\hat{X}}$, the population's distribution.

3.3 Connection with Information Theory

Perhaps the most attractive aspect of the formulation of the privacy-distortion problem in Sec. 3.2 is the strong resemblance it bears with the *rate-distortion problem* in the field of information theory. We shall see that our formulation is a generalization of a well-known, extensively studied information-theoretic problem with half a century of maturity. Namely, the problem of lossy compression of source data with a distortion criterion, first proposed by Shannon in 1959 [16].

To emphasize the connection, briefly recall that the simplest version of the problem of lossy data compression, shown in Fig. 3, involves coding of identically distributed (i.i.d.) copies $X_1, X_2, \ldots$ of a generic r.v. X. To this end, an n-letter deterministic quantizer maps blocks of n copies $X_1, \ldots, X_n$ into quantization indices Q in the set $\{1, \ldots, \lfloor 2^{n\mathcal{R}} \rfloor\}$, where $\mathcal{R}$ represents the coding rate in bits per sample. An estimation $\hat{X}_1, \ldots, \hat{X}_n$ of the source data vector is recovered to minimize the expected distortion per sample $\mathcal{D} = \frac{1}{n} \sum_i \mathrm{E}\, d(X_i, \hat{X}_i)$, according

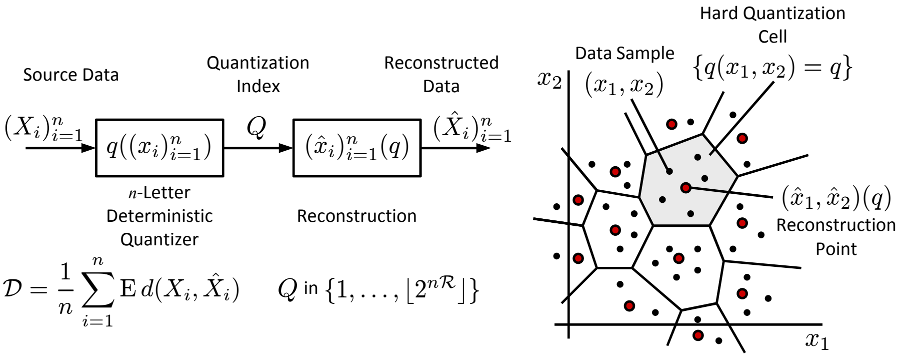


Fig. 3. Information-theoretic formulation of the rate-distortion problem

to some distortion measure $d(x, \hat{x})$. Intuitively, a rate of zero bits may only be achieved in the uninteresting case when no information is conveyed, whereas in the absence of distortion, the rate is maximized. Rate-distortion theory deals with the characterization of the optimal tradeoff between the rate $\mathcal{R}$ and the distortion $\mathcal{D}$, allowing codes with arbitrarily large block length n. Accordingly, the *rate-distortion function* is defined as the infimum of the rates of codes satisfying a distortion constraint.

A surprising and fundamental result of rate-distortion theory is that such function, defined in terms of blocks of samples, can be expressed in terms of a single copy of the source data vector [11], often more suitable for theoretical analysis. More precisely, the *single-letter characterization of the rate-distortion function* is

$$\mathcal{R}(\mathcal{D}) = \inf_{\substack{p_{\hat{X}|X} \\ \mathrm{E}\, d(X,\hat{X}) \leqslant \mathcal{D}}} \mathrm{I}(X;\hat{X}) = \inf_{\substack{p_{Q|X},\, \hat{x}(q) \\ \mathrm{E}\, d(X,\hat{X}) \leqslant \mathcal{D}}} \mathrm{I}(X;Q), \tag{2}$$

represented in Fig. 4. Aside from the fact that the equivalent problem is expressed in terms of a single letter X rather than n copies, there are two additional differences. First, the quantizer is randomized, and determined by a conditional distribution $p_{Q|X}$. Secondly, the rate is no longer the number of bits required to index quantization cells, or even the lowest achievable rate using an ideal

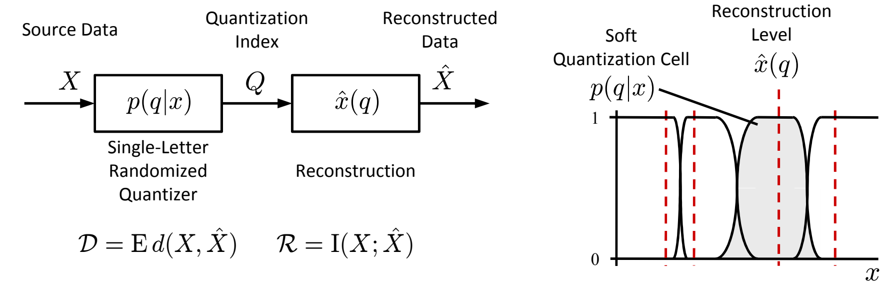


Fig. 4. Single-letter characterization of the rate-distortion problem

entropy coder, namely the entropy of the quantization index $\mathrm{H}(Q)$. Instead, the rate is a mutual information $\mathcal{R} = \mathrm{I}(X;\hat{X})$.

Interestingly, the single-letter characterization of the rate-distortion function (2) is almost identical to our definition of privacy-distortion function (1), except for the fact that in the latter there is an extra variable W, the confidential attributes, in general different from X, the key attributes. It turns out that some of the information-theoretic results and methods for the rate-distortion problem can be extended, with varying degrees of effort, to the privacy-distortion problem formulated in this work. Some of these extensions are discussed in the next section.

4 Theoretical Analysis

All theoretical claims in this section are detailed and proven in [15].

Similarly to the rate-distortion function, the privacy-distortion function (1) is decreasing, convex, and continuous in the interior of its domain. Furthermore, the optimization problem determining (1), with $p_{\hat{X}|W}$ as unknown variable, is itself convex. This means that any local minimum is also global, and makes the powerful tools of convex optimization [17] applicable to compute numerically but efficiently the privacy-distortion function. In Sec. 5, an example of numerical computation will be discussed.

While a general closed-form expression for privacy-distortion function has not been provided, the Shannon lower bound for the rate-distortion function can be extended to find a closed-form lower bound under certain assumptions. Furthermore, the techniques used to prove this bound may yield an exact closed formula in specific cases. A closed-form upper bound is also presented in this section.

Suppose that W and X are real-valued r.v.'s (random scalars), and that MSE is used as distortion measure, thus $\mathcal{D} = \mathrm{E}(X - \hat{X})^2$. Define the normalized distortion $d = \frac{\mathcal{D}}{\sigma_X^2}$, where σ_X^2 denotes the variance of X. Let σ_W^2 be the variance of W, ρ_{WX} the correlation coefficient of W and X, and $\mathrm{h}(W)$ the differential entropy [11] of W. Then,

$$\mathcal{R}(\mathcal{D}) \geqslant \mathcal{R}_{\mathrm{QGLB}}(\mathcal{D}) = \mathrm{h}(W) - \frac{1}{2}\log\left(2\pi e\left(1-(1-d)\rho_{WX}^2\right)\sigma_W^2\right) \qquad (3)$$

for $0 \leqslant d \leqslant 1$ (for $d \geqslant 1$, clearly $\mathcal{R} = 0$). We shall call the bounding function $\mathcal{R}_{\mathrm{QGLB}}(\mathcal{D})$ the *quadratic-Gaussian lower bound* (QGLB).

With the same assumptions, namely scalar r.v.'s and MSE distortion measure, consider the two trivial cases $d = 0$ and $d = 1$. The former can be achieved with $\hat{X} = X$, yielding $\mathcal{R}(\mathcal{D}) = \mathrm{I}(W;X)$, and the latter with $\hat{X} = \mu_X$, the mean of X, for which $\mathcal{R}(\mathcal{D}) = 0$. Now, for any $0 \leqslant d \leqslant 1$, set $\hat{X} = X$ with probability $1-d$, and $\hat{X} = \mu_X$ with probability d. Convexity properties of the mutual information guarantee that the privacy-distortion performance of this setting cannot lie above the segment connecting the two trivial cases. Since the setting is not necessarily optimal, it may be concluded that

$$\mathcal{R}(\mathcal{D}) \leqslant \mathcal{R}_{\mathrm{MIUB}}(\mathcal{D}) = \mathrm{I}(W;X)(1-d). \qquad (4)$$

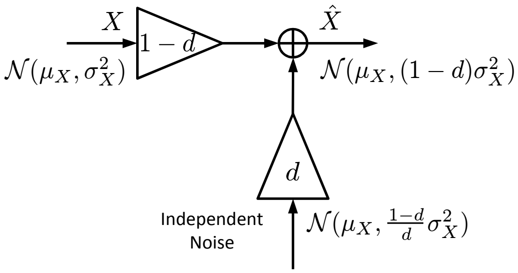

X
$\mathcal{N}(\mu_X, \sigma_X^2)$
$1-d$
$\hat{X}$
$\mathcal{N}(\mu_X, (1-d)\sigma_X^2)$
d
Independent Noise
$\mathcal{N}(\mu_X, \frac{1-d}{d}\sigma_X^2)$

Fig. 5. Optimal randomized perturbation in the quadratic-Gaussian case

We shall call this bounding function the *mutual-information upper bound* (MIUB). The $p_{\hat{X}|X}$ determined by the combination of the two trivial cases for intermediate values of d may be a simple yet effective way to initialize numerical search methods to compute the privacy-distortion function, as it will be shown in Sec. 5.

Provided that W and X are jointly Gaussian, real-valued r.v.'s, and that MSE is used as distortion measure, the QGLB (3) is tight:

$$\mathcal{R}(\mathcal{D}) = -\frac{1}{2} \log \left(1 - (1-d)\rho_{WX}^2\right), \tag{5}$$

with $d = \frac{\mathcal{D}}{\sigma_X^2} \leqslant 1$ as before. The optimal randomized perturbation rule achieving this privacy-distortion performance is represented in Fig. 5. Observe that the perturbed data $\hat{X}$ is a convex combination of the source data X and independent noise, in a way such that the final variance achieves the distortion constraint with equality.

5 Numerical Computation Example

In this section, we illustrate the theoretical analysis of Sec. 4 with experimental results for a simple, intuitive case. Specifically, W and X are jointly Gaussian random scalars with correlation coefficient ρ (after zero-mean, unit-variance normalization). In terms of the database microaggregation problem, W represents sensitive information, and X corresponds to key attributes that can be used to identify specific individuals. These variables could model, for example, the plasma concentration of LDL cholesterol in adults, which is approximately normal, and their weight, respectively. MSE is used as a distortion measure. For convenience $\sigma_X^2 = 1$, thus $\mathcal{D} = d$. Since the privacy-distortion function is convex, minimization of one objective with a constraint on the other is equivalent to the minimization of the Lagrangian cost $\mathcal{C} = \mathcal{D} + \lambda\mathcal{R}$, for some positive multiplier λ. We wish to design randomized perturbation rules $p_{\hat{X}|X}$ minimizing $\mathcal{C}$ for several values of λ, to investigate the feasibility of numerical computation of the privacy-distortion curve, and to verify the theoretic results for the quadratic-Gaussian case of Sec. 4.

We implement a slight modification of a simple optimization technique, namely the steepest descent algorithm, operating on a sufficiently fine discretization of the variables involved. More precisely, p_{WX} is the joint PMF obtained by discretizing the PDF of W and X, where each variable is quantized with 31 samples in the interval $[-3, 3]$. The starting values for $p_{\hat{X}|X}$ are convex combinations of the extreme cases corresponding to $d = 0$ and $d = 1$, as described in Sec. 4 when the MIUB (4) was discussed. Only results corresponding to the correlation coefficient $\rho = 0.95$ are shown, for two reasons. First, because of their similarity with results for other values of ρ. Secondly, because for high correlation, the gap between the MIUB (which approximates the performance of the starting solutions) and the QGLB (3) is wider, leading to a more challenging problem.

The definitions of distortion and privacy risk in Sec. 3 for the finite-alphabet case become

$$\mathcal{D} = \sum_x \sum_{\hat{x}} p(x) p(\hat{x}|x) d(x, \hat{x}), \qquad \mathcal{R} = \sum_w \sum_{\hat{x}} p(w) p(\hat{x}|w) \ln \frac{p(\hat{x}|w)}{p(\hat{x})}.$$

The conditional independence assumption in the same section enables us to express the PMFs of $\hat{X}$ in the expression for $\mathcal{R}$ as $p(\hat{x}) = \sum_x p(\hat{x}|x)p(x)$ and $p(\hat{x}|w) = \sum_x p(\hat{x}|x)p(x|w)$, in terms of the optimization variables $p(\hat{x}|x)$. Our implementation of the steepest descent algorithm uses the exact gradient with components $\frac{\partial \mathcal{C}}{\partial p(\hat{x}|x)} = \frac{\partial \mathcal{D}}{\partial p(\hat{x}|x)} + \lambda \frac{\partial \mathcal{R}}{\partial p(\hat{x}|x)}$, where $\frac{\partial \mathcal{D}}{\partial p(\hat{x}|x)} = p(x) d(x, \hat{x})$ and

$$\frac{\partial \mathcal{R}}{\partial p(\hat{x}|x)} = p(x) \left(\sum_w p(w|x) \ln p(\hat{x}|w) - \ln p(\hat{x}) \right)$$

(after simplification [15]).

Two modifications of the standard version of the steepest descent algorithm [17] were applied. First, rather than updating $p_{\hat{X}|X}$ directly according to the negative gradient multiplied by a small factor, we used its projection onto the affine set of conditional probabilities satisfying $\sum_{\hat{x}} p(\hat{x}|x) = 1$ for all x, which in fact gives the steepest descent within that set. Secondly, rather than using a barrier or a Lagrangian function to consider the constraint $p(\hat{x}|x) \geqslant 0$ for all x and $\hat{x}$, after each iteration, we reset possible negative values to 0 and renormalized the probabilities accordingly. This may seem unnecessary since the theoretical analysis in Sec. 4 gives a strictly feasible solution (i.e., probabilities are strictly positive), and consequently the constraints are inactive. However, the algorithm operates on a discretization of the joint distribution of W and X in a machine with finite precision. The fact is that precision errors in the computation of gradient components corresponding to very low probabilities activated the nonnegativity constraints. Finally, we observed that the ratio between the largest and the smallest eigenvalue of the Hessian matrix was large enough for the algorithm to require a fairly small update factor, 10^{-4}, to prevent significant oscillations.

The privacy-distortion performance of the randomized perturbation rules $p_{\hat{X}|X}$ found by our modification of the steepest descent algorithm is shown in Fig. 6, along with the bounds established in Sec. 4, namely the QGLB (3) and the MIUB (4). On account of (5), it can be shown that $\lambda = 2\sigma_X^2 \left(1/\rho^2 - 1 + d\right)$. Accordingly, we set λ approximately to 0.72, 1.22 and 1.72, which theoretically corresponds to $d = 0.25, 0.5, 0.75$. A total of 32000 iterations were computed for each value of λ,

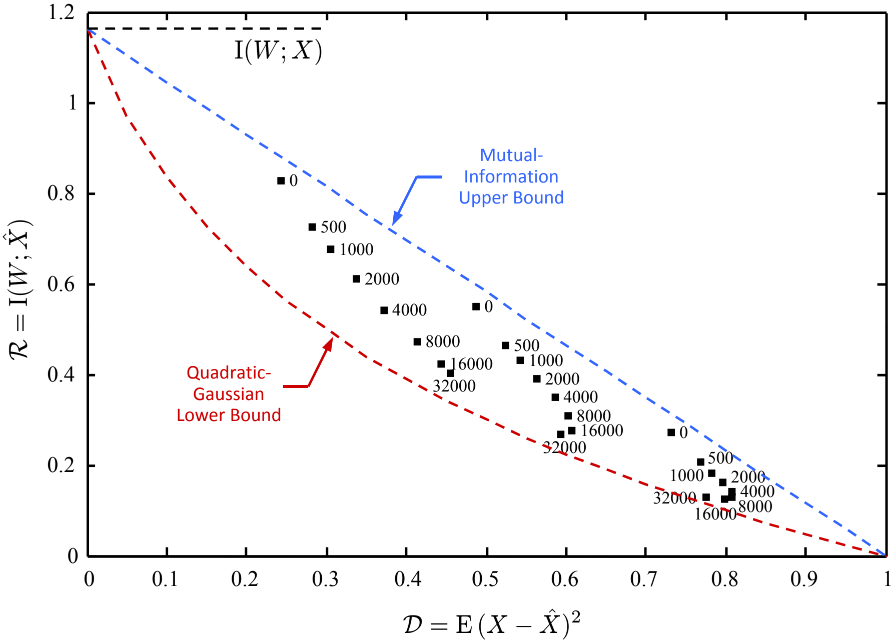


Fig. 6. Privacy-distortion performance of randomized perturbation rules found by a modification of the steepest descent algorithm

at about 16 iterations per second on a modern computer[1]. The large number of iterations is consistent with the fact that the Hessian is ill-conditioned and the small updating step size. Obviously, one would expect that methods based on Newton's technique [17] converge to the optimal solution in less iterations (at the cost of higher computational complexity per iteration), but our goal was to check the performance of one of the simplest optimization algorithms. In all cases, the conditional PMFs found had a performance very close to that described by (5) in Sec. 4. Their shape, depicted in Fig. 7, roughly resembled the Gaussian shape predicted by the theoretical analysis as the number of iterations increased. Specifically, Fig. 7 corresponds to $\lambda \simeq 1.22$, was obtained after 32000 iterations, and the number of discretized samples of X and W was increased from 31 to 51. Increasing the number of iterations to 128000 resulted in an experimental solution shaped almost identically to the optimal one, although the one in Fig. 7, corresponding to a fourth of the number of iterations, already achieves values of $\mathcal{C}$ reasonably optimal.

[1] Implementation used Matlab R2007b on Windows Vista SP1, on an Intel Core2 Quad Q6600 CPU at 2.4 GHz.

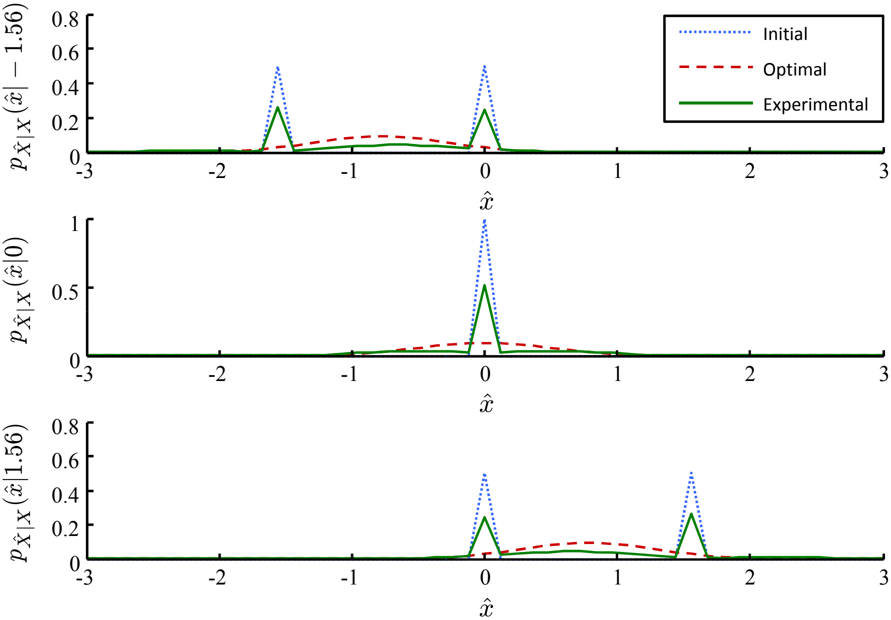


Fig. 7. Shape of initial, optimal, and experimental randomized perturbation rules $p_{\hat{X}|X}$ found by the steepest descent algorithm

6 Conclusion

An information-theoretic formulation of the privacy-distortion tradeoff in applications such as microdata anonymization and location privacy in location-based services is provided. Following the *t*-closeness model, the privacy risk is measured as the mutual information between perturbed key attributes and confidential attributes, equivalent to the KL divergence between posterior and prior distributions. We consider the problem of maximizing privacy (that is, minimizing the above mutual information) while keeping the perturbation of data within a pre-specified bound to ensure that data utility is not too damaged. We establish a strong connection between this privacy-perturbation problem and the rate-distortion problem of information theory and extend of a number of results, including convexity of the privacy-distortion function and the Shannon lower bound. A closed formula is obtained for the quadratic-Gaussian case, proving that the optimal perturbation is randomized rather than deterministic, which justifies the use of PRAM in the case of attributes with finite alphabets or noise addition in the general case.

Acknowledgments and Disclaimer

We would like to thank the anonymous reviewers for their valuable comments. This work was partly supported by the Spanish Government through projects

CONSOLIDER INGENIO 2010 CSD2007-00004 "ARES", TSI2007-65393-C02-02 "ITACA" and TSI2007-65406-C03-01 "E-AEGIS", and by the Government of Catalonia under grants 2005 SGR 00446 and 2005 SGR 01015. The third author is with the UNESCO Chair in Data Privacy, but his views do not necessarily reflect the position of UNESCO nor commit that organization.

References

1. Dalenius, T.: Finding a needle in a haystack - or identifying anonymous census records. Journal of Official Statistics 2(3), 329–336 (1986)
2. Samarati, P.: Protecting respondents identities in microdata release. IEEE Transactions on Knowledge and Data Engineering 13(6), 1010–1027 (2001)
3. Samarati, P., Sweeney, L.: Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression. Technical report, SRI International (1998)
4. Li, N., Li, T., Venkatasubramanian, S.: t-closeness: Privacy beyond k-anonymity and l-diversity. In: Proc. IEEE Int. Conf. Data Eng (ICDE), Istanbul, Turkey, April 2007, pp. 106–115 (2007)
5. Gouweleeuw, J.M., Kooiman, P., Willenborg, L.C.R.J., DeWolf, P.P.: Post randomisation for statistical disclosure control: Theory and implementation, Research paper no. 9731 (Voorburg: Statistics Netherlands) (1997)
6. de Wolf, P.P.: Risk, utility and PRAM. In: Domingo-Ferrer, J., Franconi, L. (eds.) PSD 2006. LNCS, vol. 4302. Springer, Heidelberg (2006)
7. Domingo-Ferrer, J., Torra, V.: Ordinal, continuous and heterogenerous k-anonymity through microaggregation. Data Mining and Knowledge Discovery 11(2), 195–212 (2005)
8. Truta, T.M., Vinay, B.: Privacy protection: p-sensitive k-anonymity property. In: 2nd International Workshop on Privacy Data Management PDM 2006, Atlanta, GA, p. 94. IEEE Computer Society, Los Alamitos (2006)
9. Machanavajjhala, A., Gehrke, J., Kiefer, D., Venkitasubramanian, M.: L-diversity: privacy beyond k-anonymity. In: Proceedings of the IEEE ICDE 2006 (2006)
10. Domingo-Ferrer, J., Torra, V.: A critique of k-anonymity and some of its enhancements. In: Proceedings of ARES/PSAI 2008, pp. 990–993. IEEE Computer Society, Los Alamitos (2008)
11. Cover, T.M., Thomas, J.A.: Elements of Information Theory. Wiley, New York (1991)
12. Kooiman, P.L., Willenborg, L., Gouweleeuw, J.: PRAM: A method for disclosure limitation of microdata. Research Rep. 9705, Statistics Netherlands, Voorburg, NL (1998)
13. de Waal, T., Willenborg, L.: Information loss through global recoding and local suppression. Netherlands Official Stat. 14, 17–20 (1999)
14. Willenborg, L., de Waal, T.: Elements of Statistical Disclosure Control. Springer, New York (2001)
15. Rebollo-Monedero, D., Forné, J.: An information-theoretic formulation of the privacy-distortion tradeoff. Research rep., Tech. Univ. of Catalonia (UPC) (June 2008)
16. Shannon, C.E.: Coding theorems for a discrete source with a fidelity criterion. IRE Nat. Conv. Rec. 7(4), 142–163 (1959)
17. Boyd, S., Vandenberghe, L.: Convex Optimization. Cambridge University Press, Cambridge (2004)

Robustification of Microdata Masking Methods and the Comparison with Existing Methods

Matthias Templ[1,2] and Bernhard Meindl[1]

[1] Department of Methodology, Statistics Austria, Guglgasse 13, 1110 Vienna, Austria
bernhard.meindl@statistik.gv.at

[2] Department of Statistics and Probability Theory, Vienna University of Technology,
Wiedner Hauptstr, 8-10, 1040 Vienna, Austria
templ@statistik.tuwien.ac.at

Abstract. The aim of this study was to compare different microdata protection methods for numerical variables under various conditions. Most of the methods used in this paper have been implemented in the R-package *sdcMicro* which is available for free on the comprehensive R archive network (http://cran.r-project.org). The other methods used can be easily applied using other R-packages. While most methods work well for homogeneous data sets, some methods fail completely when confidential variables contain outliers which is almost always the case with data from official statistics. To overcome these problems we have robustified popular methods such as microaggregation or shuffling which is based on a regression model. All methods have beed tested on bivariate data sets featuring different outlier scenarios. Additionally, a simulation study was performed.

Keyword: Statistical disclosure control, numerical data protection methods, robustness, simulation.

1 Introduction

One of the most important steps in the process of data anonymization is to anonymize the categorical variables. This means to anonymize indirect identifiers with respect to the sampling weights.

However, an attacker may try to identify statistical units using numerical variables by using linking and/or matching procedures as well. The anonymisation of numerical variables makes it more difficult for the attacker to successfully match or merge underlying data with other data sources. Therefore, the anonymisation of numerical variables is of high interest too.

Please note that R-package *sdcMicro* contains a large amount of methods for the protection of categorical data (see, e.g., [1] and [2] for a practical application). In the following work we will focus on methods for anonymizing numerical variables such as microaggregation, adding noise, swapping and shuffling.

Serveral perturbation methods for numerical data (microaggregation via z-transformation, rank swapping, resampling, generation of synthetic data based

J. Domingo-Ferrer and Y. Saygın (Eds.): PSD 2008, LNCS 5262, pp. 113–126, 2008.

on samples that are generated from the empirical mean and covariance structure of the data) have been compared within a simulation framework using bivariate data, for example, by [3].

However, [4] notes that better microaggregation procedures exist. Furthermore, rank swapping destroys the multivariate data structure ([4]). Synthetic data are approximately normal distributed but the original data do not follow and can not be transformed to follow a normal distribution in general.

Another simulation study was carried out in [5] where swapping was compared to shuffling on data sets generated from bivariate normal distributions featuring different correlations. A similar simulation study was conducted in [6] where the distribution of the confidential variables followed other theoretical distributions, but no outliers were included.

The results of [3] are based on random data sampled from theoretical distributions without contamination. But in real world applications we often can not assume that the data follow a theoretical distribution. Therefore, perturbation methods must also give reasonable results when outlier exist in the data.

In the next section we introduce the methods that we have evaluated. Details can be found in the references and the implementation can be found in the R-package *sdcMicro* ([7], [8], [1]) which can be downloaded on the comprehensive R archive network (CRAN, `http://cran.r-project.org`, see [9]).

1.1 Methods Used in This Study

We now give an overview on the methods that have been investigated. The methods can be classified into four groups: methods based on sorting, methods based on grouping, methods based on adding noise and methods for synthetic data generation.

In Table 1 we give references for each method as well as the the name of the corresponding function in R-package *sdcMicro*. Some functions (e.g. addNoise, microaggregation) are wrapper functions for several methods which are listed in Table 1 as well. Furthermore, the parameters used in this study are given in the last column of the table since they are different from the default ones. The parameters and their default choices are described in detail in the package manual of *sdcMicro*. Please note that more method especially on synthetic data generation are implemented in *sdcMicro*.

Methods based on adding noise: One possible procedure is to add additive noise (cited as *noise* in Tables and Figures) to each numerical variable

$$Y = X + \epsilon \quad ,$$

where $X \sim (\mu, \Sigma), \epsilon \sim N(0, \Sigma_\epsilon), \quad \Sigma_\epsilon = \alpha \cdot diag(\sigma_1^2, \sigma_2^2, \ldots, \sigma_p^2), \alpha > 0$, $Cov(\epsilon_i \neq \epsilon_j)$ $\forall i \neq j$ and p is equal the dimension of the numerical variables which should be perturbed (see e.g. also in [10]).

It is clear that multivariate measures such as the correlation coefficient can not be preserved after adding additive (uncorrelated) noise. Correlation coefficients can however be preserved when correlated noise is added. In this case the covariance matrix of the masked data is $\Sigma_Y = (1 + \alpha)\Sigma_X$ (see e.g. in [10]).

Table 1. Investigated methods. Only those references to papers are listed which have been used for the (re-)implementation of the methods in *sdcMicro*

Method description	Reference	Function in *sdcMicro*	Method	Parameter used
adding additive noise	[10]	addNoise	additive	noise=200
correlated noise	[10]	addNoise	correlated	noise=200
correlated noise 2	[11], [12]	addNoise	correlated2	noise=200
restricted correlated noise	[13]	addNoise	restr	noise=200
ROMM	[14]	addNoise	ROMM	noise= $200, p = 0.01$
adding noise based on multivariate outlier detection	[4]	addNoise	outdect	noise=200
rank swapping	[15]	swappNum	-	p=15, p=40
ma, sort on a single variable	-	microaggregation	single	default
ma, individual ranking	[16]	microaggregation	onedims	default
ma, influence	[4]	microaggregation	influence	default
ma, sorting on first principal component	-	microaggregation	pca	default
ma, sorting with projection pursuit pca	[4], [8]	microaggregation	pppca	default
ma, sorting with projection pursuit pca on clustered data	[4], [8]	microaggregation	clustpppca	default
ma, mdav	[12]	microaggregation	mdav	default
ma, multivariate microaggregation based on robust Mahalanobis distances	[8]	microaggregation	rmd	default
gadp	[17], [6]	shuffle (output *gadp*)	gadp	default
shuffling	[6]	shuffle (output *shuffle*)	shuffle	default
shuffling based on cgadp	[6]	shuffle2	shuffle2	default
robust gadp	this paper	robShuffle (output *gadp*)	robGadp	default
robust shuffling	this paper	robShuffle (output *shuffle*)	robShuffle	default

For method *correlated2* $d = \epsilon(1 - \alpha^2)$ and then $x_j d + \alpha z_j$ is calculated where z_j are random numbers drawn from $N(\frac{(1-d)\bar{x}_j}{\alpha}, s_j)$ with s_j being the standard deviation of X_j (see e.g. in [11] or [12]).

The restricted correlated noise method (implemented as method *restr* in *sdcMicro*) is a similar method that takes the sample size into account ([13]).

Furthermore, method ROMM (Random Orthogonal Matrix Masking, [14]) has been considered in the simulation study. In this method perturbed data are obtained by the transformation $Y = AX$ where A is randomly generated and fulfills the orthogonality condition $A^{-1} = A^T$. To obtain a orthogonal matrix as described in [14] the Gram-Schmidt procedure was chosen in the computional implementation of method *ROMM*.

In order to be able to deal with inhomogeneous data sets including outliers the first author of the paper has implemented a method in which outliers are detected. Observations with large robust Mahalanobis distances are treated as outliers as well as observations that exhibit univariate outliers, i.e. where the values of a variable x are greater than robust measure of location (e.g. the median) plus a robust measure of scatter (usually the Median Absolute Deviation (see [18])).

Outliers should be protected more (by default by adding additive noise) than the rest of the observations because outliers show a higher risk for re-identification than non-outliers. This method is denoted *outdect* in package *sdcMicro* and in the following text.

Methods based on sorting: Rank swapping (see [15] and [19]) is a method based on sorting a variable by their numerical values (ranking). Each ranked value is then swapped with another ranked value that has been chosen randomly within a restricted range. This means for example that the rank of two swapped values cannot differ by more than p percent of the total number of observations. Rank swapping must be applied to each variable separately.

Methods based on sorting and grouping: A familiar definition of microaggregation can be found at http://neon.vb.cbs.nl/casc/Glossary.htm: "Records are grouped based on a proximity measure of variables of interest, and the same small groups of records are used in calculating aggregates for those variables. The aggregates are released instead of the individual record values." The choice of the "proximity" measure is the most challenging and most important part in microaggregation since the multivariate structure of the data is only preserved if similar observations are aggregated. Sorting data based on one single variable in ascending or descending order (method *single* in *sdcMicro*), sorting the observations in each cluster (after clustering the data) by the most influencial variable in each cluster (method *influence*, see [4]) as well as sorting (and re-ordering the data after aggregation) in each variable (individual ranking method, see [16]) is not optimal for multivariate data (see [8]).

Projection methods which sort the data according to the first principal component (method *pca*) or its robust counterpart (method *pppca*, see [4]) can be improved when these methods are applied to clustered data (method *clustpppca*, see [1], for example). In order to estimate the first principal component in a robust way, it is necessary to obtain a a robust estimate of the covariance matrix. However, this is only feasible for small or medium sized data sets using methods like M-estimation [20], the MVE estimator [21] or the orthogonalized Gnanadesikan-Kettering (OGK) estimator [22]. Furthermore, all principal components must be estimated when using classical approaches for PCA. Method *pppca* avoids this and estimates the first (robust) principal component without the need of estimating the covariance.

The Maximum Distance to Average Vector (MDAV) method is an evolution of the multivariate fixed-size microaggregation (see [23], for example). This method (*mdav* in *sdcMicro*) is based on Euclidean distances in a multivariate space.

The algorithm has been improved by the first author of the paper by replacing Euclidean distances with robust Mahalanobis distances. In [1] a new algorithm called RMDM (**R**obust **M**ahalanobis **D**istance based **M**icroaggregation) was proposed for microaggregation where MDAV [23] is adapted in several ways. The proposed procedure (details can be found in [8]) is a more natural approach than MDAV since multivariate data are dealt with by taking the (robust) covariance structure of the data into account.

Methods based on models: GADP (method *gadp* in *sdcMicro*) is based on the model $Y = S\beta + \epsilon$ and was originally proposed by [17]. Y are perturbed variables, S are non-confidential variables and $\epsilon \sim MVN(0,\ \Sigma_{XX} - \Sigma_{XS}(\Sigma_{SS})^{-1}\Sigma_{SX})$. The regression coefficients β are estimated by $\hat{\beta} = (S^{'}S)^{-1}S'X$ using X, the confidential variables. The procedure is described in detail in http://gatton.uky.edu/faculty/muralidhar/CDAC42.ppt.

Since correlations between S and X have to be estimated in the procedure, [6] refers to choose the Spearman rank correlation measure when dealing with non-normal distributed data. Shuffling is finally done by replacing the rank ordered values of Y (generated by GADP) with the rank ordered values of X.

Another approach is the copula based GADP which is described in [6] and which was investigated too. It is assumed that the data follow a theoretical distribution from which the inverse distribution function can be expressed in analytical form. This approach is only feasible if the data follow a theoretical distribution approximately. However, in real world data sets this assumption can hardly be justified since data almost always include outliers. Therefore, similar conclusions as in the GADP approach have been obtained since this copula based approach cannot deal with outlying observations.

These very simple methods are under US-Patent (7200757) and so the re-implementation of the method is not included in package *sdcMicro*. Nevertheless, the scientific work must go on and therefore the first author of the paper has implemented a - not patented - extension of the procedure which can deal with outliers as well (see section 2).

Synthetic data: In package *sdcMicro* a method based on the Cholesky decomposition for fast generation of synthetic data has been re-implemented with which multivariate normal distributed data can be generated with respect to the covariance structure of the original data (details can be found in [24]). However, this does not reflect the distribution of real complex data since such data are generally not multivariate normal distributed and include outliers. Thus, this method was not considered in the simulation study.

Another method which is based on regression is called the IPSO synthetic data generators (see [25] or [26] for an application). However, the generation of synthetic data using regression models will fail when dealing with inhomogeneous data sets including outliers. This method gives reasonable results only for data that are multivariate normal distributed. For this reason, this method is not further addressed in this paper.

A well known method is blanking and imputation ([27]). Since outlying observations may be identified easier than non-outliers it is clear that such a method must blank and impute the outliers in any case. However, an imputation of outliers without destroying the multivariate structure of the data is difficult.

[28] suggested to generate a completely synthetic data set based on the original survey data and multiple imputation. This method is not addressed in this simulation study as well.

Latin Hypercube Sampling ([29], [30], [31]) was also not considered in the paper because this method gives worst results and we are not sure if this results from a mistake in our re-implementation of the method. Unfortunately, also the results based on a MATLAB code from another author are worst ([32]). This is not a surprise because the inverse of the distribution function must be available for this method. An analytical form of the inverse of the distribution function is rather complex or impossible to find for inhomogeneous data sets.

1.2 Information Loss

First we want to overview existing measures of information loss which are almost always univariate measures. In this paper we concentrate on multivariate measures of information loss which evaluate the multivariate structure of the original and the perturbed data.

One measure of information loss, called IL1s (see e.g. in [33] or [34]) and is based on aggregated distances from original data points to corresponding values from the perturbed data divided by the standard deviation for each variable. Unfortunately, this measure is large even if only one outlier was perturbed highly and all values are exactly the same as in the original data set.

Other measures are considered in [12] and [4]. Measures of information loss which compare univariate statistics of the original data and the perturbed data are for example the sum of the differences of the mean or medians. Measures which compare multivariate statistics of the original data and the perturbed data evaluate differences of the correlation matrices or loadings in principal component analysis. In our study we compare the eigenvalues of the classical covariance and the robust covariance [21] of the original data with the ones from the perturbed data. Other kinds of measures of information loss are discussed in [3].

Improved measures of information loss has been suggested by [35] and are also implemented in *sdcMicro* as well as robust measures (see in the package manual of *sdcMicro*).

1.3 Disclosure Risk

In [36] a measure of disclosure risk is proposed which is based on distances and assumes that an intruder has additional information (disclosure scenarios) so that he can link the masked record of an observation to its original value (see e.g. [34]). Given the value of a masked variable it is checked whether the corresponding original value falls within an interval centered on the masked value. The width of the interval is based on the rank of the variable or on its standard deviation [36]. However, this interval does not depend on the scale of the actual value and therefore the length of the interval is equal for non-outlying and outlying values. However, outlying observations should be much more perturbed than non-outliers.

Another type of measures of disclosure risk - the value disclosure risk - is extensively used, e.g. by [5]. The main goal of this measure is to evaluate the gain in explanation of parameters or variables when perturbed data are released.

[37] suggests a new and more realistic measures of disclosure risk which accounts for outlying observations by using robust Mahalanobis distances. The robustification was done using the MCD-estimator [21].

2 Robustification of GADP and Shuffling

When classical methods are fitted to data which include outliers one may get unrealistic estimation results. On the other hand robust methods are able to describe the majority of the data and to detect outliers. In the context of gadp and shuffling it is then possible to perturbate this majority of the data reasonable but it is not possible to perturbe or generate the outliers in a reasonable way. Therefore high data utiltiy cannot be reached with non-robust nor with robust procedures. Therefore, we propose to use robust procedure both for the non-outliers and for the outliers seperately. Please note that when using mixture models to describe the data the same problems with outliers occur.

It is easy to see that classical shuffling procedures are influenced by outlying observations because within the procedure a rank based covariance matrix estimation and a least squares regression fitting is applied.

In order to robustify shuffling one has to modify the following conditions:

- Choose a robust regression method with a high breakdown point instead of least squares estimation. The *breakdown point* of an estimator measures the maximal percentage of the data points which may have been contaminated before the estimates become completely corrupted.

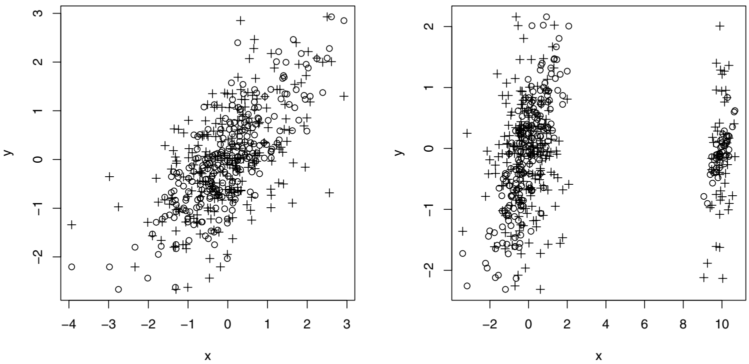


Fig. 1. LEFT: The original bivariate normal distributed data (circles) and the shuffled data (crosses) do have a quite similar behavior. RIGHT: The original data consists of bivariate normal distributed data (non-outliers) plus a shifted outlier group and the shuffled data (crosses) show a quite dissimilar behavior.

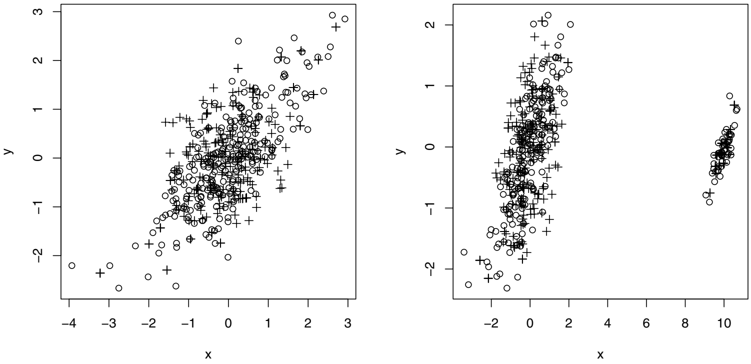


Fig. 2. LEFT: The original bivariate normal distributed data (circles) and the shuffled data (crosses) do have a quite similar behavior. RIGHT: The original data consists of bivariate normal distributed data (non-outliers) plus a shifted outlier group and the shuffled data (crosses) also show a quite similar behavior.

– Choose a robust estimator with a high breakdown point to estimate the correlation between the confidential and the non-confidential variables.
– Define outliers to be the observations with robust Malahanobis distances larger than $\sqrt{\chi^2_{(0.975,\ p)}}$ with p being the number of variables.
– Apply shuffling to non-outlying observations.
– Apply another perturbation method to the outlying observations.

In Figure 1 as well as in Figure 2 it is shown that shuffling (and therefore gadp as well) will be seriously influenced by outliers while the robust shuffling procedure gives reasonable results (Fig. 2; we had applied RMDM microaggregation for the outlier part) also when the data are contaminated.

3 Results Based on Specific Artificial Data Sets

Most of the methods under consideration were already evaluated and compared based on real data in [4], [1] and [8] (see also in the R online help files of package *sdcMicro* [7]). In this section we will investigate artificial data sets featuring different outlier scenarios.

Each method was applied to several bivariate data sets which are visualized at the top of Table 2. The first two data sets follow a bivariate normal distribution with uncorrelated (first data set) and correlated variables (second data set). The only difference between data sets 1 and 2 and data sets 3 and 4 is the inclusion of a single outlier. Furthermore we test all methods on a data set that features an outlier group.

Table 2 provides detailed information about the performance of the methods under consideration with respect to data utility and data protection. Columns *prot.* in Table 2 indicate the performance of the microdata protection methods regarding data protection while columns *qual.* show the performance of the methods with respect to the fact whether the original data structure had been destroyed after applying a procedure.

Possible values of these columns are *pass*, *part.* and *fail* which mean that the method passes, partly passes or fails the criteria. Regarding data utility we simple look at the bivariate original data and compare it to the masked data similar to the evaluation of shuffling in 1 and 2. When evaluating disclosure risk we pay special attention if and how the data are perturbed, especially if the outliers are protected sufficiently.

The classification of the masking procedure given a certain data scenario into *pass*, *part.* and *fail* is rather subjective but more powerful than using one measure of information loss or a traditional measure of disclosure risk.

This fact becomes clear when thinking of the evaluation of disclosure risk regarding to the third graphic in Figure 2. If a method only fails to protect the outlier a usual global measure of disclosure risk (for example measure SDID, see [34]) would be very low. However, the protection of the probably most interesting observation for a data intruder is essential.

Furthermore, when using certain measures the evaluation of the methods depends on the chosen measures and therefore we prefer the explorative approach of visually assessing the performance of the masking methods.

Table 2 shows that almost all methods work well, providing good quality and data utility when the procedures are applied to normally distributed data featuring different covariances. If outliers are included in the data most of the classical methods have problems with respect to data utility while using robust procedures avoid these problems. Using shuffling it is not possible to protect the single outlier in graphic 3 and 4 since the outlier of the generated data (e.g. with method gadp) has the same rank as the outlier of the original data for sure and the swapped value is exactly the same as the original value. Thus it is not possible to perturbe such an outlier using method *shuffle* while the robust version avoids this problem.

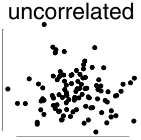
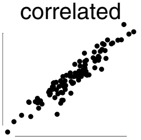
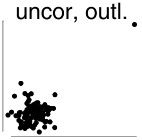
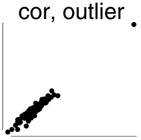
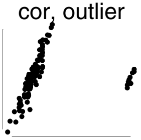
When working with real complex data a minimum requirement of a method is that they should only feature *pass* in Table 2. Hence, only methods *clustppca*, *mdav*, *rmd*, *outdect*, *robShuffle* and *robGapdp* should be applied on real data sets.

4 Simulation

4.1 Design of the Simulation Study

As already indicated, all methods discussed above have been applied to synthetic, bivariate data sets that have been sampled from a multivariate normal distribution with mean vector $\mu = (0, 0)$ and covariance matrix Σ ($\sigma_{ii} = (1, 1), \sigma_{ij} = (0.8, 0.8)$). The data sets that have been used to assess the quality of the microaggregation procedures feature different proportions of outliers (from 0% up to 40%). In the simulation study we considered shifted outliers that have also been generated by a multivariate normal distribution. However, the mean vector of the outlying observations is different ($\mu_{out} = (10, 0)$) to the mean vector of

Table 2. Detailed check if the methods can deal with special data configurations, i.e. which methods protect the underlying data well and which methods preserve the data structure

method	uncorrelated 		correlated 		uncor, outl. 		cor, outlier 		cor, outlier 	
	prot.	qual.	prot.	qual.	prot.	qual.	prot.	qual.	prot.	qual.
additive	pass	pass	pass	pass	part.	part.	part.	part.	pass	pass
correlated	pass	pass	pass	part.	pass	fail	pass	fail	pass	fail
correlated2	pass	pass	pass	pass	part.	part.	part.	part.	pass	pass
restr	pass	pass	pass	pass	pass	pass	pass	pass	pass	fail
ROMM	pass	pass	pass	pass	pass	pass	pass	pass	pass	fail
outdect	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass
swappNum (p=15)	pass	pass	pass	fail	pass	fail	pass	fail	pass	part.
swappNum (p=40)	pass	pass	pass	fail	pass	fail	pass	fail	pass	fail
single	pass	fail	pass	fail	pass	fail	pass	fail	pass	part.
onedims	pass	pass	pass	pass	pass	part.	pass	part.	fail	pass
influence	pass	pass	pass	pass	pass	part.	pass	part.	pass	pass
pca	pass	pass	pass	pass	pass	part.	pass	pass	pass	fail
clustpppca	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass
mdav	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass
rmd	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass
gadp	pass	pass	pass	pass	pass	pass	fail	pass	pass	fail
shuffle	pass	pass	pass	pass	fail	pass	fail	pass	pass	fail
robGadp	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass
robShuffle	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass

the non-outliers. Furthermore, different correlations between the variables have been considered by adjusting the covariance matrix.

We also assess the stability of the different microaggregation methods with respect to measures for data utility and risk. Thus, a total number of 300 data sets was generated for each combination of outlier percentage and correlation between the two variables. All methods were then applied to all data sets featuring a given correlation and outlier percentage. Analyzing the results it is possible to discuss which methods are providing stable outcomes in terms of data quality (protection) and data utility.

4.2 Simulation Results

The following results are based on a simulation study using a total of 1000 simulation runs. The results show again how the existence of outliers influences some microprotection procedures. All the following graphs show the median of the 1000 simulation runs given the outlier fraction.

The first results (displayed in the graphics at the top of Figure 3) are summarizing results of information loss and disclosure risk measures which are based on distances between the perturbed and the original data. The graphics at the

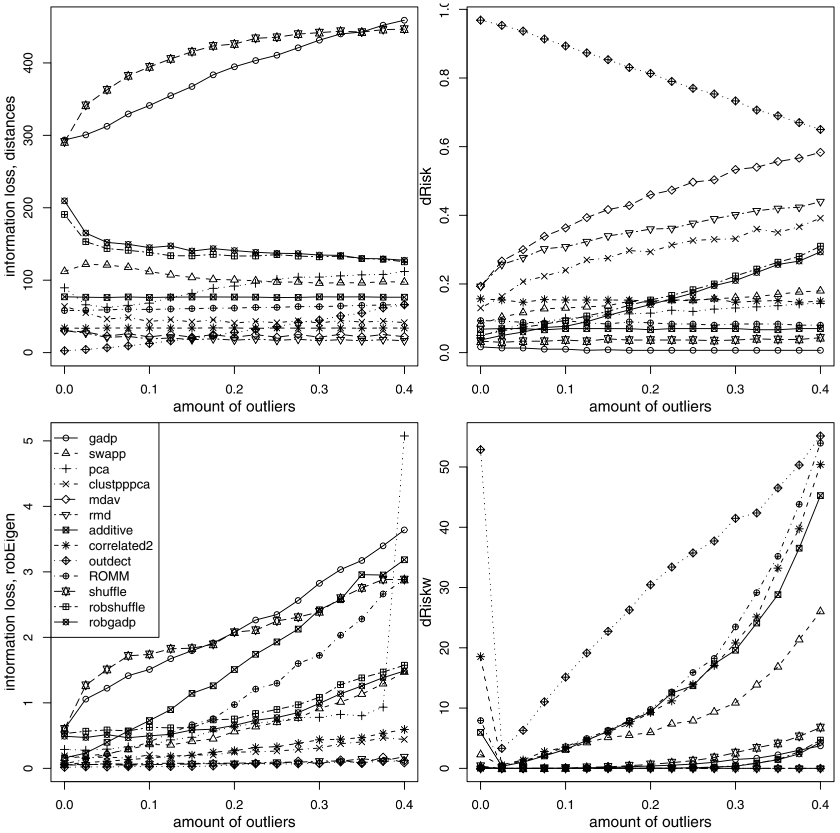


Fig. 3. TOP LEFT: IL1s information loss measure versus the amount of shifted outliers. TOP RIGHT: disclosure risk ([34]) versus the amount of outliers in the generated data sets. BOTTOM LEFT: Information loss based on differences of the eigenvalues of the robust covariances between original and perturbed data versus the amount of outliers. BOTTOM RIGHT: weighted disclosure risk based on robust Mahalanobis distances versus the amount of outliers.

bottom of Figure 3 summarize the simulation results based of an information loss criteria which is based on absolute differences between the eigenvalues of the robust covariance matrices in the original and the perturbed data and disclosure risk criteria which is based on robust Mahalanobis distances and neighbourhood comparisons (see [37]).

The graphic at the top left of Figure 3 shows the influence of outliers to various methods based on the IL1 information loss measure. This measure does not evaluate how well certain statistics are preserved. It only evaluates distances between the original and the preserved data. Therefore, *shuffling* and *gadp* exhibit the highest "information loss" and low "disclosure risk". If the amount of outliers in the data increases, also the information loss criteria shows higher values. This indicates the influence of outliers to these methods. This is not the case for *robust shuffling*. It is also clearly visible that the *RMDM* (denoted as *rmd*) method performs better than *mdav*. Naturally, these two methods have high “risk of disclosure” since only distances are evaluated but the advantages of microaggregation - the aggregation of observations which provides a good protection by itself - is not considered in this risk measure.

The simulation results shown in the bottom of Figure 3 are based on more realistic measures of information loss and disclosure risk. These measures take the multivariate behavior and the risk with respect to robust Mahalanobis distances into account. Microaggregation methods *rmd* and *mdav* perform again very well and also *robust shuffling* gives quite good results. *Shuffling*, *ROMM*, *gadp* and *adding noise* are highly influenced by outliers and perform poorly. The “shift” between 0 percent outliers and 2.5% outliers in the graphic at the bottom right of Figure 3 occurs for many methods. This “shift” indicates that even a few outliers have influence on these methods which results in poor perturbation quality if the data contains outliers. It is self-evident that poor results regarding information loss often indicate low disclosure risk, i.e. if the (multivariate) structure of the data is completely destroyed not much can be uncovered by an intruder.

5 Conclusion

We have conducted a large simulation study considering various outlier scenarios and different correlations between the variables. Only a few results in a very comprehensive form could be presented in this paper in order to stay within the limit of pages.

Outliers are virtually present in every data set from official statistics and therefore perturbation methods for numerical variables must be able to deal with inhomogeneous data sets. Furthermore, outlying observations surely possess a higher risk for re-identification and it is essential that the methods protect these outliers properly. On the other hand, a protection method should not destroy the multivariate data structure. We showed that many methods are heavily influenced by outliers which results in poor quality regarding data utility and protection.

We showed that some classical methods are not able to deal with special data configurations (see Table 2). These methods may not be suitable for applications to real world data sets. Nevertheless, some of the most popular methods which fail under such data configurations have been included in the simulation study together with robust modifications of these methods. However, some methods (mostly methods for synthetic data generation) were excluded from the simulation study because it is clear that these methods can not deal with data that feature outliers.

The results of the simulation study showed some procedures performed poorly when applied to data that are contaminated with outliers. Analyzing the results it turned out that methods *rmd*, *clustpppca*, *mdav* and *robust shuffling* performed very well. In many situations *rmd* outperformed all other methods (see Fig. 3).

With model based procedures one may run into serious problems when masking real complex data including outliers. The robustification of such methods, like our proposed robust shuffling procedure, makes it possible to deal with data contamination in an efficient way.

References

1. Templ, M.: sdcMicro: A package for statistical disclosure control in R. In: Bulletin of the International Statistical Institute, 56th Session (2007)
2. Meindl, B., Templ, M.: The anonymisation of the CVTS2 and income tax dataset. an approach using R-package sdcMicro. In: Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality, Monographs of Official Statistics (to appear, 2007)
3. Karr, A., Oganian, A., Reiter, J., Woo, M.J.: New measures of data utility. Technical report (2006)
4. Templ, M.: Software development for SDC in R. In: Domingo-Ferrer, J., Franconi, L. (eds.) PSD 2006. LNCS, vol. 4302, pp. 347–359. Springer, Heidelberg (2006)
5. Muralidhar, K., Sarathy, R., Dankekar, R.: Why swap when you can shuffle? a comparison of the proximity swap and data shuffle for numeric data. In: Privacy in Statistical Databases. LNCS, pp. 164–176. Springer, Heidelberg (2006)
6. Muralidhar, K., Sarathy, R.: Data shuffling- a new masking approach for numerical data. Management Science 52(2), 658–670 (2006)
7. Templ, T.: sdcMicro: Statistical Disclosure Control methods for the generation of public- and scientific-use files, R package version 2.4.7 (2008)
8. Templ, M.: sdcMicro: A new flexible R-package for the generation of anonymised microdata - design issues and new methods. In: Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality, Monographs of Official Statistics (to appear, 2007)
9. R Development Core Team: R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria (2008) ISBN 3-900051-07-0
10. Brand, R., Giessing, S.: Report on preparation of the data set and improvements on sullivans algorithm. Technical report (2002)
11. Kim, J.: A method for limiting disclosure in microdata based on random noise and transformation. In: Proceedings of the Section on Survey Research Methods. American Statistical Association, pp. 303–308 (1986)
12. Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Lenz, R., Longhurst, J., Schulte-Nordholt, E., Seri, G., De Wolf, P.P.: Handbook on statistical disclosure control version 1.01 (2007)
13. Brand, R.: Microdata protection through noise addition. In: PSD 2004. LNCS, pp. 347–359. Springer, Heidelberg (2004)
14. Ting, D., Fienberg, S., Trottini, M.: ROMM methodology for microdata release. In: Monographs of official statistics, Work session on statistical data confidentiality, Eurostat, Luxembourg (2005)
15. Dalenius, T., Reiss, S.: Data-swapping: A technique for disclosure control. In: Proceedings of the Section on Survey Research Methods, vol. 6, pp. 73–85. American Statistical Association (1982)
16. Defays, D., Nanopoulos, P.: Panels of enterprises and confidentiality: the small aggregates method. In: Proceedings of the 1992 Symposium on Design and Analysis of Longitudinal Surveys, Statistics Canada, Ottawa, pp. 195–204 (1993)

17. Muralidhar, K., Parsa, R., Sarathy, R.: A general additive data perurbation method for database security. Management Science 45, 1399–1415 (1999)
18. Huber, P.: Robust Statistics. Wiley and Sons, New York (1981)
19. Moore, R.: Controlled data-swapping techniques for masking public use microdata sets. Technical report (1996)
20. Maronna, R.: Robust M-estimators of multivariate location and scatter. The Annals of Statistics 4(1), 51–67 (1976)
21. Rousseeuw, P.: Multivariate estimation with high breakdown point. In: Mathematical Statistics and Applications, Akademiai Kiado, Budapest, pp. 283–297 (1985)
22. Maronna, R., Zamar, R.: Robust multivariate estimates for highdimensional datasets. Technometrics 44, 307–317 (2002)
23. Domingo-Ferrer, J., Mateo-Sanz, J.: Practical data-oriented microaggregation for statistical disclosure control. IEEE Trans. on Knowledge and Data Engineering 14(1), 189–201 (2002)
24. Mateo-Sanz, J., Martínez-Ballesté, A., Domingo-Ferrer, J.: Fast generation of accurate synthetic microdata. In: Domingo-Ferrer, J., Torra, V. (eds.) PSD 2004. LNCS, vol. 3050, pp. 298–306. Springer, Heidelberg (2004)
25. Burridge, J.: Information preserving statistical obfuscation. Statistics and Computing 13, 321–327 (2003)
26. Torra, V., Abowd, J., Domingo-Ferrer, J.: Using mahalanobis distance-based record linkage for disclosure risk assessment. In: Domingo-Ferrer, J., Franconi, L. (eds.) PSD 2006. LNCS, vol. 4302, pp. 233–242. Springer, Heidelberg (2006)
27. Griffin, R., Navarro, A., Flores-Baez, L.: Disclosure avoidance for the 1990 census. In: Proceedings of the Section on Survey Research Methods, pp. 516–521. American Statistical Association (1989)
28. Rubin, D.: Discussion of statistical disclosure limitation. Journal of Official Statistics 9(2), 461–468 (1993)
29. Iman, R., Conover, W.: A distribution-free approach to inducing rank correlation among input variables. Communications in Statistics B11, 311–334 (1982)
30. Stein, M.: Large sample properties of simulations using latin hypercube sampling. Technometrics 29, 143–151 (1987)
31. Wyss, G., Jorgensen, K.: Sandia's latin hypercube sampling software. Technical report sand98-0210, Sandia National Laboratories, Albuquerque, NM (1998)
32. Minasny, B.: Sampling methods for uncertainty analysis, Matlab Toolbox for Latin Hypercube Sampling (2003)
33. Yancey, W., Winkler, W., Creecy, R.: Disclosure risk assessment in perturbative microdata protection. In: Inference Control in Statistical Databases. LNCS, pp. 49–60. Springer, Heidelberg (2002)
34. Mateo-Sanz, J.M., Sebe, F., Domingo-Ferrer, J.: Outlier protection in continuous microdata masking. In: Domingo-Ferrer, J., Torra, V. (eds.) PSD 2004. LNCS, vol. 3050, pp. 201–215. Springer, Heidelberg (2004)
35. Mateo-Sanz, J., Domingo-Ferrer, J., Sebé, F.: Probabilistic information loss measures in confidentiality protection of continuous microdata. Data Mining and Knowledge Discovery 11, 181–193 (2005)
36. Domingo-Ferrer, J., Mateo-Sanz, J., Torra, V.: Comparing sdc methods for microdata on the basis of information loss and disclosure risk. In: Pre-Proccedings of ETK-NTTS, vol. 2, pp. 807–826. Springer, Heidelberg (2001)
37. Templ, M., Meindl, B.: Robust statistics meets SDC: New disclosure risk measures for continuous microdata masking. In: Domingo-Ferrer, J., Saygın, Y. (eds.) PSD 2008. LNCS, vol. 5262, pp. 177–189. Springer, Heidelberg (2008)

A Preliminary Investigation of the Impact of Gaussian Versus t-Copula for Data Perturbation

Mario Trottini, Krish Muralidhar, and Rathindra Sarathy

University of Alicante, Apartado de Correos 99, Alicante, Spain
University of Kentucky, Lexington, KY 40506
Oklahoma State University, Stillwater, OK 74078
mario.trottini@ua.es, krishm@uky.edu, rathin.sarathy@okstate.edu

Abstract. In this paper, we provide a preliminary investigation of t-copulas for perturbing numerical confidential variables. A perturbation approach using Gaussian copulas has been proposed earlier. However, one of the problems with the Gaussian copulas is that it does not preserve tail dependence. In this investigation, we show that the t-copula can be used effectively to provide all the benefits that a Gaussian copula provides and, in addition, maintain tail dependence as well. We illustrate this approach using two examples. We hope to perform a comprehensive investigation of this approach in the future.

Keywords: Data Masking, Data Perturbation; Copula; Tail Dependence.

1 Introduction

Suppose that a certain Information Organization (IO) has collected information on M confidential variables $\mathbf{X} = (X_1, \ldots, X_M)$, and L nonconfidential variables, $\mathbf{S} = (S_1, \ldots, S_L)$ for a set of N units. The resulting micordata, D_0, can be represented as an $N \times (M + L)$ partitioned matrix

$$\mathbf{D_0} = (\tilde{\mathbf{X}} \vdots \tilde{\mathbf{S}}) \tag{1}$$

where $\tilde{\mathbf{X}}$ is the $N \times M$ matrix containing the collected information on the confidential variables, and $\tilde{\mathbf{S}}$ is the $N \times L$ matrix containing the collected information on the nonconfidential variables. The i-th row of $\tilde{\mathbf{X}}$ and $\tilde{\mathbf{S}}$, that we will denote by $x_i = (x_{i1}, \ldots, x_{iM})$ and $s_i = (s_{i1}, \ldots, s_{iL})$, represent the M-dimensional and the L-dimensional vectors of the confidential and non confidential values for the i-th unit in the dataset.

Because of confidentiality concerns the microdata, $\mathbf{D_0}$, cannot be released to the users. Some type of masking must be applied to $\mathbf{D_0}$ before release. One possibility is to replace the confidential values $\tilde{\mathbf{X}}$ with perturbed values $\tilde{\mathbf{Y}}$ and release $\mathbf{D_M}$,

$$\mathbf{D_M} = (\tilde{\mathbf{Y}} \vdots \tilde{\mathbf{S}}).$$

J. Domingo-Ferrer and Y. Saygın (Eds.): PSD 2008, LNCS 5262, pp. 127–138, 2008.

Ideally we would like to find a perturbation method to generate $\tilde{\mathbf{Y}}$ such that disclosure risk is minimized and data utility is maximized. Within the conditional distribution approach (CDA) of [6] the optimal procedure generates $\tilde{\mathbf{Y}}$ from the conditional distribution of $\mathbf{X}$ given $\mathbf{S}$. Such an optimal procedure, however, requires complete knowledge of the joint distribution of $(\mathbf{X}, \mathbf{S})$ and the ability of simulating from the conditional $\mathbf{X}|\mathbf{S} = \mathbf{s}$. An approximation to this optimal procedure, known as C-GADP, is presented in [10]. It is based on a Gaussian copula, and it only requires knowledge of the marginal distributions of the original data. Rather then generating $\tilde{\mathbf{Y}}$ from the conditional distribution of $\mathbf{X}$ given $\mathbf{S}$, C-GADP generates $\tilde{\mathbf{Y}}$ from the conditional distribution of $\mathbf{X}\prime$ given $\mathbf{S}\prime$, where $(\mathbf{X}\prime, \mathbf{S}\prime)$ is a random vector whose distribution is expressable in terms of a Gaussian copula with correlation matrix, ρ, estimated from the Spearman rho correlation matrix $\rho_S^{original}$ of the original data, using the relationship $\rho_{ij} = 6/\pi \cdot \arcsin[\rho_S^{original}(i,j)/2]$, and marginals estimated from the original data as well. In [10] it is shown that when the sample size of the original data is sufficiently large, the C-GADP perturbation procedure preserves approximately the margins and the rank-order correlation (and thus monotonic relationships) of the original data while minimizing disclosure risk (at least within the CDA framework of [6]). As noted by the authors, however, the C-GADP approach cannot capture and correctly reproduce the phenomenon of dependence in extreme values (or *tail dependence*). In this paper we address this issue implementing a copula based perturbation procedure, that we called *t-copula perturbation* (TCP), which, as the name indicates, is based on the t-copula. We show that TCP can be used effectively to provide all the benefits that the C-GADP method would provide and, in addition, maintain some important type of tail dependence.

The idea of masking data by estimating its distribution from a given data set and generate fake data from the estimated distribution is not exclusive of the CDA approach that underlies the derivation of the TCP procedure proposed here. The multiple imputation approach suggested by [9], its variant adopted by [5], and their generalizations (see, for example, [8]), or the IPSO approach of [2], to mention just a few, are also based on the same idea. The relation of the CDA approach that we use here to these and other masking procedures is discussed in details in [6] (section 4). We do not investigate this relation any further in the paper. The interested reader can refer to [6] and the references therein.

In Section 2 we briefly introduce the notion of copula, and describe with some details relevant features of the Gaussian and t-copula. Section 3 formalizes the notion of tail dependence and illustrates the poor perfomance of C-GADP when tail dependence is present with an example. In Section 4 we present the new TCP perturbation procedure that extends C-GADP and that can adjust the perturbation depending on the degrees of tail dependence in the original data. Section 5 illustrates the performance of the proposed TCP procedure with two examples with and without tail dependence. Section 6 summarizes the main results of the paper and outlines ideas of future work.

2 Gaussian and t-Copulas

A p dimensional copula, C is a p-dimensional distribution function on the unit cube $[0,1]^p$ with uniformly distributed marginals. Sklar's Theorem (Sklar 1959) states that every joint distribution function F of a p-dimensional random vector, with marginals CDF $F_1, \ldots, F_p$ can be written as,

$$F(x_1, \ldots, x_p) = C(F_1(x_1), \ldots, F_p(x_p) \tag{2}$$

for some copula C, which is uniquely determined on $[0,1]^p$ for distributions F with absolutely continuous margins. Conversely any copula C may be used to "join" any collection of univariate CDF $F_1, \ldots, F_p$ to create a multivariate CDF, F, with margins $F_1, \ldots, F_p$.

Copulas have been playing a prominent role in multivariate analysis due to the fact that several types of dependence of interest in applications are copula based, that is they only depend on the copula function C and are independent of the margins. These types of dependence include, among others, the well known Spearman's rho and Kendall's tau measures of concordance and the coefficients of lower and upper tail dependence that we describe with some details in the next section. As such copulas provide a natural way to study and measure dependence between random variables. The primary applications of copulas have been to combine specified (arbitrary) marginal distributions into joint distributions that exhibit certain specified dependence behaviour. [7] serves as a good introduction. A variety of copula functions have been proposed and investigated in the research literature. The selection of a copula function depends on the specific problem under consideration. Here we restrict our attention to the Gaussian and the t-copula that we define next. A k-dimensional Gaussian copula parameterized with product moment correlation matrix ρ can be written as:

$$C_\rho^{Ga}(\mathbf{u}) = \Phi_\rho(\Phi^{-1}(u_1), \ldots, \Phi^{-1}(u_k)),$$

where Φ_ρ is the joint CDF of a k-dimensional multivariate standard normal distribution with correlation matrix ρ and Φ^{-1} is the quantile function of the univariate standard normal distribution. Similarly, a k-dimensional t-copula parameterized with product moment correlation ρ, and degrees of freedom ν can be written as,

$$C_{\rho,\nu}^t(\mathbf{u}) = t_{\nu,\rho}(t_\nu^{-1}(u_1), \ldots, t_\nu^{-1}(u_n)), \tag{3}$$

where $t_{\nu,\rho}$ represents the joint CDF of a k-variate Student t distribution with location $\mathbf{0}$, scale and correlation matrix ρ, ν degrees of freedom and t_ν^{-1} is the quantile function of a univariate t distribution with ν degrees of freedom. Note that the Gaussian copula can be obtained as limiting of the t-copula as $\nu \to \infty$. An important property of the Gaussian and t-copula (and more in general of elliptical copulas) that we will use later in the paper is the following:

Preposition. Let $\mathbf{X}$ be a k-dimensional random vector whose distribution can be expressed either in terms of a Gaussian copula, C_ρ^{Ga} or in terms of a t-copula $C_{\rho,\nu}^t$. Let also $\rho_\tau = \{\rho_\tau(ij)\}_{i,j=1,\ldots,k}$ be the $k \times k$ Kendall Tau correlation matrix of $\mathbf{X}$. Then we have:

$$\rho_\tau(i,j) = \frac{2}{\pi} \cdot \arcsin(\rho_{ij}) \tag{4}$$

where ρ_{ij} is the (i,j) element of ρ.

The Gaussian (t-) copula can be thought of as representing the dependence structure implicit in a multivariate Gaussian (t) distribution. Gaussian copulas have been extensively used in modeling non normal data due to their flexibility and analytical tractability (see, for example, [3]). As we commented in the introduction the C-GADP method in [10], which approximately preserves marginals and monotonic relationships in the original data, is based on a Gaussian copula with correlation matrix ρ estimated from the Spearman rho correlation matrix $\rho_S^{original}$ of the original data, using the relationship $\rho_{ij} = 6/\pi \cdot \arcsin[\rho_S^{original}(i,j)/2]$, and marginals estimated from the original data as well. One important limitation of the Gaussian copula, however, is that it is not able to capture the phenomenon of dependence in extreme values. The t-copula, on the other hand, does have such ability and this, in part, explains its success in applications involving, for example, modeling of financial data such as return data for which dependence in extreme values is often observed.

In the next section, we introduce the notion of tail dependence and illustrate the ability of the t-copula to maintain this property better than the Gaussian copula. Consequently, we also show that a perturbation approach based on the t-copula preserves tail dependence (if any) present in the original data better than a perturbation approach based on the Gaussian copula.

3 Tail Dependence

The notion of bivariate tail dependence relates to the amount of dependence in the upper-quadrant tail or in the lower-quadrant tail of a bivariate distribution. It is a concept relevant to dependence in extreme values. The following definition formalizes the notion of *lower* and *upper tail dependence* ([7], page 214).

Definition: Let X and Y be continuous random variables with distributions functions F and G respectively. The *upper tail dependence parameter* λ_U is the limit (if it exists) of the conditional probability that Y is greater than the 100α percentile of G given that X is greater than the 100α percentile of F as α approaches 1, i.e.,

$$\lambda_U = lim_{\alpha\to 1^-} P(Y > G^{-1}(\alpha) | X > F^{-1}(\alpha)).$$

Similarly the *lower tail dependence parameter* λ_L is defined as

$$\lambda_L = lim_{\alpha\to 0^+} P(Y \leq G^{-1}(\alpha) | X \leq F^{-1}(\alpha)).$$

It can be shown that these measures are copula based, i.e. they only depend on the copula C of (X_1, X_2) regardless of the margins. When these coefficients are strictly greater than zero the copula tends to generate joint extreme events. If $\lambda_L > 0$, for example, we talk of tail dependence in the lower tail; if $\lambda_L = 0$ we talk of independence in the lower tails. For the copula of an elliptically symmetric distribution, like the Gaussian or the t copula, the two measures λ_U and λ_L coincide, and are denoted simply by λ. For the Gaussian copula the value of λ is zero, for the t- copula is given by ([4], page 114):

$$\lambda = 2 \cdot t_{\nu+1}(-\sqrt{\nu+1} \cdot \sqrt{1-\rho}/\sqrt{1-\rho}). \tag{5}$$

While tail dependence, as presented here is an asymptotic concept, the next example illustrates its practical implications and the poor performance of C-GADP when tail dependence is present.

3.1 Example 1

To illustrate the behaviour of the C-GADP approach in the presence of tail dependence we simulated a microdata M_1 consisting of 10000 i.i.d. draws from a three dimensional random vector (X_1, X_2, S_1) with distribution function expressable in terms of a t-copula $C^t_{\rho,\nu}$ with $\nu = 2$ degrees of freedom, correlation matrix $\rho = \{\rho_{ij}\}_{i,j=1,2,3}$, $\rho_{12} = 0.61$, $\rho_{13} = 0.65$, $\rho_{23} = 0.49$ and normal marginals

$$X_1 \backsim N(0,2); \quad X_2 \backsim N(3,1); \quad S_1 \backsim N(4,3); \tag{6}$$

where $N(\mu, \sigma)$ denotes a normal distribution with mean μ and standard deviation σ. By construction the microdata M_1 does have tail dependence. In particular, using (5), we have that the coefficients of tail dependence for M_1 are:

$$\lambda_{X_1,X_2} = 0.46; \quad \lambda_{X_1,S_1} = 0.48; \quad \lambda_{X_2,S_1} = 0.39.$$

Bivariate distributions of the original microdata M_1 are shown in Fig. 1. To facilitate the visualization of the tail dependence in the microdata M_1, the vertical and horizontal lines in Fig. 1 mark the 0.005 and the 0.995 quantiles of the marginal distributions. We applied C-GADP to the original microdata M_1 using X_1 and X_2 as confidential variables and S_1 as nonconfidential variable. Spearman's rho, ρ_S, and Kendall's tau, ρ_τ, for original and C-GADP masked data are reported in table 1.

Table 1. ρ_S and ρ_τ for original and C-GADP masked data: Example 1

$\rho_\tau^{Original}$			ρ_τ^{C-GADP}			$\rho_S^{Original}$			ρ_S^{C-GADP}		
1.00	0.42	0.44	1.00	0.40	0.42	1.00	0.56	0.59	1.00	0.57	0.60
0.42	1.00	0.33	0.40	1.00	0.32	0.56	1.00	0.45	0.57	1.00	0.46
0.44	0.33	1.00	0.42	0.32	1.00	0.59	0.45	1.00	0.60	0.46	1.00

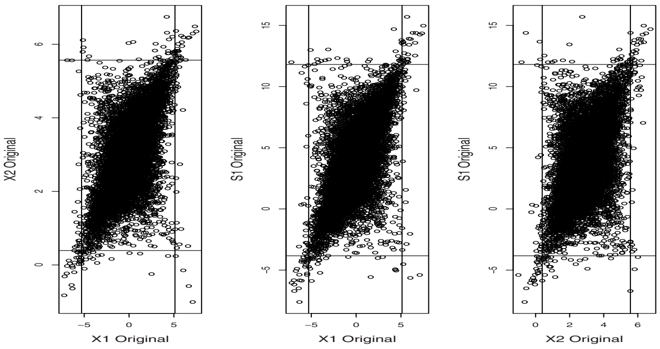


Fig. 1. Bivariate distributions of the original data: Example 1

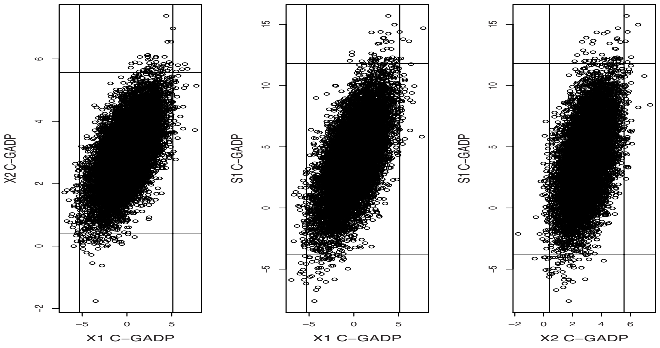


Fig. 2. Bivariate distributions, C-GADP masked data: Example 1

Fig. 2 shows the bivariate distributions of the confidential and nonconfidential variables in the C-GADP masked data. As expected C-GADP preserves well the two measures of concordance, ρ_S and ρ_τ as well as the marginal distributions (not shown). In terms of preserving bivariate distributions, however, C-GADP performs quite poorly. The practical implications of tail dependence, in this example, can be seen by comparing joint quantile exceedance probabilities. Suppose, for example, that the two sensitive variables X_1 and X_2 represent daily returns of two stocks with correlation 0.58. Based on the original data the probability p_1 that on any day the two returns would drop below the 0.5% quantile of their marginal distribution (evaluated as the fraction of points that follows in the lower quadrant of the bottom left plot in Fig. 1) would be 0.0027. The estimation of p_1 using C-GADP is 0.0003. This means that using the Gaussian copula a data user would estimate that in the long run such an event would happen once every 3333 days on average, i.e roughly once every 13 years (assuming 260 days in the stock market year). In the true model, however, the event occurs with a probability that is 9 times higher or roughly once every 1.42 years.

4 The t-Copula Perturbation

Using the notation in the introduction, let F_i, $i = 1, \ldots M$ be the CDF of the i^{th} component of the confidential random vector $\mathbf{X}$, and let G_j, $j = 1, \ldots L$ be the CDF of the j^{th} component of the nonconfidential random vector $\mathbf{S}$. It is assumed that both the F_i's and the G_j's are strictly increasing. The t-copula for the perturbation case can be written as:

$$C^t_{\rho,\nu}(\mathbf{u}) = t_{\nu,\rho}(\mathbf{X}^*, \mathbf{S}^*)$$

where $\mathbf{X}^*$ and $\mathbf{S}^*$ are defined as follows:

$$X^*_i = t^{-1}_\nu(F_i(X_i)), \;\; i = 1, \ldots, M \tag{7}$$

$$S^*_j = t^{-1}_\nu(G_j(S_j)), \;\; j = 1, \ldots, L. \tag{8}$$

Under the assumption that the joint distribution of the original variables can be expressed in terms of the copula in (3), the joint distribution of $(\mathbf{X}^*, \mathbf{S}^*)$ is multivariate Student-t with location $\tilde{\mu}$, scale matrix $\tilde{\Sigma}$ and ν degrees of freedom,

$$(\mathbf{X}^*, \mathbf{S}^*) \backsim St(\tilde{\mu}, \tilde{\Sigma}, \nu) \tag{9}$$

where

$$\tilde{\mu} = \mathbf{0}; \quad \tilde{\Sigma} = \rho.$$

The parameters ρ and ν need to be estimated from the data. It is useful to represent the mean vector $\tilde{\mu}$ and the scale matrix $\tilde{\Sigma}$ using the partition:

$$\tilde{\mu} = (\tilde{\mu}_X, \tilde{\mu}_S); \;\; \tilde{\Sigma} = \begin{pmatrix} \tilde{\Sigma}_{XX} & \tilde{\Sigma}_{XS} \\ \tilde{\Sigma}_{SX} & \tilde{\Sigma}_{SS} \end{pmatrix}.$$

From (9) and basic properties of the multivariate Student-t distribution (see, for example, [1], proposition 4), the conditional distribution of $\mathbf{X}^*$ given $\mathbf{S}^* = \mathbf{s}$ is still a multivariate Student t of dimension M, with location $\tilde{\mu}_{X.S}$, scale matrix $\tilde{\Sigma}_{XX.S}$ and degrees of freedom $\nu_{X.S}$,

$$X^* | S^* = s \backsim St(\tilde{\mu}_{X.S}, \tilde{\Sigma}_{XX.S}, \nu_{X.S}) \tag{10}$$

where:

$$\tilde{\mu}_{X.S} = \tilde{\mu}_X + \tilde{\Sigma}_{XS}\tilde{\Sigma}^{-1}_{SS}(s - \tilde{\mu}_S), \quad \nu_{X.S} = \nu + L,$$

$$\tilde{\Sigma}_{XX.S} = [\frac{\nu + (s - \tilde{\mu}_S)^T \tilde{\Sigma}^{-1}_{SS}(s - \tilde{\mu}_S)}{\nu + L}] \cdot [\tilde{\Sigma}_{XX} - \tilde{\Sigma}_{XS}\tilde{\Sigma}^{-1}_{SS}\tilde{\Sigma}_{SX}].$$

Under the assumption that the marginal CDF of the original variables are strictly increasing, the rank order correlation matrix of the original data is the same as the correlation matrix of the transformed data $\mathbf{D_0}^*$,

$$\mathbf{D_0}^* = \begin{pmatrix} x^*_1 & s^*_1 \\ \ldots & \ldots \\ x^*_N & s^*_N \end{pmatrix} \tag{11}$$

where for any $i \in \{1, \ldots, N\}$,

$$x_i^* = (x_{i1}^*, \ldots, x_{iM}^*), \quad \text{with } x_{ih}^* = t_\nu^{-1}(F_h(x_{ih}))\ h = 1, \ldots, M$$

$$s_i^* = (s_{i1}^*, \ldots, s_{iL}^*), \quad \text{with } s_{ij}^* = t_\nu^{-1}(G_j(s_{ij}))\ j = 1, \ldots, L. \tag{12}$$

One can evaluate the Kendall's tau correlation matrix of the original data $\rho_\tau^{Original}$ and estimate the correlation matrix ρ in (9) by $\hat{\rho}$ using (4). Once ρ have been estimated the degrees of freedom can be estimated fitting a t-copula with correlation matrix $\hat{\rho}$ to the original data.

The algorithm to implement the t-copula perturbation would be then as follows:

T-copula Perturbation Procedure (TCP)

Step 1. Identify the marginal distributions of attributes $X_1, \ldots, X_M, S_1, \ldots, S_L$. Denote by $\hat{F}_{X_i}$ and $\hat{G}_{S_j}$, $i = 1, \ldots, M$, $j = 1, \ldots, L$ the estimated marginal CDF.

Step 2. Compute the Kendall's tau matrix of the original data set, $\rho_\tau^{Original}$.

Step 3a. Compute product moment correlation $\hat{\rho}$ using $\rho_\tau^{Original}$ and (4).

Step 3b. Fit a t-copula with correlation matrix $\hat{\rho}$ to the original data in order to obtain an estimate $\hat{\nu}$ of the t-copula's degrees of freedom ν.

Step 4. Compute the matrix S^* of transformed nonconfidential variables,

$$\mathbf{S}^* = (s_1^* \vdots s_2^* \vdots \ldots \vdots s_N^*)^T \tag{13}$$

with s_i^* as in (12), $i = 1, \ldots, N$ with F_i's and G_j's as in step 1.

Step 5. For each $i \in \{1, \ldots, N\}$ generate y_i^* from the conditional distribution of $X^*|S^* = s_i$ in (10)

Step 6. For each $i \in \{1, \ldots, N\}$ compute y_i from y_i^* as follows:

$$y_i = F_i^{-1}(t_\nu(y_i^*))$$

Step 7. In the original data in (1) replace $\tilde{\mathbf{X}}$ by Y,

$$\mathbf{Y} = (y_1 \vdots y_2 \vdots \ldots \vdots y_N)^T \tag{14}$$

and release to the users $(\mathbf{Y}, \mathbf{S})$ plus the information about: (i) the t-copula model used for the perturbation; (ii) the estimated parameters of the t-copula model $(\hat{\rho}, \hat{\nu})$.

It can be shown that the proposed procedure is equivalent to replacing the original confidential variables with i.i.d. draws from the conditional distribution of a $\mathbf{X}''|\mathbf{S}''$ of a random vector $(\mathbf{X}'', \mathbf{S}'')$ with joint distribution expressable in terms of a t-copula with marginals $\hat{F}_{X_i}$'s and $\hat{G}_{S_j}$'s, product moment correlation matrix $\hat{\rho}$ and degrees of freedom $\hat{\nu}$ as in steps 1, 3a and 3b of the TCP procedure.

For large sample size C-GADP, preserves approximately the estimated marginals ($\hat{F}_{X_i}$'s and $\hat{G}_{S_j}$'s) and the Kendall-tau correlation matrix ($\rho_\tau^{Original}$) (and thus monotonic relationships) of the original data. In addition to C-GADP, however, when the original data present dependence in extreme values that can be described by a multivariate t-copula, the TCP procedure preserves, approximately, tail dependence as well. However, as we describe in section 6, the strong symmetry of the copula might reduce the range of applications and generalizations of t-copula that introduce more asymmetry might be preferred. We illustrate the application of the t-copula in the next section.

5 Examples of Gaussian and t-Copula Based Perturbation

The following two examples illustrate the performance of the t-copula and compare it with the the C-GADP approach for data set with and without tail dependence.

5.1 Example 1: Continued

We applied the t-copula perturbation to the microdata M_1 of example 1 again using X_1 and X_2 as confidential variables and S_1 as nonconfidential variable. As observed in section 3.1 this is a microdata with tail dependence.

The estimated value of ν for this case was $\hat{\nu} = 2$. Spearman's rho, ρ_S, and Kendall's tau, ρ_τ, for original and t-copula perturbed data are reported in table 2.

Fig. 3 shows bivariate distributions of the confidential and nonconfidential variables in the t-copula masked data. As expected the t-copula pertrubation performs as well as C-GADP in preserving the two measures of concordance, ρ_S and ρ_τ and the marginal distributions (not shown here) and outperforms C-GADP in terms of preserving bivariate distributions and tail dependence.

As an illustration of practical implications of tail dependence, in example 1 we supposed that the two sensitive variables X_1 and X_2 were daily returns of two stocks. We observed that based on the original data the probability p_1 that on any day the two returns would drop below the 0.5% quantile of their marginal distribution would be 0.0027. The estimation of p_1 using C-GADP instead was 0.0003 quite far from the "true" value 0.0027. We also observed that, as a result of this discrepancy, C-GADP would provide a very poor estimate of the long run proportions of days in which the event is observed (under C-GADP the event would occur once every 13 years while under the original data the event would occur roughly once every 1.42 years). The t-perturbation, on the other hand would produce $p_1 = 0.0022$ quite close to the "true" 0.0027 leading to approximately the same conclusions as the original data in terms of long run

Table 2. ρ_S and ρ_τ for original and TCP masked data: Example 1

$\rho_\tau^{Original}$			$\rho_\tau^{t-copula}$			$\rho_S^{Original}$			$\rho_S^{t-copula}$		
1.00	0.42	0.44	1.00	0.42	0.44	1.00	0.56	0.59	1.00	0.57	0.59
0.42	1.00	0.33	0.42	1.00	0.33	0.56	1.00	0.45	0.57	1.00	0.45
0.44	0.33	1.00	0.44	0.33	1.00	0.59	0.45	1.00	0.59	0.45	1.00

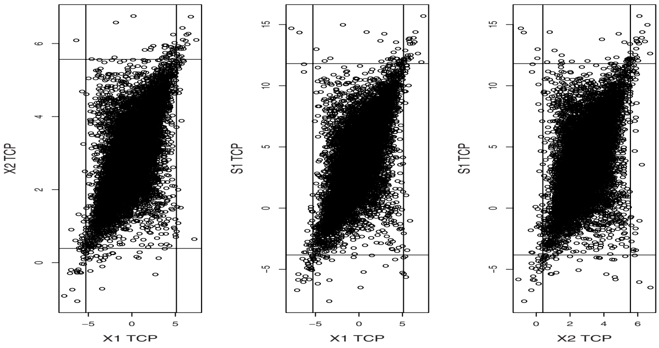


Fig. 3. Bivariate distributions for TCP masked data: Example 1

proportions of days in which the event is observed. Using the t-copula masked data, in fact, the event would occur once every 455 days (1.75 years).

5.2 Example 2: t-Perturbation without Tail Dependence

In order to illustrate and compare the performance of the Gaussian and t-copula perturbation approaches for masked data with no tail dependence, we generated an artificial microdata M_2 consisting of 10000 i.i.d. draws from a three dimensional random vector (X_1, X_2, S_1) whose joint distribution can be expressed in terms of a Gaussian copula with product moment correlation matrix ρ and normal marginals as in example 1. By construction the microdata M_2 has no tail dependence. We applied to M_2 the C-GADP and the t-copula perturbation procedures using X_1 and X_2 as confidential variables and S_1 as nonconfidential variable. For the t-copula perturbation approach the MLE estimator of ν at step 4 of the t-perturbation procedure was found using a grid $1, 2, \ldots, 1000$ for ν. The estimation of ν, for this case

Table 3. ρ_τ for original, C-GADP and TCP masked data: Example 2

ρ_τ Original			ρ_τ C-GADP			ρ_τ T-GADP		
1.00	0.42	0.45	1.00	0.43	0.45	1.00	0.41	0.44
0.42	1.00	0.33	0.43	1.00	0.33	0.41	1.00	0.32
0.45	0.33	1.00	0.45	0.33	1.00	0.44	0.32	1.00

Table 4. ρ_S for original, C-GADP and TCP masked data: Example 2

ρ_S Original			ρ_S C-GADP			ρ_S T-GADP		
1.00	0.60	0.63	1.00	0.61	0.63	1.00	0.59	0.62
0.60	1.00	0.48	0.61	1.00	0.48	0.59	1.00	0.46
0.63	0.48	1.00	0.63	0.48	1.00	0.62	0.46	1.00

was $\hat{\nu} = 1000$ (the maximum value in the grid). Spearman's rho and Kendall's tau for original, C-GADP and TCP perturbed data are reported in tables 3 and 4 respectively. Figure 4 compares bivariate distributions of the confidential variables in the M_2 microdata and in the corresponding C-GADP and t-copula masked versions. As expected the t-copula perturbation approach, in this case, is approximately equivalent to the C-GADP approach and both methods perform very well in preserving marginal (not shown here), bivariate distributions and the concordance measures ρ_S and ρ_τ.

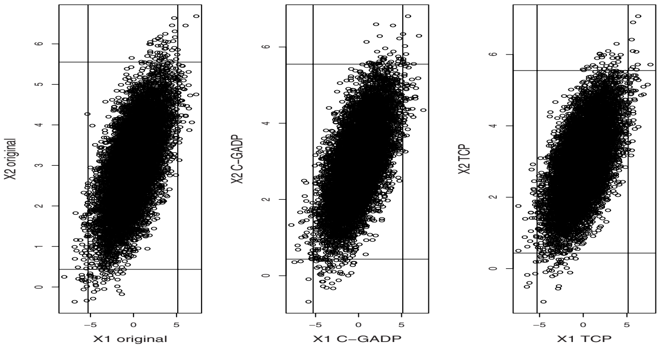


Fig. 4. Bivariate distributions of the sensitive variables under the original data and the C-GADP and t-copula perturbed data: Example 2

5.3 Discussion

The two examples presented illustrate that both C-GADP and the t-copula perturbation method preserve the marginals and concordance measures ρ_S and ρ_τ regardless of the microdata to be masked (and of the degree of tail dependence in the data).

If we focus on association between pairs of variables in the original data, however, while the C-GADP approach preserves bivariate distributions when no tail dependence is present (see Fig. 4) it performs quite poorly in the presence of tail dependence (see Fig. 2). The TCP method, on the other hand, works fairly well in both cases. The extra parameter ν in the t-copula approach, allows the TCP method to adjust for different degrees of tail dependence, at least for those cases for which tail dependence can be properly described by a t-copula. For these cases, if the size of the original data is sufficiently large, absence of tail dependence in the data will results in a estimate $\hat{\nu}$ of ν very large and the t-perturbation method coincides with the C-GADP approach (this is the case in the example 2). On the other hand for sufficiently large data sets, presence of tail dependence in the data will results in a "small" value of the estimate $\hat{\nu}$ of ν and the TCP method correctly captures the tail dependence outperforming the C-GADP approach (as in example 1).

6 Conclusions

The objective of this paper was to illustrate the usefulness of using the t-copula for perturbing numerical confidential variables. This approach extends the Gaussian copula approach suggested in the C-GADP procedure developed in [10]. The results indicate that the TCP approach provides all the benefits of the C-GADP approach. Additionally, the TCP approach maintains some important types of tail dependence that the C-GADP approach does not. It must be emphasized that even though we have used Gaussian marginals as examples, the TCP approach will maintain monotonic relationships and symmetric tail dependence among non-Gaussian marginals as well. These results also provide important information to the data provider. In general, the data provider is interested in providing the data of the highest possible quality subject to the condition that disclosure risk is minimized. The above results indicate that using the TCP approach for data perturbation would provide data of higher quality than other approaches while minimizing the risk of disclosure (at least within the CDA framework of [6]).

Despite the attractive features of the proposed TCP procedure, there are several issues that need to be addressed for a proper use of the method (strong symmetry of the t-copula, empirical assessment of disclosure risk, performance of the method for data sets with diverse dependence structures). Several extensions of the TCP approach are also possible. We hope to address this issue and evaluate the extension in the complete version of the paper.

References

1. Arslan, O.: Family of Multivariate Generalized *t* Distributions. Journal of Multivariate Analysis 89, 329–337 (2004)
2. Burridge, J.: Information Preserving Statistical Obfuscation. Statistics and Computing 13, 321–327 (2003)
3. Clemen, R.T., Reilly, T.: Correlations and Copulas for Decision and Risk Analysis. Management Science 45, 208–224 (1999)
4. Demarta, S., McNeil, A.J.: The *t* Copula and Related Copulas. International Statistical Review 73, 111–129 (2005)
5. Little, R.J.A.: Statistical Analysis of Masked Data. Journal of Official Statistics 9, 407–426 (1993)
6. Muralidhar, K., Sarathy, R.: A Theoretical Basis for Perturbation Methods. Statistics and Computing 13, 329–335 (2003)
7. Nelsen, R.B.: An Introduction to Copulas. Springer, New York (1999)
8. Reiter, J.P.: Simultaneous Use of Multiple Imputation for Missing Data and Disclosure Limitation. Survey Methodology 27, 235–242 (2004)
9. Rubin, D.B.: Discussion: Statistical Disclosure Limitation. Journal of Official Statistics 9, 462–468 (1993)
10. Sarathy, R., Muralidhar, K., Parsa, R.: Perturbing Nonnormal Confidential Attributes: The Copula Approach. Management Science 48, 1613–1627 (2002)

Anonymisation of Panel Enterprise Microdata – Survey of a German Project

Maurice Brandt[1], Rainer Lenz[2], and Martin Rosemann[3]

[1] Federal Statistical Office Germany, Gustav-Stresemann-Ring 11, 65189 Wiesbaden
maurice.brandt@destatis.de
[2] University of Applied Science, Holzstrasse 36, 55116 Mainz
rainer.lenz@fh-mainz.de
[3] Institute for Applied Economic Research, Ob dem Himmelreich 1, 72074 Tübingen
martin.rosemann@iaw.edu

Abstract. The access of the scientific community to cross-section data in the field of business statistics in Germany has considerably improved over the last few years. The purpose of the project on "Business Panel data and de facto anonymisation" is to extend the data infrastructure in Germany for panel data on local units and on enterprises, so that business statistical data can be made available to empirical researchers for use on their own workstations. This paper gives an overview of the project, describes the data sets and the anonymisation methods which are considered to create scientific use files. The work to assess the analysis potential and anonymity of the data is outlined as well. Also, an example is given, applying anonymisation methods to achieve the best possible results regarding the well known trade off between data confidentiality and analysis potential.

1 Introduction

The bases for anonymising german enterprise microdata were developed in the project on "De facto anonymisation of business microdata" (Lenz et al. 2006, Ronning et al. 2005). A major result of the project was that so called de facto anonymisation of business statistical data can be achieved on a cross-section basis. De facto anonymisation means, that the costs of trying to reidentify records in the dataset must be higher than the benefit gained by the disclosed information. In this case a rational data intruder would not even try to deanonymise the dataset, because he or she would have to put an unreasonably high amount on work, time, manpower and specialized knowledge in the data attack. The project on "Business Panel data and de facto anonymisation"[1] started at the beginning of 2006 and is intended to clearly improve both, the data infrastructure in Germany regarding panel data on local units and enterprises and the access of the scientific community to those data. The project deals with an improvement of the data supply by longitudinal linkage of statistics which so far have

[1] The project is carried out jointly by the Institute for Employment Research (IAB), the Institute for Applied Economic Research (IAW), the Research Data Centre (FDZ) of the statistical offices of the states and the Research Data Centre of the Federal Statistical Office.

J. Domingo-Ferrer and Y. Saygın (Eds.): PSD 2008, LNCS 5262, pp. 139–151, 2008.

been used mainly on a cross-section basis. At the end of the project de facto anonymised files, so called scientific use files will be created. Scientific use files are licensed files that can only be used by institutions doing independent scientific research. Those scientific use files can be released to the users after they have signed a contract with the data distributing institution. They should contain sufficient analysis potential, so that they can be used for scientific analysis. It is also planned to produce so called public use files, which can be used by everybody and hence have to be absolutely anonymised. Thus, regarding public use files the analytical validity can be quite low. Since the applied anonymisation methods are very strong here, it does not make sense to use them to gain results for scientific research, but it is still possible to use this material for the statistical education in universities to make the students familiar with panel data and statistical methods or econometric models. The project focuses on the cost structure survey in manufacturing, the monthly reports in manufacturing, the survey of investments, the industrial small units survey and the turnover tax statistics, which are processed as longitudinal data sets as part of the project. The local units panel of the Institute for Employment Research was selected for anonymisation by means of multiple imputation.

As another important element of the project, the linked longitudinal data have been complemented by information from the official business register. The main purpose of that work is to identify by means of the business register (cf. Sturm 2006) reasons for missing data, specially demographic information about enterprises, in longitudinal terms and thus to increase the analysis potential of the data. As regarding turnover tax statistics, the turnover data have been complemented – on the basis of the business register – by employees data for the years 2001 to 2005.

Panel data are demanded more and more often by scientific users because only with such data it is possible to show the dynamics, changes and processes over time. Another advantage of panel data is that unobservable heterogeneity can be considered. However, the positive aspects provided by panel data for research evaluations might also prove to be an additional challenge to anonymisation. This is because, across several waves, a structure in the data can be detected which gives additional knowledge to a potential data intruder that is helpful in reidentification attempts (Lenz 2008).

With a view to maintaining the analysis potential of panel data it must be ensured that developments over time can adequately be analysed also by means of anonymised data and that panel-econometric methods continue to produce consistent estimates (Biewen/Ronning/Rosemann 2007).

One of the questions to be answered by the project is the extent to which the anonymisation methods originally developed for cross-section data can be further developed for the anonymisation of panel data and what impact such methods have on data protection and on the analysis potential of panel data of business statistics.

The outline of the paper is as follows. Chapter 2 contains a description of the datasets and the editing of the data in this project. Chapter 3 illustrates the anonymisation methods of panel data and the impacts of these methods on analytical validity. Chapter 4 gives an overview about the possibilities to measure the disclosure risk to achieve de facto anonymity of panel data. In chapter 5 we give an example of the applied anonymisation methods and summarize the optimal results that could be achieved to level out the balance between data confidentiality and analysis potential. The paper ends with a summary and outlook on further work and projects.

2 The Data Sets of the Project

For the longitudinal linkage and the subsequent anonymisation, business data were selected for which some experience is available regarding cross-section anonymisation and which are demanded most often by researchers.[2]

2.1 Monthly Reports, Survey of Investments and Survey of Small Units

Based on the local units as a unit of analysis, the monthly reports in manufacturing, mining and quarrying are a longitudinal linkage of the years from 1995 to 2005. They contain information on employees, wages and salaries, and turnover (Statistisches Bundesamt 2007a). The survey of investments, however, provides information on highly different types of investments (Statistisches Bundesamt 2007c) and basically contains the same local units as the monthly reports. The monthly reports represent a complete enumeration of the local units with 20 or more employees.[3] The range of data is complemented by the survey of small units of the years 1995 to 2002, which supplies information from local units with 19 or fewer employees.

For the panel data set, the individual data supplies have been aggregated to form an annual data supply. The data contain information on employees, turnover (domestic and foreign turnover), hours worked, wages and salaries, and investments (Konold 2007). Wagner (2007) contains some examples of comments on the research potential of the monthly reports.

2.2 Cost Structure Survey

The cost structure survey is a stratified sample with almost 18,000 enterprises each year. The data of the cost structure survey in manufacturing, mining and quarrying are designed as a panel data set for the years from 1995 to 2005. The cost structure survey is suited for manifold structural analyses (Fritsch et al. 2004) and provides comprehensive information on output, the production factors used, and on the value added of enterprises with at least 20 employees (cf. Statistisches Bundesamt 2007d). The panel data set contains about 43,000 observations for the years 1995 to 2005. The way of processing allows to perform analyses both on a cross-section basis for the reference year and on a longitudinal basis. For the period from 1995 to 2005, there are about 2,000 enterprises which were questioned every year. A large part of those enterprises come from areas fully covered (branches with few cases, large enterprises). For the years 1999 to 2002, there are still just under 13,300 enterprises which were questioned every year, thus providing sufficient potential for scientific analyses and shall cover the period for the scientific use file (Brandt et al. 2008).

[2] In consequence of the project on “Anonymisation of business microdata”, further enterprise statistics such as the structure of earnings survey, could be anonymised (cf. Hafner and Lenz 2007).

[3] An exception is 14 economic branches with 10 or more employees (cf. Statistisches Bundesamt 2007b/c).

2.3 Turnover Tax Statistics

The longitudinal linkage of turnover tax statistics comprises a data set of a total of some 4.3 million observations, about 1.9 million of which can be linked for the period from 2001 to 2005 to form a real panel data set with observations for each year. In a first step, the panel data set for 2001-2005 was established for special analyses at the Federal Statistical Office and for remote data purposes. For every case, the file contains a data set with a total of 156 variables for 5 reference years, with differing numbers of variables actually occurring, depending on the existence of the enterprise in the relevant year. Turnover tax statistics contains information on all taxable turnovers, turnover tax, prior tax, and duration of tax liability (Statistisches Bundesamt 2005).

2.4 IAB Panel of Local Units

The IAB (Institute for Employment Research) panel of local units is a representative survey among employers on local unit items influencing employment and covers a stratified sample of all local units with at least one employee subject to social insurance contributions in Germany. The panel contains information allowing to perform analyses of the development of labour demand on the labour market in Germany. Items covered include information on the employment trend, weekly hours worked, turnover, and export share, investments and innovation in the local unit, public subsidies, staff structure, vocational training and apprenticeship positions, staff recruited and staff leaving, search for new staff, wages and salaries, hours worked in the local unit, advanced training and continuing education. The local units panel has been produced every year since 1993 in western Germany and since 1996 in eastern Germany by the IAB research unit “Local units and employment”. The local units panel contains information of the various waves on about 4,300 to a maximum of some 16,000 local units (Bellmann 2002).

3 Anonymisation Methods for Panel Data and the Analytical Validity of Anonymised Panel Data

In the last decade a broad variety of anonymisation methods is described in literature (see for example Brand (2000), Höhne (2003), Ronning et al. (2005) and Rosemann (2006)). Anonymisation methods may be subdivided into two groups: methods reducing the information, and more recent methods modifying the values of numerical data (data perturbing methods). When an anonymisation concept for business micro data is developed a mix of these two approaches often seems to be the best solution. Information reducing methods such as the suppression of variables or presenting key variables in broader categories should be preferred, provided that the analyses of interest to the users can still be made. However, if it seems inevitable to additionally apply anonymisation measures which modify the data, a method has to be agreed upon and the parameters of that method need to be balanced appropriately (Lenz et al. 2006).

In Ronning et al. (2005) most known anonymisation procedures have been rated both with regard to data protection and to informational content left after perturbation. In particular micro aggregation or stochastic noise has been found convenient for continuous variables whereas “Post Randomization” (PRAM) can be recommended with some reservations for discrete variables. Additionally, most recently multiple imputation has been suggested by Rubin (1993) for data protection.

The basic idea of (deterministic) micro aggregation is to form groups of similar objects and to substitute the original values by the arithmetic mean of this group (Mateo-Sanz and Domingo-Ferrer 1998).[4] The variants of deterministic micro aggregation principally differ with regard to the question whether the micro aggregation is performed jointly for all numerical variables or separately for each variable.[5] In the first case therefore the same groups are formed for different variables when determining the averages. In the second case (individual ranking) the groups are formed for the several variables separately.

In the case of panel data we have r variables, T periods and N observations. So we can perform the micro aggregation (a) separately for all variables and all periods (Individual Ranking), (b) separately for all variables but jointly for all periods, (c) separately for all periods but jointly for all variables and (d) jointly for all periods and all variables.

Micro aggregation preserves the expected values original but leads to a decreasing variance in a finite sample. Therefore Höhne (2004a) develops a variant of individual ranking that preserves the variances too. He builds up groups of size four. Then for two of these observations in group i anonymised values are given by

$$x^{a}_{i,1/2} = \bar{x}_{i.} - sd(x_i) \tag{1}$$

whereas for the two other anonymised values

$$x^{a}_{i,3/4} = \bar{x}_{i.} + sd(x_i) \tag{2}$$

is used where $\bar{x}_{i.}$ is the average of the variable x in group i and sd(x_i) is the standard deviation of x in this group.

The alternative approach of addition or multiplication of stochastic noise is one of the most important data perturbating methods. In the additive case the noise variable usually is assumed to be normally distributed with expectation zero. To increase the data security one can use a mixture distribution of normal distributions where the expectations of the underlying component distributions are unequal to zero. In the case of anonymisation we can restrict ourselves to a mixture distribution of two normal distributed components with expectations −μ and μ (Roque 2000, Yancey et al. 2002, Höhne 2004b and Ronning et al. 2005).

We achieve better protection for larger firms if we use multiplicative noise (Ronning et al. 2005). In this case the expectation of the noise variable should be one and the values of the noise variable should be limited to the positive band. Several distributions can be used, e.g. lognormal or uniform distribution. As an alternative, also in the multiplicative case a mixture distribution of two normal distributions is used,

[4] For stochastic micro aggregation see Rosemann (2006).

[5] Also used are variants where the set of numerical variables is subdivided into groups first and where the variables of a group are then micro aggregated jointly (Ronning et al. 2005).

where the expectations are 1−f and 1+f. The parameter f as well as the standard deviations of the two components (which equal each other) is chosen in such a manner that the values of the noise variable remain positive.

A special variant of a mixture distribution was proposed by Höhne (2004b). The main idea of this approach is that for one observed unit all values are scaled down or scaled up. In other words, for every unit the probability to draw from a normal distribution with expectation 1−f is 0.5 and corresponds to the probability to draw from a normal distribution with expectation 1+f. If we adopt this anonymisation method on the case of panel data we can distinguish several variants for the multiplicative noise variable w_{ijt} of observation i, variable j and period t.

$$w_{ijt} = 1 + d_i f + \varepsilon_{ijt} \tag{3-1}$$

$$w_{ijt} = 1 + d_{ij} f + \varepsilon_{ijt} \tag{3-2}$$

$$w_{ijt} = 1 + d_{it} f + \varepsilon_{ijt} \tag{3-3}$$

$$w_{ijt} = 1 + d_{ijt} f + \varepsilon_{ijt} \tag{3-4}$$

In all cases we assume $\varepsilon_{ijt} \sim N(0, \sigma_\varepsilon^2)$ and the variable d takes on +1 and −1 with probability 0.5.

Another auspicious method to anonymise panel data is multiple imputation (Rubin 1993, Raghunathan et al. 2003). In 1993 Rubin suggested to generate fully synthetic data sets to guarantee confidentiality. His idea was to treat all the observations from the sampling frame that are not part of the sample as missing data and to impute them according to the multiple imputation framework. Afterwards, several simple random samples from these fully imputed datasets are released to the public.

However, the quality of this method strongly depends on the accuracy of the model used to impute the “missing” values. If the model doesn’t include all the relationships between the variables that are of interest to the analyst or if the joint distribution of the variables is mis-specified, results from the synthetic data set can be biased. Furthermore, specifying a model that considers all the skip pattern and constraints between the variables can be cumbersome if not impossible

To overcome these problems, a related approach suggested by Little (1993) replaces observed values with imputed values only for variables that are publicly available in other databases (key variables) or for variables that contain especially sensitive information leaving most of the data unchanged. This approach has been adopted for some data sets in the US. In our project both approaches are tested in time with data of the IAB establishment panel (first results can be found in Drechsler et al. (2007) and Reiter and Drechsler (2007)).

The methods described above should ensure confidentiality of panel data at the same time the usefulness of data should be gained. The analytic potential is limited on the one hand by the fact that certain analyses are excluded from the start by the anonymisation procedures it selves because either the issue in question cannot be analysed anymore or the method to be used and equivalent methods cannot be applied anymore. This could be the main problem in the case of using methods reducing the information. On the other hand, such limits result form anonymised data producing results which differ from those based on the original data. When anonymisation procedures are assessed which modify the data, the focus is on the second aspect.

When we use data perturbating methods we have to ensure that distributional properties of the data do not change too much. However, in the project "Business Panel data and de facto anonymisation" the impacts of data perturbating methods on analysis using special qualities of panel data are in focus. On the one hand the project analyses the impacts of the described data perturbating methods on descriptive distribution measures where cross-sectional measures are supplemented by special aspects of panel data, for instance measures relating to the rates of change. On the other hand we focus on the effects of these methods on the estimation of econometric panel models, particularly if we use the within-estimator to control for individual unobservable heterogeneity. These analyses include theoretical derivations as well as simulation experiments and examples with data of official statistics. First results of this work are available.

Biewen et al. (2007) show that the within estimator is consistent in the case of anonymisation by individual ranking. These results correspond to the results of Schmid (2006) for the OLS-estimator. Biewen (2007) derives a consistent within-estimator in the case of anonymisation by multiplicative stochastic noise. The paper focuses on the case of no autocorrelation. Ronning (2007) deals with the effects of stochastic noise using a mixture distribution, for instance the method proposed by Höhne (2004b). In the case of panel data he focuses on the variant described in formula (3-1). However such a distribution will imply correlation of measurement errors. This is of special concern if linear (or nonlinear) models are estimated from data anonymised in such a way. This case so far had not received much attention since usually measurement errors are assumed to be independent across variables. It can be shown that the measurement error of the dependent variable in this case no longer can be considered as harmless to estimation. A consistent fixed effects estimator using the method of Höhne can be found in Ronning (2007) as well as in Biewen (2007). Ronning and Rosemann (2007) present a special approach of the simulation extrapolation estimator (SIMEX estimator) to tackle these problems also in nonlinear models. Biewen and Ronning (2007) expound the problems of serial auto correlation in the case of multiplicative stochastic noise. Actual project work deals with the method of instrument variable estimation to tackle the problems in case of auto correlated regressor variables.

4 Approaches to Assessing De Facto Anonymity

In order to evaluate the degree of anonymity of previously anonymised micro data, it was necessary to develop a technique for simulating data intrusion scenarios a potentially attacking data intruder might apply. One important constellation is the so-called database cross match scenario. In a database cross match scenario, an attacking data intruder tries to assign as many external database units as possible (additional knowledge) uniquely to units of an anonymised target database in order to extend the external database by target database information.

In a first phase, the database cross match scenario was mathematically modelled as a multicriteria assignment problem, which was then converted, by way of suitable parameterisation, into an assignment problem with one target function to be minimised. Then, the main concern was to choose the best-fitting coefficients of this target function. Whereas in the past a distance measure, generated across all matching variables of the two data sources (key variables and overlaps), proved to be well suited for the examination of cross-sectional data (Lenz 2006), it turned out that the examination of panel data requires the use of additional, more elaborated measures. As the information on variables, which in the case of panel data is available to a potential data intruder, has been collected in several waves, it seems obvious that this more complex structure should be reflected in the coefficients of the linear program as well. With that goal in mind we have implemented and tested several promising approaches. A more detailed description of these approaches can be found in Lenz (2008).

4.1 Conventional Distance Based Approach

For every numerical key variable v_i and every pair of records *(a,b)* in the two data sources, the standardised square deviation is calculated. Afterwards, these component deviations are summed up. It may be advisable in some cases to assign additional weights to the various deviations on variable level. However, a weakness of that measure becomes apparent in cases where the definition of some key variable slightly differs between the two data sources, for example, if a variable such as "number of employees" relates to the number of all employees in absolute terms in one data set, whereas that number is converted into full-time workers in the other data set.

4.2 Correlation-Based Approach

Let $v^e_1,\dots , v^e_k$ and $v^t_1,\dots, v^t_k$ be ordinal key variables of the external and target data, respectively. We define $\boldsymbol{v}^e$ and $\boldsymbol{v}^t$ as variables from which k realisations have been drawn and calculate the empirical correlation $corr(\boldsymbol{v}^e; \boldsymbol{v}^t)$ using Spearman's coefficient. The less this coefficient deviates from 1 the more likely the record pair *(a,b)* belongs to the same enterprise. Note that this coefficient can be applied either in case of numerical (and hence also ordinal) variables or in case of categorical variables, whose range forms a well-ordered set.

4.3 Distribution Based Approach

In a panel data situation we can take it for granted that an attacking data intruder will have information over several years for every key variable, for example, total turnover of an enterprise from 1999 to 2002. In general, we can assume the existence of a bias between the two sources of data in these variables. In order to counteract this problem, we consider the annual changes of a key variable and treat them like a frequency distribution of a discrete variable. Hence, we can apply statistical methods in order to measure the "similarity" of the frequency distributions on either side, external and target data.

4.4 Collinearity Approach

A data intruder might have information on two key variables over a period of n years in both sources of data, e.g., “total turnover” $(u_1, \ldots, u_n)$ and “number of employees” $(b_1, \ldots, b_n)$ of an enterprise. If we interpret the pairs of values (u_i,b_i) as realisations of two random variables, those units that belong together in the different data sources can be expected to reveal empirical correlation cofficients that are ‘similar’. It should, however, be considered that what is measured by correlation is just the linear interrelation of two variables. In special cases the two estimated correlation coefficients can diverge from each other very clearly, even if the variables are linked by a direct functional relationship.

Overall, risk tables can serve as a basis for the decision whether an experimentally anonymised data file can be rated as de facto anonymous. For example, table 1 below can be used to enter risks of reidentification by size class of employees and approach to the coefficients of the target function.

Referring to the subsections 4.1 - 4.4 the approaches to the coefficients are denoted by A_{conv}, A_{corr}, A_{dist} and A_{coll}.

Table 1. Risks of re-identification by size class of employees and coefficient approach

Class\Strategy	A_{conv}	A_{corr}	A_{corr}	A_{coll}
20 - 49	0.25	0.08	0.08	0.06
50 - 99	0.26	0.11	0.13	0.07
100 - 249	0.34	0.10	0.11	0.06
250 - 499	0.61	0.18	0.23	0.15
500 - 999	0.72	0.52	0.63	0.28
1 000 and more	0.85	0.37	0.45	0.30

Once the coefficients $d(a_i,b_j)$ are calculated, one can solve the linear assignment problem using classical established methods such as the simplex method. For larger data blocks (typically generated when dealing with tax statistics) it is recommendable for reasons of efficiency that approximation heuristics should be used. Fortunately, the usage of appropriate heuristics yields results near the optimum solution of the assignment problem, see Lenz (2003).

5 Example: German Cost Structure Survey 1999 - 2002

In this section, we focus on 13,300 target records of the German Cost Structure Survey which have been observed over four years 1999 to 2002. These data are anonymised applying Höhne’s method (see section 3) setting $f=0.11$ and $\sigma_\varepsilon=0.03$ as parameters.

Choosing this parameter constellation on the one hand results in the fact that 62 percent of the values change by more than 10 percent. On the other hand one can show that the anonymisation does not influence the descriptive distribution measures in a considerable way. Due to the specific construction of Höhne’s method only two percent of the rates of change deviate by more than ten percent from the original ones.

The mean of the variables as well as the mean of the rates of change over all variables and points in time do not vary by more than ten percent. Only three percent of the standard deviations (of the variables and the rates of change) change by more than 10 percent. The variation of correlation coefficients (Bravais-Pearson) caused by the anonymisation does not exceed 0.1. The same results hold for Spearman's rank correlation coefficient if we ignore the rates of change. Actually, variation is not higher than 0.05. But also for the rates of change only five percent of Spearman's rank correlation coefficients vary by more than 0.05 if we focus on the correlations between variables. For correlations between several periods of time, in only eight percent of the cases the limit of 0.05 is exceeded.

In order to simulate data intrusion scenarios we generated as external data an excerpt of about 9,500 records of the so-called MARKUS database. Both, target and external survey, share the following key variables according to the years 1999 to 2002: branch of economic activity at the 2-digit sector level (NACE 2), regional key (recoded to two categories eastern and western Germany), total turnover and total number of employees. Additionally, we introduced five employee size classes '20-49', '50-99', '100-249', '250-499' and 'at least 500'. As a matching result, for each combination (NACE 2/region/employee size class) we observed disclosure risks far below 0.5, so that we decided to rate the target data to be de facto anonymously. Finally, we carried out single individual recherches regarding the three dominating enterprises of each sector.

6 Outlook

Already within the scope of the project on "Business Panel data and de facto anonymisation", some panel data sets were supplied. They are already used in some research projects. The cost structure survey for the years 1995 to 2005, the monthly reports from 1995 to 2005, the survey of investments from 1995 to 2005 and the survey of small units for the years 1995 to 2002 in manufacturing as well as the data of the turnover tax statistics for 2001 to 2005 are available through remote data access and by using safe scientific workstations at the statistical offices.

One result of the project is that de facto anonymisation of panel data can be achieved. First Scientific Use Files for data utilisation on one's own workstation will presumably be made available at the beginning of 2009. The project should permit to automate the processing and anonymisation of other business statistics over time. This should make it easier and faster to release business panel data to the users. Also, the experience thus acquired will be used for further projects such as the integration of business data from various surveys and years. This is the subject of the new project "Integrated enterprise data for Germany" and will be the next step in the field of data production. Moreover, the central business register allows joining data from different sources and years. This produces complex data sets with a lot more business information to analyse than ever before in Germany. Regarding access to those data there is a new challenge for the statistical offices concerning data confidentiality. It seems almost impossible to produce scientific use files for integrated business statistics over time because all the new information obtained has to be anonymised or in the worst case the information has to be taken out of the data set. That means in the end that on the one hand the statistics will be enriched with information from different sources and on the other hand this information would have to be deleted afterwards caused by anonymisation. For this reason, it seems necessary to find next to scientific use files alternative ways to access also complex data sets. Those circumstances and the international development in this area show that there is no way around remote data access. With remote data access it is possible to work with the original data; the challenge is to keep the produced output safe to ensure confidentiality. This topic is part of the latest project in the Research Data Centre of the Federal Statistical Office in Germany “An informational infrastructure for E-Science Age”, which has already been requested. The goal of this project is to develop and automatise remote data access in Germany. One part deals with the production of so called data structure files (these are absolutely anonymised files possessing the same structure as the original data; they are sent to the researcher so that he or she can develop his or her analysis programs). The second part of the planned project deals with the development of procedures for automatic output checking. These are the first steps in the direction of real remote data access and the project shall provide the methodology to implement the concept in a technical solution. The project benefits from the work done in “Business Panel data and de facto anonymisation” in several ways. Particularly anonymisation methods for panel data developed for the creation of scientific use files also can be used to construct data structure files.

References

Bellmann, L.: Das IAB-Betriebspanel. Konzeption und Anwendungsbereiche. In: Allgemeines Statistisches Archiv., Bd. 86, H. 2, pp. 177–188 (2002)

Biewen, E.: Within-Schätzung bei anonymisierten Paneldaten, IAW-Discussion Paper 34 (2007)

Biewen, E., Ronning, G., Rosemann, M.: Estimation of Linear Panel Models with Anonymised Business Data. In IAW-Report 1/2007, 87–114 (2007)

Biewen, E., Ronning, G.: Estimation of Linear Models with Anonymized Panel Data. In: The Workshop Methodical aspects of the anonymisation of panel data, Tübingen (November 2007) (unpublished article)

Brand, R.: Anonymität von Betriebsdaten – Verfahren zur Erfassung und Maßnahmen zur Verringerung des Reidentifikationsrisikos. Beiträge zur Arbeitsmarkt- und Berufsforschung, 237 (2000)

Brandt, M., Oberschachtsiek, D., Pohl, R.: Neue Datenangebote in den Forschungsdatenzentren – Betriebs- und Unternehmensdaten im Längsschnitt. In: Wirtschafts- und Sozialstatistisches Archiv., 2008 (to appear, 2008)

Drechsler, J., Dundler, A., Bender, S., Rässler, S., Zwick, T.: A new approach for disclosure control in the IAB establishment panel. Multiple imputation for a better data access.Tech. rep., IAB Discussion Paper, No.11/2007 (2007)

Fritsch, M., Görzig, B., Hennchen, O., Stephan, A.: Cost Structure Surveys in Germany, Schmollers Jahrbuch. Journal of Applied Social Science Studies 124, 557–566 (2004)

Hafner, H.-P., Lenz, R.: Anonymisation of linked employer employee datasets using the example of the german structure of earnings survey. Work session on statistical data confidentiality, Manchester (2007)

Höhne, J.: Methoden zur Anonymisierung wirtschaftsstatistischer Einzeldaten. In: Gnoss, R., Ronning, G. (eds.) Anonymisierung wirtschaftsstatistischer Einzeldaten, Wiesbaden, pp. 69–94 (2003)

Höhne, J.: Weiterentwicklung von Mikroaggregationsverfahren (2004a)

Arbeitspapier des Projekts Faktische Anonymisierung wirtschaftsstatistischer Einzeldaten

Höhne, J.: Varianten von Zufallsüberlagerungen, Arbeitspapier des Projekts Faktische Anonymisierung wirtschaftsstatistischer Einzeldaten (2004b)

Konold, M.: New possibilities for economic research through integration of establishment-level panel data of German official statistics. Journal of Applied Social Science Studies (Schmollers Jahrbuch) 127(2), 321–334 (2007)

Lenz, R.: Disclosure of confidential information by means of multi-objective optimisation. In: Proceedings of the Comparative Analysis of Enterprise Data Conference (CAED), CD-ROM publication, London (2003)

http://www.statistics.gov.uk/events/caed/abstracts/lenz.asp

Lenz, R.: Measuring the disclosure protection of micro aggregated business microdata - An analysis taking the example of German Structure of Costs Survey. Journal of Official Statistics 22(4), 681–710 (2006)

Lenz, R., Rosemann, M., Vorgrimler, D., Sturm, R.: Anonymising business micro data - results of a German project. Journal of Applied Social Science Studies (Schmollers Jahrbuch) 126(4), 635–651 (2006)

Lenz, R.: Risk Assessment Methodology for Longitudinal Business Micro Data. Wirtschafts- und Sozialstatistisches Archiv. (to appear, 2008)

Little, R.: Statistical Analysis of Masked Data. Journal of Official Statistic 9, 407–426 (1993)

Mateo-Sanz, J., Domingo-Ferrer, J.: A Method for Data-Oriented Multivariate Microaggregation. In: Statistical Data Protection, Proceedings of the conference Eurostat 1999 (1998)

Raghunathan, T., Reiter, J., Rubin, D.: Multiple Imputation für Statistical Disclosure Limitation. Journal of Official Statistics 19, 1–16 (2003)

Reiter, J., Drechsler, J.: Releasing Multiply-Imputed Synthetic Data Generated in Two Stages to Protect Confidentiality. IAB Discussion Paper, No.20/2007 (2007)

Ronning, G.: Stochastische Überlagerung mit Hilfe der Mischungsverteilung, IAW-Discussion Paper 30 (2007)

Ronning, G., Rosemann, M.: SIMEX Estimation in Case of Correlated Measurement Errors. In: The Workshop Methodical aspects of the anonymisation of panel data, Tübingen (November 2007) (unpublished article)

Ronning, G., Sturm, R., Höhne, J., Lenz, R., Rosemann, M., Scheffler, M., Vorgrimler, D.: Statistik und Wissenschaft, Handbuch zur Anonymisierung wirtschaftsstatistischer Mikrodaten, Band 4 (2005)

Roque, G.: Masking Microdata Files with Mixtures of Multivariate Normal Distributions, Ph.D. thesis, University of California, Riverside (2000)

Rosemann, M.: Auswirkungen datenverändernder Anonymisierungsverfahren auf die Analyse von Mikrodaten. IAW-Forschungsbericht, Nr. 66, Tübingen (2006)

Rubin, D.: Discussion: Statistical Disclosure Limitation. Journal of Official Statistics 9(2), 461–468 (1993)

Schmid, M.: Estimation of a linear model under microaggregation by individual ranking, Allgemeines Statistisches Archiv. 90(3) (2006)

Statistisches Bundesamt, Umsatzsteuerstatistik, Qualitätsbericht (2005)

Statistisches Bundesamt, Monatsbericht für Betriebe des Verarbeitenden Gewerbes sowie des Bergbaus und der Gewinnung von Steinen und Erden, Qualitätsbericht (2007a)

Statistisches Bundesamt, Produktionserhebungen, Qualitätsbericht (2007b)

Statistisches Bundesamt, Investitionserhebung bei Unternehmen und Betrieben des Verarbeitenden Gewerbes sowie des Bergbaus und der Gewinnung von Steinen und Erden, Qualitätsbericht (2007c)

Statistisches Bundesamt, Kostenstrukturerhebung im Verarbeitenden Gewerbe, im Bergbau sowie in der Gewinnung von Steinen und Erden, Qualitätsbericht (2007d)

Sturm, R., Tümmler, T.: Das statistische Unternehmensregister - Entwicklungsstand und Perspektiven. In: Wirtschaft und Statistik 10/2006, pp. 1021–1036 (2006)

Wagner, J., Kaiser, U.: Neue Möglichkeiten zur Nutzung vertraulicher amtlicher Personen- und Firmendaten. RDC-discussion paper 20 (2007)

Yancey, W., Winkler, W., Creezy, R.: Disclosure Risk Assessment in Perturbative Micro Data Protection. In: Domingo-Ferrer, J. (ed.) Inference Control in Statistical Databases. LNCS, vol. 2316, pp. 135–152. Springer, Heidelberg (2002)

Towards a More Realistic Disclosure Risk Assessment

Jordi Nin, Javier Herranz, and Vicenç Torra

IIIA, Artificial Intelligence Research Institute
CSIC, Spanish National Research Council
Campus UAB s/n
08193 Bellaterra, Catalonia, Spain
{jnin,jherranz,vtorra}@iiia.csic.es

Abstract. The *score* was introduced in 2001 in order to compare different perturbative methods for statistical database protection. It measures the trade-off between utility (information loss) and privacy (disclosure risk of the released data). Since its introduction, the score has been widely accepted and used in the statistical database community. In particular, some methods are sometimes prefered to others depending on the obtained results in the original computation of the score.

In this paper we argue that some original aspects of the score computation, specially those related to the disclosure risk, should be revisited. Informally, the reason is that they do not consider the best possible situation for the intruder, and so they do not measure the real level of privacy. We add some experimental results which support our claims. More importantly, we propose some modifications which can/should lead in the future to a more fair, realistic and useful computation of the score.

1 Introduction

There are many real situations where confidential data of people (respondents) is published by statistical agencies, to be used by decision makers, politicians, researchers, etc. The proliferation of such datasets in the Internet, for example, is easy to check. This dissemination of confidential information should ensure, however, that the privacy of the respondents is protected in some way, to be in accordance with current laws and regulations. For example, a person would not be happy if some published dataset contained a record with some attributes which identify him univocally, concatenated with some confidential attributes such as the income or the diseases he has suffered from.

One approach to achieve some level of privacy in this scenario is the application of *perturbative protection methods* to the confidential data, before making them public. A large number of such methods exist (see [1], [7] and [20] for three good surveys on data protection methods). Besides protecting the privacy of the respondents, the main goal is that the data protection method preserves as much as possible the statistical utility of the original data. Of course, the values of privacy and statistical utility are inversely related.

J. Domingo-Ferrer and Y. Saygın (Eds.): PSD 2008, LNCS 5262, pp. 152–165, 2008.

A particular perturbative protection method is well-considered if it achieves a good trade-off between privacy and statistical utility. There are different ways to measure this trade-off. Maybe the most simple and intuitive one is the *score*, which was presented in [5,6]. It just measures the average between two quantities: one of them analyzes the information loss which is produced by the application of the protection method, and the other one counts the risk that an intruder may obtain any information that breaks the privacy of the data, after the protected dataset has been released. Since these two quantities are generic, meaning that they can be computed independently of the considered protection method, the score is a very good way to compare and classify different methods. This is exactly what was done in [7]: a specific computation of the score was implemented and applied to a large number of protection methods, with different parameterizations, leading to a ranking where methods were classified according to the obtained score. In this ranking, some of the first positions were occupied by different parameterizations of two protection methods: microaggregation [4] and rank swapping [2,14].

This ranking has been often considered as a criterion to choose one method or another, when protecting a statistical dataset. Furthermore, after the publication of this score ranking, subsequent works proposing new protection methods or modifications to existing ones always consider the ranking in [7] as a benchmark to compare the score of the new methods, in order to argue that they are (or are not) good enough.

The goal of this paper is to show that some aspects of the original computation of the score are debatable. These aspects are related to how one computes the disclosure risk, in particular the risk of re-identification (or record linkage): an intruder wants to link an original record with the corresponding protected record in the released dataset. One of the debatable aspects is perhaps philosophical: the original computation of the disclosure risk does not consider that an intruder will always choose the best possible re-identification method, when trying to re-identify a record and break therefore the privacy of the whole system. The other main aspect that we want to discuss is the way to define the attributes made available to an intruder for re-identification. The original definition, where the intruder is assumed to know the i first attributes of some original record(s), instead of i arbitrary attributes, can actually benefit the score of some particular methods, *e.g.* microaggregation, as we show with some experiments.

In order to eliminate these problems, we propose some modifications to the original algorithm for computing the disclosure risk of a data protection method. In our opinion, these modifications will result in a more realistic, fair and useful score. If our modifications are accepted by the statistical database community, then the new score should be computed for all the methods and parameterizations that appear in the ranking of [7], to obtain a new ranking, which would be more realistic and, for sure, different to the original one.

Organization of the paper. In Section 2 we explain the general framework of statistical database protection that we consider in this work. We review microaggregation as an example of a protection method. We also give the general overview of the definition of the score. Then in Section 3 we explain in more detail how the disclosure risk part of the score is usually computed. In Section 4 we discuss the two aspects of this computation that are not, in our opinion, completely correct. We propose some modifications to solve these problems. Section 5 shows some experiments on real data that support our arguments. Finally, we conclude our work in Section 6.

2 Preliminaries

A dataset X can be viewed as a matrix with n rows (*records*) and V columns (*attributes*), where each row contains V attributes of an individual. The attributes in a dataset can be classified in two different categories, *identifiers* (X_{id}) or *quasi-identifiers*, depending on their capability to identify unique individuals. Among the quasi-identifier attributes, we distinguish between *confidential* (X_c) and *non-confidential* (X_{nc}), depending on the kind of information that they contain.

We consider the following scenario for statistical disclosure control: (i) identifier attributes in X are either removed or encrypted, therefore we will write $X = X_{nc}||X_c$; (ii) confidential quasi-identifier attributes X_c are not modified, and so we have $X'_c = X_c$; (iii) a protection method ρ is applied to non-confidential quasi-identifier attributes, in order to preserve the privacy of the individuals whose confidential data is being released. This leads to a protected dataset $X'_{nc} = \rho(X_{nc})$. This scenario, which was used first in [7], has also been adopted in other works like [19].

2.1 An Example: Microaggregation

Microaggregation is one of the most popular, studied and used microdata protection methods. It builds small clusters of at least k elements of v attributes and replaces the original records by the centroid of the cluster to which the records belong.

The goal of a microaggregation method is to minimize the total Sum of Square Error

$$SSE = \sum_{i=1}^{c} \sum_{x_{ij} \in C_i} (x_{ij} - \bar{x}_i)^T (x_{ij} - \bar{x}_i),$$

where c is the total number of clusters, C_i is the i-th cluster and $\bar{x}_i$ is the centroid of C_i. The restriction is $|C_i| \geq k$, for all $i = 1, \ldots, c$.

If a microaggregation method is applied to all the V attributes of the original dataset X at the same time; then, the resulting protected dataset X' satisfies the property of k-anonymity: each protected record can correspond to at least k original records. However, in order to increase the statistical utility of the released (protected) information, statistical agencies usually split the whole dataset X in blocks of a few attributes, and then apply a microaggregation method to each block, independently. In this way, k-anonymity is not preserved any more [16].

In the case of univariate microaggregation ($v = 1$), there are polynomial time algorithms to obtain the optimal microaggregation [11]. However, for the multivariate case ($v > 1$), the problem of finding the optimal microaggregation is NP-hard. For this reason, multivariate microaggregation methods, like MDAV [9,12], are heuristic.

2.2 The Score: Evaluating Risk and Utility

A microdata protection method must guarantee a certain level of privacy (low disclosure risk). At the same time, since the goal is to allow third parties to perform reliable statistical computations over the released (protected) data, the protection method must ensure somehow that the protected data is statistically close to the original data.

Therefore, we have two inversely related aspects to measure for each microdata protection method: the *disclosure risk* (DR), which is the risk that an intruder obtains correct links between the protected and the original data; and the *information loss* (IL) caused by the protection method. When one of them increases, the other one decreases. The two extreme cases are the following ones: (i) if the original microdata is released, then information loss is zero, but the disclosure risk is maximal; (ii) if the original microdata is encrypted and then released, the disclosure risk is zero (if we exclude the possibility that the protected attributes are strongly statistically related to other known unprotected attributes), but the information loss is maximal.

There are different generic measures proposed in the literature to evaluate the quality of a data protection method. As we have stated in the introduction, we will use the *score*, which was introduced in [6] and used in several papers [15,21,22] to compare protection methods. The score is a simple and natural way to evaluate the trade-off between the information loss and the disclosure risk because it is defined as the average of these two values. Namely,

$$score = \frac{(IL + DR)}{2},$$

where IL denotes the information loss and DR denotes the disclosure risk. More details on the computation of DR are provided in the next section. Regarding IL, it is computed as $IL = 100\big(0.2\,IL_1 + 0.2\,IL_2 + 0.2\,IL_3 + 0.2\,IL_4 + 0.2\,IL_5\big)$, where IL_1 is the mean absolute error of the original microdata X with respect to the protected data X', IL_2 is the mean variation of the attribute average vectors, IL_3 is the mean variation of the attribute covariance matrices, IL_4 is the mean variation of the attribute variance vectors, and IL_5 is the mean variation of the attribute correlation matrices.

3 How Was DR Originally Computed?

To compute Disclosure Risk (DR), one considers two different approaches, the first one being the *interval disclosure risk*, ID, which is the average percentage of protected values falling into the intervals around their corresponding original values. The second approach is *record linkage risk* (or re-identification risk), which considers the scenario where an intruder has obtained an original record $x \in X$, possibly from a different external dataset Y, and tries to link it with the corresponding protected record $x' \in X'$. If he succeeds, then he can match the non-protected confidential information x_{nc} with the identifiers x_{id} that he obtained from Y, and so he breaks the privacy of this individual. Two standard record linkage methods are usually considered:

- Distance-based record linkage [17], where the original record is linked to the closest protected record, using for example the Euclidean distance. The average percentage of correctly linked records using this method is the *Distance based Linkage Disclosure risk*, DLD.
- Probabilistic record linkage [13], where the link is assigned in a probabilistic way, according to some criterion on some coincidence vectors (defined from the available sets of original and protected records). The average percentage of correctly linked records using this method is the *Probabilistic Linkage Disclosure risk*, PLD.

When computing disclosure risk for the score, half weight is given to record linkage and half weight is given to interval disclosure. Then, the risk of record linkage is defined as the average of the two methods. Formally, this corresponds to $DR = 0.25 \cdot DLD + 0.25 \cdot PLD + 0.5 \cdot ID$. The positive part of computing DR in this way is that all the components are generic, in the sense that they can be easily applied to a particular protection method or a particular data use. For this reason, it is quite easy to implement this measure and to compare the results (score) obtained by different data protection methods.

Since many data protection methods are probabilistic, one usually applies them many times to a certain database. Then one executes the corresponding methods for record linkage or interval disclosure, to obtain the disclosure risk, and finally one computes the average values for all these executions.

Last, but not least, when the Linkage Disclosure Risk is computed, there are many different possible situations, depending on the information (the amount of attributes, in particular) on the original record(s) $x \in X$ that is assumed to be available to the adversary. The strategy employed in the originally implemented computation of the score [7] was to consider as many different cases as attributes in the database. In the t-th case, the intruder was assumed to know only one t attributes of the original record(s), for $t = 1, \ldots, V$. Finally, the average of the linkage disclosure risks, for the V cases, was defined as the corresponding (probabilistic, or distance based) linkage disclosure risk for that protection method and that database.

4 Proposed Modifications for the Computation of DR

In this section we detect and discuss two debatable aspects of the original computation of the disclosure risk DR. Some experimental examples which support our opinion will be given in the next section. Besides arguing why these aspects are partially wrong, we propose in this section a possible way to modify them, which would lead to a new (and, in our opinion, better) definition of the disclosure risk component of the score.

4.1 Considering the Best Record Linkage Technique

As we have seen in Section 3, the 'record linkage part' of DR is computed as $0.25 \cdot DLD + 0.25 \cdot PLD$, i.e., as the average of the successful probabilities of correct record linkage when using the distance-based method and when using the probabilistic method. This seems a good and fair measure, for situations where an intruder cannot know which of the two record linkage methods works better. In this case, the best solution for him is to choose at random one of the two methods, and apply it to the specific original record(s) to be linked (maybe choosing one random method for each available original record).

But this approach is not realistic at all. In real life, intruders will know which method has been used to protect a particular dataset. Furthermore, they can artificially create their own databases to play with: they can protect them with the method they want, and they can execute different record linkage techniques for the resulting protected datasets and all the original records. As a result, they will obtain, for each analyzed protection method, which record linkage technique obtains the best re-identification results. If distance-based record linkage obtains much better results than probabilistic record linkage for a specific dataset protection method ρ, then this intruder will always use distance-based record linkage when trying to break the privacy of a dataset which has been protected using ρ. In this case, $0.25 \cdot DLD + 0.25 \cdot PLD$ is obviously not a realistic measure for the re-identification risk of ρ. Instead of this, one should consider $0.5 \cdot DLD$ in this example. In general, one should replace $0.25 \cdot DLD + 0.25 \cdot PLD$ with $0.5 \cdot MAX(DLD, PLD)$, in the computation of DR.

This argument can (and should) go even further. We are assuming that an intruder will try to correctly link original and protected records by using either the probabilistic or the distance-based record linkage techniques. But maybe a clever intruder is able to find other record linkage methods which work better than those two methods, either in general or at least for some particular dataset protection methods. An example of this fact has been shown in [15], where a new record linkage technique is presented, specifically designed for rank swapping, which finds a much larger number of correct links than the two generic techniques. Some experimental results are provided in Section 5.2.

Summing up, one should try to find, for each particular dataset protection method, which is the record linkage technique (either generic or specific) which works better for this method. If we define as *Linkage Disclosure Risk*, LDR, the average percentage of correctly linked records using this technique, then we should compute the disclosure risk DR as $DR = 0.5 \cdot ID + 0.5 \cdot LDR$. Of course, a drawback of this approach is that the value of the disclosure risk DR for a particular protection method can (and should) vary through time, depending on the progress in the area of record linkage: each time a new and better technique is found for this method, its real disclosure risk increases. Being more formal, each record linkage technique which improves the previous ones for some method gives a lower bound for the real value of LDR as well. For some methods, it is possible to give theoretical upper bounds for the real LDR. For example, any protection method which provides k-anonymity (such as microaggregation when applied to all the V attributes of the original dataset) satisfies $LDR \leq 1/k$.

4.2 Not All the Attributes Behave Equally

In the original implementation of the algorithm for computing the Linkage Disclosure risks, one considers V different scenarios, where V is the number of attributes in the dataset. For each $t = 1, \ldots, V$, the intruder is assumed to know only t original attributes of some original record(s), that he has obtained from external sources. His goal is to link these i attributes with the corresponding protected record in the released dataset X'. Once the percentages p_t of correctly linked records have been computed, for each of these scenarios $t = 1, \ldots, V$, the final Linkage Disclosure risk is computed as the average of all of them, $(\sum_{t=1}^{V} p_t)/V$.

Again, this seems to be a good and fair approach, because in real life situations, different intruders can have access to different amounts of original information. The mistakes can appear when one studies the way in which the t attributes available to the intruder are chosen. We can distinguish different possibilities:

(a) The optimal solution would consider the semantics of each attribute, analyzing which attributes are more likely to be obtained from external sources, and giving to these attributes more presence in the information held by the intruder. But this approach would be very inefficient, and very specific to each database.
(b) Another solution would consist in considering all the $\binom{V}{t}$ possible combinations of t attributes, computing the percentage of correct links for each combination, and then computing p_t either as the average/median of all of them, or as their maximum (again thinking of the best possible situation for the intruder). This algorithm, however, is quite inefficient if the number V of attributes is quite large.
(c) Finally, a good enough approach seems to be choosing t attributes in a complete random way, and defining p_t as the percentage of correct links that an intruder finds with this random combination of t available attributes.

The original computation of DR, in [7], was done by following the last approach (c). The problem is that the considered ‘random’ combination of t attributes was always formed by the first t attributes of the database. This choice was justified with the argument that this combination of t attributes behaves as randomly as any other one. But this is not completely true, if we take into account the protection method which has been applied to the original database. This can be easily understood with a particular example: microaggregation.

Suppose a database X contains 9 attributes, and that X is protected by applying k-microaggregation twice: the first time for the first 5 attributes, and the second time for the last 4 attributes (actually, in the microaggregation methods analyzed in [7], blocks were formed exactly in this way: a first block with the first v_1 attributes, a second block with the following v_2 attributes, and so on). In this case, if an intruder knows the first t attributes of some original record(s), for $t = 1, 2, 3, 4, 5$, then he will be able to execute his record linkage techniques only on the first block of 5 microaggregated attributes, because he does not have any information about the last 4 attributes. In particular, the percentage of correct links will be always $p_t \leq 1/k$. In a more realistic (or at least, average) case, the intruder will know t random attributes of the original record(s), belonging to the two blocks of aggregated attributes. Now, the intruder will be able to use information of the two blocks of microaggregated attributes, which will lead to a more successful percentage of correct links. For example, the bound $p_t \leq 1/k$ will not be valid any more.

Summing up, the specific way in which the selection of the t attributes available to the intruder was done in [7] can benefit some protection methods, for example, microaggregation. We provide in Section 5 some experimental results to exemplify this statement. To repair this small mistake, we suggest to use another approach to select the t attributes of the original record(s) available to the intruder. Namely, using a single but truly random combination of t attributes, as in (c), or considering all the combinations of t attributes, and computing p_t as the maximum or average percentage of these combinations, as in (b). Some examples of the values obtained with these approaches, for some instances of microaggregation, are also given in next Section 5.3.

5 Experiments

In this section we explain some experiments that we have run on real data, and which support the arguments that we have presented in the previous section. Namely, we first show how the real score of a particular protection method, rank swapping, varies when the linkage disclosure risk is computed by considering only the best record linkage technique for this method. Then, we also explain how the score of microaggregation can change when we consider different ways to define the i attributes which are available to an intruder who tries to identify original and protected records.

5.1 Data

The datasets used in this work, called Census, contains 1080 records consisting of 12 or 13 numerical attributes. Census was extracted using the Data Extraction System of the U.S. Census Bureau [3]. A complete description about the details of the construction of this dataset can be found in [8].

The data used to create this dataset was extracted from the file-group 'March Questionnaire Supplement - Person Data Files' of the data source 'Current Population Survey of the year 1995'. Not all the records of this survey were selected. Records with zero or missing values for at least one of the attributes were discarded to obtain the final 1080 records. The attributes selected to build the

Table 1. Attributes (or variables) of the Census dataset

id	Name	Description
$v1$	AFNLWGT	Final weight (2 implied decimal places)
$v2$	AGI	Adjusted gross income
$v3$	EMCONTRB	Employer contribution for health insurance
$v4$	ERNVAL	Business or farm net earnings in 19
$v5$	FEDTAX	Federal income tax liability
$v6$	FICA	Social security retirement payroll deduction
$v7$	INTVAL	Amount of interest income
$v8$	PEARNVAL	Total person earnings
$v9$	POTHVAL	Total other persons income
$v10$	PTOTVAL	Total person income
$v11$	STATETAX	State income tax liability
$v12$	TAXINC	Taxable income amount
$v13$	WSALVAL	Amount: Total wage & salary

Census dataset are described in Table 1. For the experiments with rank swapping (Section 5.2), we will use the 13 attributes. For experiments with MDAV (Section 5.3), we will use the 12 first attributes of the dataset.

5.2 Best Record Linkage for Rank Swapping

In this section we give a very illustrative example which supports the argument that we have presented in Section 4.1: the real score should be computed by taking into account the most effective record linkage technique.

The protection method of rank swapping, with parameter p and with respect to an attribute $attr_j$, can be defined as follows: first, the records of X are sorted in increasing order of the values x_{ij} of the considered attribute $attr_j$. For simplicity, assume that the records are already sorted, that is $x_{ij} \leq x_{\ell j}$ for all $1 \leq i < \ell \leq n$. Then, each value x_{ij} is swapped with another value $x_{\ell j}$, randomly and uniformly chosen in the set of still unswapped values, in the limited range $i < \ell \leq i + p$. Finally, the sorting step is undone. Usually, when rank swapping is applied to a dataset, the algorithm explained above is run for each attribute to be protected in a sequential way.

Table 2 shows the traditional score computation for different instances of rank swapping applied to the whole Census database, with 13 attributes. DR_Old is defined as $0.5 * ID + 0.25 * DLD + 0.25 * PLD$.

Table 2. Original Score calculation for rank swapping, with parameters $p = 2, 8, 16$

rs-p	IL	DLD	PLD	ID	DR_Old	Score_Old
rs-2	3.89	73.52	71.28	93.98	83.19	43.54
rs-8	16.54	32.13	11.74	62.11	42.02	29.28
rs-16	35.16	13.59	1.29	40.78	24.11	29.63

Table 3. New Score calculation for rank swapping, with parameters $p = 2, 8, 16$

rs-p	IL	RSLD	ID	DR_Max	Score_Max
rs-2	3.89	77.73	93.98	85.85	44.87
rs-8	16.54	41.28	62.11	51.69	34.12
rs-16	35.16	13.81	40.78	27.29	31.22

However, in [15], a new record linkage technique was introduced, specifically designed for rank swapping. This technique always finds more correct links between original and protected records than the two generic (probabilistic and distance-based) techniques. In other words, if we denote as RSLD (for Rank Swapping Linkage Disclosure) risk the average percentage of correct links found by this new technique, we have that RSLD is always greater than DLD and PLD. For this reason, it is natural to think that an intruder trying to break the privacy of a database that has been protected using rank swapping will always use this record linkage technique. Therefore, the real value of the score should be computed by considering only RSLD. In other words, we will have DR_Max= $0.5 * ID + 0.5 * RSLD$, in this case. Table 3 shows the values of RSLD, DR_Max and Score_Max for the considered parameterizations of rank swapping, applied to the Census dataset.

The result is a significant increase of the score values for rank swapping, specially for those parameterizations (in particular, for $p = 8$) which obtained best (i.e. lower) scores in the original computation. In general, we believe that many positions of the original ranking of protection methods, obtained with the first implementation of the score in [7], would be modified if our ideas were used to compute a new version of the score, leading to a new and more realistic ranking of dataset protection methods.

5.3 Different Combinations of i Attributes

In this section we compare and study the consequences of the different Score calculations. Firstly, we have protected the Census dataset with different instances of the MDAV microaggregation algorithm. In particular, we have split the Census dataset in four blocks of three attributes: $((v1, v2, v3)$, $(v4, v5, v6)$, $(v7, v8, v9)$, $(v10, v11, v12))$, then, we have applied the MDAV algorithm to each block with $k = 5, 15, 25$.

In Table 4 we show the original Score values for the different MDAV instances, note that the disclosure risk values (*i.e.* DLD and PLD) are computed in the classical way. As we have said before, it is very inefficient to compute all $\binom{V}{t}$ possible combinations of attributes available for an intruder, for $V = 12$ and $t = 1, \ldots, 12$. For this reason we also present in Table 5 the simplified Score defined in [18], the unique difference of this simplified score with regard to the classical one is that PLD is not computed, because the probabilistic record linkage algorithm is very costly. Therefore, the disclosure risk is calculated as $DR = 0.5 \cdot DLD + 0.5 \cdot ID$. Using this simplified score we can efficiently compare

Table 4. Original Score of different MDAV parameterizations

k	IL	DLD	PLD	ID	DR	Score
5	7.64	27.06	34.53	85.52	58.16	32.90
15	9.99	16.00	22.67	79.63	49.48	29.74
25	11.12	11.49	18.32	77.63	46.27	28.69

Table 5. Simplified original Score of different MDAV parameterizations

k	IL	DLD	ID	DR	Score
5	7.64	27.06	85.52	56.29	31.97
15	9.99	16.00	79.63	47.82	28.90
25	11.12	11.49	77.63	44.56	27.84

different ways to compute the disclosure risk, assuming that the intruder knows other combinations of attributes.

In Table 6 we show the disclosure risk values and the changes in the final Score assuming that the intruder knows the best possible combination of attributes in each p_t calculation. As we can observe comparing the differences between the DLD columns in tables 5 and 6, the DLD value is four times larger in the new computation than in the traditional one. This increment is only produced assuming that the set of attributes known by the intruder is the most favorable for his interests.

Tables 7 and 8 show the DLD and Score values obtained by computing the average and the median of all possible combinations of t attributes known by the intruder. The final value of DLD is computed as the average of these values, for all $t = 1, \ldots, 12$. Again, in both cases the disclosure risk is larger than in the classical Score calculation. Obviously, the differences are not so significant compared with the maximum Score; however, such differences indicate that the traditional score computation underestimates the real disclosure risk of microaggregation.

For simplicity of the experiments, we have considered again only DLD in this case, because the distance-based algorithm for record linkage is very efficient, and so we have been able to run it over all the $\binom{12}{t}$ combinations of attributes, for $t = 1, \ldots, 12$. But in this particular case of MDAV, the best record linkage technique is the probabilistic one, as shown in Table 4. Similar (but very more costly) experiments could/should be run to compute the average or maximum percentage of correct links found by using Probabilistic Record Linkage, for every $t = 1, \ldots, 12$. This would give, in this case, the real value of LDR.

Table 6. Simplified maximum score of different MDAV parameterization selecting the largest disclosure risk variable selection

k	IL	DLD	ID	DR	Score
5	7.64	70.24	85.52	77.88	42.76
15	9.99	51.63	79.63	65.63	37.81
25	11.12	41.83	77.63	59.73	35.42

Table 7. Simplified score of different MDAV parameterizations selecting the average disclosure risk variable selection

k	IL	DLD	ID	DR	Score
5	7.64	34.00	85.52	59.76	33.70
15	9.99	20.30	79.63	49.96	29.98
25	11.12	14.79	77.63	46.21	28.66

Table 8. Simplified score of different MDAV parameterizations selecting the median disclosure risk variable selection

k	IL	DLD	ID	DR	Score
5	7.64	31.26	85.52	58.39	33.02
15	9.99	18.04	79.63	48.84	29.41
25	11.12	13.50	77.63	45.57	28.34

6 Conclusions

In this paper we have revisited the original implementation to compute the score of dataset protection methods [7]. In particular, we have argued that some details related to the way in which the 'real' Linkage Disclosure risk of a method is computed deserve more discussion. We propose to modify some parts of the original definition. The proposed modifications are very slight, but they can lead to significant changes in the resulting scores of different methods. In our opinion, the resulting ranking of protection methods will be more realistic and fair.

Whereas one of our modifications (how to select the t attributes available to the intruder, when re-identifying) closes this problem very quickly, the other proposed modification (considering only the most effective record linkage technique) has more implications. For example, one should always consider the possibility that specific record linkage techniques are designed for each particular protection method, leading to a more effective re-identification than the generic (distance-based or probabilistic) techniques. For this reason, it is important to find (or better, to discard the existence of) such specifically designed record linkage techniques, before trying to conclude a more or less realistic value for the Disclosure Risk of each dataset protection method.

Acknowledgements

Partial support by the Spanish MEC (projects ARES – CONSOLIDER INGENIO 2010 CSD2007-00004 – and eAEGIS – TSI2007-65406-C03-02), the Generalitat de Catalunya (grant 2005-SGR-00093). Jordi Nin wants to thank the Spanish Council for Scientific Research (CSIC) for his I3P grant.

References

1. Adam, N.R., Wortmann, J.C.: Security-control for statistical databases: a comparative study. ACM Computing Surveys 21, 515–556 (1989)
2. Dalenius, T., Reiss, S.P.: Data-swapping: a technique for disclosure control. Journal of Statistical Planning and Inference 6, 73–85 (1982)
3. Data Extraction System, U.S. Census Bureau, http://www.census.gov/
4. Defays, D., Anwar, M.N.: Micro-aggregation: A Generic Method. In: Proceedings of the 94 International Seminar on Statistical Confidentiality, Luxembourg, Office for Official Publications of the European Communities (1995)
5. Domingo-Ferrer, J., Mateo-Sanz, J., Torra, V.: Comparing SDC methods for microdata on the basis of information loss and disclosure risk. In: Pre-proceedings of ETK-NTTS 2001, vol. 2, pp. 807–812. Eurostat, Luxembourg (2001)
6. Domingo-Ferrer, J., Torra, V.: Disclosure control methods and information loss for microdata [10], pp. 91–110 (2001)
7. Domingo-Ferrer, J., Torra, V.: A quantitative comparison of disclosure control methods for microdata [10], pp. 111–133 (2001)
8. Domingo-Ferrer, J., Torra, V., Mateo-Sanz, J.M., Sebé, F.: Systematic measures of re-identification risk based on the probabilistic links of the partially synthetic data back to the original microdata. Technical report (2005)
9. Domingo-Ferrer, J., Martínez-Ballesté, A., Mateo-Sanz, J.M., Sebé, F.: Efficient multivariate data-oriented microaggregation. The VLDB Journal 15, 355–369 (2006)
10. Doyle, P., Lane, J., Theeuwes, J., Zayatz, L. (eds.): Confidentiality, disclosure, and data access: theory and practical applications for statistical agencies. Elsevier Science, Amsterdam (2001)
11. Hansen, S., Mukherjee, S.: A Polynomial Algorithm for Optimal Univariate Microaggregation. Trans. on Kwnoledge and Data Engineering 15(4), 1043–1044 (2003)
12. Hundepool, A., Van de Wetering, A., Ramaswamy, R., Franconi, L., Capobianchi, A., DeWolf, P.-P., Domingo-Ferrer, J., Torra, V., Brand, R., Giessing, S.: μ-ARGUS version 3.2 Software and User's Manual. Statistics Netherlands, Voorburg NL (February 2003)
13. Jaro, M.A.: Advances in Record Linkage Methodology as Applied to Matching the 1985 Census of Tampa, Florida. Journal of the American Statistical Society 84(406), 414–420 (1989)
14. Moore, R.A.: Controlled data-swapping techniques for masking public use microdata sets. Statistical Research Division Report Series, RR96-04, U.S. Bureau of the Census (1996)
15. Nin, J., Herranz, J., Torra, V.: Rethinking rank swapping to decrease disclosure risk. Data & Knowledge Engineering 64(1), 346–364 (2008)
16. Nin, J., Herranz, J., Torra, V.: How to Group Attributes in Multivariate Microaggregation. International Journal on Uncertainty, Fuzziness and Knowledge-Based Systems 16(1), 121–138 (2008)
17. Pagliuca, D., Seri, G.: Some results of individual ranking method on the system of enterprise accounts annual survey, Esprit SDC Project, Deliverable MI-3/D2 (1999)
18. Sebé, F., Domingo-Ferrer, J., Mateo-Sanz, J.M., Torra, V.: Post-Masking Optimization of the Tradeoff between Information Loss and Disclosure Risk in Masked Microdata Sets. In: Domingo-Ferrer, J. (ed.) Inference Control in Statistical Databases. LNCS, vol. 2316, pp. 187–196. Springer, Heidelberg (2002)

19. Torra, V., Abowd, J.M., Domingo-Ferrer, J.: Using Mahalanobis distance-based record linkage for disclosure risk assessment. In: Domingo-Ferrer, J., Franconi, L. (eds.) PSD 2006. LNCS, vol. 4302, pp. 233–242. Springer, Heidelberg (2006)
20. Willenborg, L., Waal, T.: Elements of Statistical Diclosure Control. Lecture Notes in Statistics. Springer, Heidelberg (2001)
21. Winkler, W.E.: Re-identification methods for masked microdata. In: Domingo-Ferrer, J., Torra, V. (eds.) PSD 2004. LNCS, vol. 3050, pp. 216–230. Springer, Heidelberg (2004)
22. Yancey, W.E., Winkler, W.E., Creecy, R.H.: Disclosure risk assessment in perturbative microdata protection. In: Domingo-Ferrer, J. (ed.) Inference Control in Statistical Databases. LNCS, vol. 2316, pp. 135–152. Springer, Heidelberg (2002)

Assessing Disclosure Risk for Record Linkage

Chris Skinner

Southampton Statistical Sciences Research Institute
University of Southampton
Southampton SO17 1EF, United Kingdom

Abstract. An intruder seeks to match a microdata file to an external file using a record linkage technique. The identification risk is defined as the probability that a match is correct. The nature of this probability and its estimation is explored. Some connections are made to the literature on disclosure risk based on the notion of population uniqueness.

Keywords: identification; log-linear model; match; misclassification; uniqueness.

1 Introduction

Statistical agencies are obliged to protect confidentiality when they release outputs. One potential threat to confidentiality is the use of record linkage methods [1, 2, 3]. The concern is that an 'intruder' might link an element of an agency's output to a known individual (or other unit) in some external data source and, if the link is correct, succeed in identifying an individual who provided data upon which the output is based. Such identification (identity disclosure) might lead to the disclosure of further information about this individual.

This threat is most natural to consider when the output consists of a microdata file. In this paper we suppose the agency releases a file containing records for a sample of units, with each record containing the values of various variables. These values may have been masked by statistical disclosure control (SDC) methods, although we suppose there remains a one to one correspondence between the records and the units which provided the data. Thus, identification of these units could, in principle, occur via record linkage to an external file of known units. We suppose that linkage takes place by matching the values of a subset of the variables, 'key variables', shared between the microdata and the external file.

The main aim of this paper is to consider approaches to measuring and estimating the risk of identity disclosure in this setting. A secondary aim is to link this work with other approaches in the literature to assessing identification risk which have centred on concerns about the existence of 'population uniques', i.e. records which are unique in the population with respect to their values of the key variables.

Possibly the earliest contribution to assessing the identification risk arising from record linkage is by Spruill in [4]. She considers linkage methods which match by minimizing a distance measure and combines the definition of risk with the method for assessing it. The approach is based upon a re-identification experiment where each record in a microdata file, which has been masked by an SDC method, is matched to

J. Domingo-Ferrer and Y. Saygın (Eds.): PSD 2008, LNCS 5262, pp. 166–176, 2008.

the original unmasked file and the closest record in the latter file selected. The risk is defined essentially as the proportion of such matches which are correct. She also notes that account might be taken of ‘near matches’. This broad approach has been adopted or discussed in much subsequent literature, e.g. [1, 5, 6, 7].

There are, however, some problems with using the empirical proportion of correct matches as a measure of risk. First, the original unmasked file is acting as a surrogate for an external file held by the intruder in such approaches. The use of this file represents a highly conservative approach to risk assessment since it ignores the protective effect of sampling and, even if there are some common units in the microdata and external files, the values of the variables for these units in the two files are likely to differ for many practical reasons e.g. differences in measurement. To address this concern, the original unmasked file might be replaced by an alternative surrogate external file constructed by the agency. For example, it is reported in [8] that the US National Center for Education Statistics uses certain commercially available school files. Agencies may also consider using other datasets which they collect (from other surveys) or constructing synthetic files from the original unmasked file which take account of sampling and measurement error.

A second more conceptual problem with this approach is that it can fail to reflect adequately the information available to the intruder. Suppose, for example, that the overall proportion of correct matches is 5% and that the agency considers this sufficiently low. Suppose, however, that the intruder could determine which 5% of his claimed matches are correct and which 95% are incorrect. Then the intruder could claim some matches with 100% confidence and this might be deemed an unacceptable disclosure risk. On the other hand, suppose the agency chooses to calculate its proportions separately according to different areas and observes that the proportions vary across areas from 0% to 70%. It might deem the release of data for those areas with proportions as high as 70% as unacceptable. However, if the intruder could only determine that the overall rate of a correct match was 5% (in practice, the intruder will have difficulty determining the proportion of correct matches since it requires knowledge of the true identities of the records in the microdata, information unavailable to the intruder) and was unable to identify areas where it was higher, the agency’s judgment would be over-conservative.

In this paper we suppose that it is necessary for the intruder to have evidence that the link is ‘likely’ to be correct. Identification risk is defined as the probability that a match is correct, conditional on data assumed available to the intruder, c.f. [9, 10], and it is required that this probability can be estimated reliably from these data. We suppose that the agency might use empirical proportions of correct matches as a means of validating these estimates but not as a direct means of estimation.

We focus in this paper on probabilistic record linkage methods (based on the approach of Fellegi and Sunter in [11] (hereafter referred to as FS) rather than methods based on distance measures. These probabilistic methods are most naturally adapted to assess the probability of a correct match. Indeed, part of conventional record linkage methodology is the estimation of false match rates and one might, as a first approach, take one minus the estimated false match rate as a measure of identification risk. However, in conventional applications of record linkage, incorrect matches (false positives or false negatives) are only of interest because of their statistical consequences for samples as a whole. FS (p. 1196) state that ‘we are not concerned with the *probability* of [these two kinds of erroneous matches]…but rather with the *proportion* of occurrences of these two events in the long run'. In contrast, requirements to protect the confidentiality of every individual imply that an agency may be interested in the probability of a correct match for a single individual.

The paper is organized as follows. First, a framework for the use of record linkage for identification is set out in Section 2. Expressions for the probability of a correct match are obtained in Section 3. After briefly considering issues relating to key variables in Section 4, the estimation of the probability of a correct match is considered in Section 5.

2 The Use of Record Linkage to Achieve Identification

Consider a survey microdata file containing records for a sample of responding units s_1 drawn from a finite population P. Each record will include variables needed by genuine users of the file, but is supposed not to include directly identifying variables like name and address. Suppose an intruder has access to this file and wishes to identify one or more units in s_1. The intruder matches the file to an external file of records for another sample of units $s_2 \subset P$, for which the identities are known and for which it is feasible that the intersection $s_{12} = s_1 \cap s_2$ is non-empty. (We assume that the definition of the population P is public and that the intruder can thus remove any records in the external file which do not belong to P – hence we do not need to allow for s_1 and s_2 to be drawn from different populations.)

Suppose matching is based upon the values of variables, which appear in both files: the *key variables* [12]. Let $\tilde{X}_a$ denote the value of the vector of key variables for unit a in the microdata ($a \in s_1$) and X_b the corresponding value for unit b in the external database ($b \in s_2$). The difference in notation between $\tilde{X}$ and X allows for the possibility that the variables are recorded in a different way in the two data sources. This difference might arise from various reasons, including measurement error (in either source) or the application of a perturbative SDC method to the microdata file. Following FS, suppose the intruder undertakes linkage by calculating a comparison vector $\gamma(\tilde{X}_a, X_b)$ for pairs of units $(a,b) \in s_1 \times s_2$, where the function $\gamma(.,.)$ takes values in some finite comparison space Γ.

Example 1: Exact Matching on Categorical Key Variables

Suppose $\tilde{X}$ and X take only K possible values, denoted $\{1,...,K\}$ without loss of generality. Let $\Gamma = \{1,2,...,K+1\}$ and define the comparison vector by $\gamma(\tilde{X}, X) = j$ if $\tilde{X} = X = j$, $j = 1,2,...,K$, $\gamma(\tilde{X}, X) = K+1$ otherwise. In this case, an intruder might consider any pair $(a,b) \in s_1 \times s_2$ for which $\gamma(\tilde{X}_a, X_b) \le K$ as a potential match, but rule out of consideration any pair for which $\gamma(\tilde{X}_a, X_b) = K+1$.

Suppose the intruder seeks to use the comparison vectors to identify one or more pairs $(a,b)\in s_1\times s_2$ which contain identical units, i.e. are of the form (a,a) where $a\in s_{12}$. Since the number of pairs in $s_1\times s_2$ may be very large, the intruder may only consider pairs which fall in a set $\tilde{s}\subset s_1\times s_2$. Partition $\tilde{s}$ into $M=\{(a,b)\in\tilde{s}\mid a=b, a\in s_{12}\}$, the pairs of common units, and $U=\{(a,b)\in\tilde{s}\mid a\in s_1, b\in s_2, a\neq b\}$, the pairs of different units. The problem faced by the intruder is how to use comparison vector values to classify pairs from $\tilde{s}$ into M or U. An optimum strategy is shown by FS to be based upon a comparison of the probability distributions of the comparison vector between M and U, i.e. a comparison of

$$m(\gamma)=\Pr[\gamma(\tilde{X}_a,X_b)=\gamma\mid(a,b)\in M]\ , \qquad (1)$$

and $$u(\gamma)=\Pr[\gamma(\tilde{X}_a,X_b)=\gamma\mid(a,b)\in U]\ , \qquad \gamma\in\Gamma . \qquad (2)$$

We discuss the nature of these probabilities in the next section. FS show that an optimal approach for the intruder is to order pairs in $\tilde{s}$ according to the likelihood ratios $m(\gamma)/u(\gamma)$, treating pairs with higher values of this ratio as more likely to belong to M. Our aim is to explore the probability of a correct match for pairs selected in this way.

3 The Probability of a Correct Match

Given a pair (a,b), linked using its value of the comparison vector as described after (1) and (2), the probability that the pair represents a correct match, that is $a=b$, may be defined as $p_{M|\gamma}=\Pr[(a,b)\in M\mid\gamma(\tilde{X}_a,X_b)]$, i.e. the conditional probability that the pair is in M given that it is in $\tilde{s}$ and that the comparison vector takes the value γ. To express $p_{M|\gamma}$ in terms of $m(\gamma)$ and $u(\gamma)$, let:

$$p=\Pr[(a,b)\in M]\ , \qquad (3)$$

be the probability that the pair is in M given that it is in $\tilde{s}$ and, using Bayes theorem, we obtain

$$p_{M|\gamma}=m(\gamma)p/[m(\gamma)p+u(\gamma)(1-p)]\ . \qquad (4)$$

Sorting pairs according to this ‘posterior’ probability is equivalent to sorting according to the likelihood ratio $m(\gamma)/u(\gamma)$. From the SDC perspective, expression (4) may be interpreted as the identification risk for a pair (a,b), i.e. the probability that a and b are identical, given the value of the comparison vector. From the record linkage perspective, expression (4) is the probability of a correct match or alternatively one minus the probability of a false match [13].

Expressions (1), (2) and (3) are, of course, dependent on the way the probabilities are defined. Our basic approach in this paper is to suppose that the probabilities are defined with respect to the following three processes:

(i) a random selection (with equal probability) of the pair (a,b) from $\tilde{s} = M \cup U$;

(ii) a random process of generating $\tilde{X}_a$;

(iii) a specified probability design for the selection of s_1 from P ;

where the population P and the values X_a for units in the population are treated as fixed. Evaluating the probabilities over (i), holding s_1 and the $\tilde{X}_a$ fixed, we may write

$$m(\gamma) = E[n_{M\gamma} / n_M] \ , \ u(\gamma) = E[n_{U\gamma} / n_U] \ , \tag{5}$$

where n_M and n_U are the numbers of pairs in M and U respectively, $n_{M\gamma}$ and $n_{U\gamma}$ are the corresponding numbers of these pairs for which the comparison vector takes the value γ and the expectation is with respect to (ii) and (iii). We may thus interpret $m(\gamma)$ and $u(\gamma)$ as the expected relative frequencies of the different comparison vectors within M and U respectively. Similarly, we may write

$$p = E(n_M / \tilde{n}) , \tag{6}$$

where $\tilde{n}$ is the number of pairs in $\tilde{s}$ and the expectation is with respect to (iii). To explore the form of $p_{M|\gamma}$ further under (i), (ii) and (iii), consider two special cases.

Example 1(continued) Exact matching with no misclassification

Suppose exact matching is used as defined earlier and that: $\tilde{X}_a = X_a$ for all units $a \in P$ (i.e. no misclassification); $s_2 = P$ and $\tilde{s} = s_1 \times s_2$. Let $n_1 = | s_1 |$ and $N = | P |$. Noting that $n_M = n_1$ and $\tilde{n} = n_1 N$, we obtain from (5) and (6):

$$m(j) = E[f_j / n_1] \ , \qquad u(j) = E\left(\frac{f_j (F_j \text{-}1)}{n_1 (N-1)} \right) , \qquad j = 1,\ldots,K$$

$$p = E[n_1 / (n_1 N)] = 1/N \ , \tag{7}$$

where f_j and F_j are the numbers of units with $X_a = j$ in s_1 and P respectively. Using Bayes theorem we obtain:

$$\Pr[(a,b) \in M \mid \gamma(\tilde{X}_a, X_b) = j] = 1/F_j \ . \tag{8}$$

This result if free of any assumptions about the sampling scheme. Expression (8) is familiar in the disclosure risk literature, e.g. [14]. It is common to argue, however, that agencies should design release strategies so that an intruder could not know the value of F_j from external information [10]. Note that, in particular, this requires assuming that $s_2 \neq P$. Otherwise, the intruder could determine F_j from knowledge of X_a for $a \in P$. If F_j is unknown to the intruder, the uncertainty about F_j needs to be integrated out of the expression for the identification risk, subject to conditioning on the information available to the intruder. This integration is most naturally done by revising the probability mechanisms (i)-(iii) above to include a process which generates the values X_a for units in the population. Under this extended probability mechanism, the identification risk becomes $E(1/F_j \mid data)$, where *data* represents the data available to the intruder. We shall return to this issue in Section 5. First, we extend the result in (8) to the case when $\tilde{X}_a$ may be derived from X_a by a process of misclassification and s_2 may be any proper subset of P .

Example 1 (continued) Exact matching with misclassification

Suppose again that exact matching is used and that $\tilde{s} = s_1 \times s_2$. We now allow s_2 to be any proper subset of P and suppose that each $\tilde{X}_a$ is determined from X_a as follows

$$\Pr(\tilde{X}_a = j \mid X_a = k) = \theta_{jk} \text{ , for all } a \in P \text{ ,} \tag{9}$$

where θ_{jk} is an element of a misclassification matrix with columns which sum to 1. We now obtain

$$m(j) = E[f_j^{12} / n_{12}] \text{ , } u(j) = E\left(\frac{\tilde{f}_j f_j - f_j^{12}}{n_1 n_2 - n_{12}} \right), \qquad j = 1,...,K$$

$$p = E[n_{12} / (n_1 n_2)],$$

where f_j^{12} is the number of units in s_{12} with $X_a = j$ and $\tilde{X}_a = j$, $\tilde{f}_j$ is the number of units in s_1 with $\tilde{X}_a = j$ and f_j is the number of units in s_2 with $X_a = j$. If we suppose that Bernoulli sampling is employed with inclusion probability π we have $n_{12} \doteq n_2 n_1 / N$ so that $p \doteq 1/N$ and $n_1 n_2 - n_{12} \doteq (N-1) n_{12}$. It follows that

$$\Pr[(a,b) \in M \mid \gamma(\tilde{X}_a, X_b) = j] \doteq E\left(\frac{f_j^{12}}{\tilde{f}_j f_j} \right),$$

where the expectation is with respect to both the sampling and the misclassification mechanisms. We have $E(f_j^{12}) = \pi \theta_{jj} f_j$ and $E(\tilde{f}_j) = \pi \tilde{F}_j$, where $\tilde{F}_j$ is the number of units in P with $\tilde{X}_a = j$ (imagining that the misclassification takes place before the sampling). Hence we may write

$$\Pr[(a,b) \in M \mid \gamma(\tilde{X}_a, X_b) = j] \doteq \frac{\theta_{jj}}{\tilde{F}_j} \text{ .} \tag{10}$$

Note that this expression applies for any choice of s_2, which may be selected arbitrarily. The expression in (4) for the probability of a correct match and the special cases in (8) and (10) apply to a pair of records (a,b) with a specific agreement pattern γ. This notion may be extended to apply to a class of pairs, $\hat{M}$, for which the likelihood ratio is above some threshold, say $\hat{M} = \{(a,b) \mid \gamma(\tilde{X}_a, X_b) \in \Gamma_M\}$, where Γ_M is the set of agreement patterns γ for which $m(\gamma)/u(\gamma)$ is above a threshold specified by the intruder as determining which pairs to declare as links.

A key issue for identification risk assessment is how to estimate $p_{M|\gamma}$ and, more specifically, how to estimate $p, m(\gamma)$ and $u(\gamma)$. We discuss this in section 5. Before then, we consider the record linkage approach further.

4 Taking Account of Key Variable Structure

In practice it is usual to base the comparison vector $\gamma(\tilde{X}_a, X_b)$ upon the separate comparisons of C key variables. Letting $\tilde{X} = (\tilde{X}^1,...,\tilde{X}^C)$ and $X = (X^1,...,X^C)$ we write

$$\gamma(\tilde{X}_a, X_b) = [\gamma^1(\tilde{X}_a^1, X_b^1),...,\gamma^C(\tilde{X}_a^C, X_b^C)], \tag{11}$$

where $\gamma^c(\tilde{X}^c, X^c)$ denotes the comparison vector for the c^{th} key variable.

Example 2. Comparison vectors for simple agreements between continuous or categorical key variables, c.f. [15]

Let $\gamma^c(\tilde{X}^c, X^c) = 1$ if $\tilde{X}^c \sim X^c$ and $\gamma^c(\tilde{X}^c, X^c) = 0$, otherwise, $c = 1,2,...,C$, where $\sim$ is a specified agreement relation. Then
$\Gamma = \{(\gamma^1, \gamma^2,...,\gamma^C) \mid \gamma^c = 0,1;\ c = 1,2,...,C\} = \{0,1\}^C$ and $|\Gamma| = 2^C$.

Example 3. Comparison vectors for agreements between categorical key variables

Suppose $\tilde{X}^c$ and X^c are categorical, taking values $j^c = 1,2,...,t^c$, and $\gamma^c(\tilde{X}^c, X^c) = j^c$ if $\tilde{X}^c = X^c = j^c$, $j^c = 1,2,...,t^c$, $\gamma^c(\tilde{X}^c, X^c) = t^c + 1$ otherwise, $c = 1,2,...,C$. Then

$$\Gamma = \{(\gamma^1, \gamma^2,...,\gamma^C) \mid \gamma^c = 1,...,t^c + 1, c = 1,2,...,C\} \text{ and } |\Gamma| = \prod_{c=1}^{C}(t^c + 1).$$

Given the large potential size of Γ when C is at all large, it is common to restrict attention to a subspace Γ^* of Γ. A common approach is to partition the set of possible values of a specified subset of the key variables into blocks (e.g. [16]) so that the intruder only examines pairs for matching for which the values of these key variables fall in the same block. This constraint is typically equivalent to restricting attention to a subspace Γ^* of Γ.

The estimation of $m(\gamma)$ and $u(\gamma)$ is challenging if $|\Gamma|$ is large, as is likely in Examples 2 and 3 if C is at all large. It is therefore common to make simplifying assumptions, in particular, following FS, to treat the C agreement patterns in (11) as independent within M and U, i.e.

$$m(\gamma) = m_1(\gamma^1)m_2(\gamma^2)...m_C(\gamma^C) \text{ and } u(\gamma) = u_1(\gamma^1)u_2(\gamma^2)...u_C(\gamma^C) \tag{12}$$

where $m_c(\gamma^c) = \Pr[\gamma^c(\tilde{X}_a^c, X_b^c) = \gamma^c \mid (a,b) \in M]$ and

$u_c(\gamma^c) = \Pr[\gamma^c(\tilde{X}_a^c, X_b^c) = \gamma^c \mid (a,b) \in U]$, $c = 1,2,...,C$. We refer to this assumption as *independence of agreement patterns***.** In the categorical variable case of Example 3 with misclassification defined as in (9), a sufficient condition for the independence of agreement patterns is that misclassification operates independently, variable by variable, and that the key variables are themselves independent.

5 Estimation

In this section we consider the estimation of the probability of a correct match, $p_{M|\gamma}$, defined in section 3. We assume that the estimator is a function only of data which is available to the intruder and thus rule out the possibility of using a training sample, c.f. [13]. In this case, one approach would be to use a *mixture model*, where $p, m(\gamma)$ and $u(\gamma)$ are treated as unknown parameters in a model for the observed values of the comparison vectors. The model is a mixture of models for M and U, treated as latent classes, and maximum likelihood estimation is used for parameter estimation (e.g. FS Method 2; [15, 17]). This modelling approach has found some success in record linkage applications where very strong identifying information, such as name and address, is available. On the other hand, it has been less successful when the distributions of the comparison vectors for M and U are not well-separated or are not each unimodal [15, 18] and this may be the case in practice in many SDC contexts, e.g. for social survey data. This is a matter for further empirical investigation, however, which we do not attempt in this paper.

Instead, we approach the estimation problem more directly by considering expressions for $p_{M|\gamma}$ in terms of our assumed probability mechanisms, as in section 3, and then considering how to estimate these expressions, from the data available to the intruder as well as possible additional external sources. This approach is analogous to Method 1 of FS. Since $p_{M|\gamma}$ is a function of $p, m(\gamma)$ and $u(\gamma)$, we also discuss the problem of estimating these parameters to gain a better understanding of the general estimation problem. We first return to the two examples in Section 3.

Example : Exact matching with no misclassification

We obtained $p_{M|\gamma} = 1/F_j$ in expression (8) but argued, following this expression, that a more suitable measure will usually be $E(1/F_j \mid data)$. The evaluation of this conditional expectation is discussed in [19] under the assumption that the F_j are generated from a Poisson log-linear model and that the sample frequencies f_j represent the *data*. Treating the pairs (f_j, F_j) as independent, the conditional probability may then be expressed as $E(1/F_j \mid f_j)$ and a closed form expression may be obtained under the Poisson log-linear model and a Bernoulli sampling assumption. The conditional probability will be highest for cases which are unique in the sample, i.e. $f_j = 1$. The conditional probability may be estimated by estimating the log-linear model parameters and plugging these estimates into the expression for the conditional probability.

Example 1: Exact matching with misclassification

We obtained the approximate expression $p_{M|\gamma} \doteq \theta_{jj} / \tilde{F}_j$ in expression (10) . As above, we may argue that in practice $\tilde{F}_j$ will be unknown and a more suitable measure is $\theta_{jj} E(1/\tilde{F}_j \mid \tilde{f}_j)$. The second component of this expression, $E(1/\tilde{F}_j \mid \tilde{f}_j)$, may be estimated by applying the methodology of [19] to the observed microdata. The misclassification probability θ_{jj} might be estimated by making some approximating assumptions and using external evidence on the misclassification process. One assumption may be that some of the key variables are subject to no misclassification, as is commonly assumed for blocking variables, and that misclassification on the remaining variables is not dependent upon the values of such correctly classified variables. A further assumption may be that the remaining key variables are misclassified independently. This may be related to but is not the same as the earlier assumption of independence of agreement patterns. Under the independence of misclassification assumption, θ_{jj} may be expressed as a product of correct classification probabilities for the different key variables. This may need to be modified to allow for the possibility that the values of some key variables are missing.

To better understand the nature of the general estimation problem, now consider the separate estimation of $p, m(\gamma)$ and $u(\gamma)$. Consider p first. If $\tilde{n}$ is large we have from (6) that $p \doteq n_M / \tilde{n}$. The intruder knows the value of $\tilde{n}$ and so needs to estimate n_M in order to estimate p . We know $n_M \leq n_{12}$, where $n_{12} = | s_{12} |$. And if we take the worst case, where the intruder selects $\tilde{s}$ in such a way that it includes all possible common pairs (i.e. all (a,a) where $a \in s_{12}$) then we have $n_M = n_{12}$. Thus, in order to estimate p , it suffices to estimate n_{12} . We *s*uppose the intruder can determine inclusion probabilities $\pi_i = \Pr(i \in s_1)$ for $i \in s_2$. This is plausible. Often inclusion probabilities are equal in social surveys or else they will vary by strata which may be known for units in s_2 . Since we have $n_{12} = E(\sum_{i \in s_2} \pi_i)$, where the expectation is with respect to the sampling scheme for s_1 , the intruder can estimate n_{12} by $\hat{n}_{12} = \sum_{i \in s_2} \pi_i$ and hence estimate p by $\hat{p} = \hat{n}_{12} / \tilde{n}$. Note also that some adjustment will usually be necessary for nonresponse (e.g. by multiplying π_i by a response rate). Often in social surveys the inclusion probabilities π_i will be small and so $\hat{n}_{12}$ is only likely to be to have reasonable relative precision as an estimator if the size of the external database is large, representing a substantial proportion of the population. The extent to which p may be estimated reliably also, of course, depends upon this condition.

Let us now turn to the estimation of $m(\gamma)$ and $u(\gamma)$. Consider Example 1 with misclassification again, where we wish to estimate $m(\gamma)$ and $u(\gamma)$ for $j = 1,...,K$. We may write $m(j) = \theta_{jj} E[n_{12j} / n_{12}]$, where n_{12j} is the number of units in s_{12} with $\gamma = j$. And under Bernoulli (or equal probability) sampling we may write

$E[n_{12j}/n_{12}] = f_j/n_2$, so that $m(j) = \theta_{jj} f_j/n_2$. And to first approximation (Jaro, 1989) we have: $u(j) \doteq (\tilde{f}_j/n_1)(f_j/n_2)$. The right hand side of this expression provides an estimator of $u(j)$ which should be reliable when $\tilde{f}_j$ and f_j are not small. However, in many disclosure problems of interest this will not be the case. In these circumstances, a modelling approach such as using log-linear models [19] or the independence of agreement patterns approach in section 4 seems needed. Note that to estimate $p_{M|\gamma}$ in (4) we only need to estimate the ratio $m(j)/u(j)$, which we may approximate in this case by $m(j)/u(j) = \theta_{jj}/(\tilde{f}_j/n_1)$. The factor f_j/n_2 cancels out and the key unknown required to estimate $m(j)$ is θ_{jj}. We suggest that it will normally not be realistic to expect that the intruder will be able to estimate this parameter reliably from the available data (although the mixture model approach merits further investigation). Thus, we suggest that a more realistic approach is that it is estimated by making some approximating assumptions and using external evidence on the misclassification process, as discussed above.

6 Conclusion

This risk of identification may be defined as the probability of a correct match for attacks where the intruder uses record linkage. It has been shown that expressions for this probability may be obtained for probabilistic record linkage in some special cases. In particular, expressions for the probability in the case of categorical key variables have close connections to those in other literature on disclosure risk, such as [10]. It has also been shown that an intruder may be able to estimate these probabilities reliably under certain assumptions.

References

[1] Domingo-Ferrer, J., Torra, V.: A quantitative comparison of disclosure control methods. In: Doyle, P., Lane, J.I., Theeuwes, J.J.M., Zayatz, L.V. (eds.) Confidentiality, Disclosure and Data Access. North-Holland, Amsterdam (2001)

[2] Domingo-Ferrer, J., Torra, V.: Disclosure risk assessment in statistical microdata protection via advanced record linkage. Statistics and Computing 13, 343–354 (2003)

[3] Fienberg, S.E.: Privacy and confidentiality in an e-commerce world: data mining, data warehousing, matching and disclosure limitation. Statistical Science 21, 143–154 (2006)

[4] Spruill, N.L.: Measures of confidentiality. Proc. Surv. Res. Sect. Amer. Statst. Ass., 260–265 (1982)

[5] Lambert, D.: Measures of disclosure risk and harm. Journal of Official Statistics 9, 313–331 (1993)

[6] Winkler, W.E.: Masking and re-identification methods for public use microdata: overview and research problems. In: Domingo-Ferrer, J., Torra, V. (eds.) PSD 2004. LNCS, vol. 3050, pp. 231–246. Springer, Heidelberg (2004)

[7] Torra, V., Abowd, J.M., Domingo-Ferrer, J.: Using Mahalanobis distance-based record linkage for disclosure risk assessment. In: Domingo-Ferrer, J., Franconi, L. (eds.) PSD 2006. LNCS, vol. 4302, pp. 233–242. Springer, Heidelberg (2006)

[8] Federal Committee on Statistical Methodology Statistical Policy Working Paper 22 (2nd Version): Report on Statistical Disclosure Limitation Methodology, Office of Management and Budget, Washington, D.C (2005)

[9] Reiter, J.: Estimating risks of identification disclosure in microdata. Journal of the American Statistical Association 100, 1103–1112 (2005)

[10] Skinner, C.J.: The probability of identification: applying ideas from forensic science to disclosure risk assessment. Journal of the Royal Statistical Society, Series A 170, 195–212 (2007)

[11] Fellegi, I.P., Sunter, A.B.: A theory for record linkage. Journal of American Statistical Association 64, 1183–1210 (1969)

[12] Bethlehem, J.G., Keller, W.J., Pannekoek, J.: Disclosure control for microdata. Journal of the American Statistical Association 85, 38–45 (1990)

[13] Belin, T.R., Rubin, D.B.: A method for calibrating false-match rates in record linkage. Journal of American Statistical Association 90, 694–707 (1995)

[14] Duncan, G., Lambert, D.: The risk of disclosure for microdata. Journal of Business and Economic Statistics 7, 207–217 (1989)

[15] Larsen, M.D., Rubin, D.B.: Iterative automated record linkage using mixture models. Journal of American Statistical Association 96, 32–41 (2001)

[16] Jaro, M.A.: Probabilistic linkage of large public health data files. Statistics in Medicine 14, 491–498 (1995)

[17] Jaro, M.A.: Advances in record linkage methodology as applied to matching the 1985 Census of Tampa, Florida. Journal of American Statistical Association 84, 414–420 (1989)

[18] Herzog, T.N., Scheuren, F.J., Winkler, W.E.: Data Quality and Record Linkage Techniques. Springer, New York (2007)

[19] Skinner, C.J., Shlomo, N.: Assessing disclosure risk in survey microdata using log-linear models. Journal of American Statistical Association (to appear, 2008)

Robust Statistics Meets SDC: New Disclosure Risk Measures for Continuous Microdata Masking

Matthias Templ[1,2] and Bernhard Meindl[1]

[1] Department of Methodology, Statistics Austria, Guglgasse 13, 1110 Vienna, Austria
bernhard.meindl@statistik.gv.at

[2] Department of Statistics and Probability Theory, Vienna University of Technology,
Wiedner Hauptstr. 8-10, 1040 Vienna, Austria
templ@statistik.tuwien.ac.at

Abstract. The aim of this study is to evaluate the risk of re-identification related to distance-based disclosure risk measures for numerical variables. First, we overview different - already proposed - disclosure risk measures. Unfortunately, all these measures do not account for outliers. We assume that outliers must be protected more than observations near the center of the data cloud. Therefore, we propose a weighting scheme for each observation based on the concept of robust Mahalanobis distances. We also consider the peculiarities of different protection methods and adapt our measures to be able to give realistic measures for each method. In order to test our proposed distance based disclosure risk measures we run a simulation study with different amounts of data contamination. The results of the simulation study shows the usefulness of the proposed measures and gives deeper insights into how the risk of quantitative data can be measured successfully. All the methods proposed and all the protection methods plus measures used in this paper are implemented in R-package *sdcMicro* which is freely available on the comprehensive R archive network (http://cran.r-project.org).

Keyword: Statistical disclosure control, Distance based disclosure risk, Outlier, Simulation study.

1 Introduction

For many applications the measurement of disclosure risk is based on the idea of uniqueness, rareness, k-anonymity ([1]), base individual risk estimation ([2], [3]) or on certain models ([4]).

However, if the data consists of continuous scaled variables (e.g. business data on enterprises) other definitions of disclosure risk must be considered.

Applying the concept of uniqueness and k-anonymity on these quantitative variables results that every observation in the data set is unique.

If detailed information about a value of a numerical variable is available, one may be able to identify and eventually gain further information about an

J. Domingo-Ferrer and Y. Saygın (Eds.): PSD 2008, LNCS 5262, pp. 177–189, 2008.

individual. So, an attacker may be able to identify statistical units by using for example linking procedures. The anonymization of numerical variables should avoid the successful merging of underlying data with other data sources.

We assume that an intruder has information about a statistical unit which is included in the data and the intruder's information about some values of certain variables overlap with some variables in the data, i.e. we assume that the intruder's information can be merged with the data. In addition to that we assume that the intruder is sure that the link to the data is correct, except for microaggregated data. In this case the intruder can never be sure because at least k observations have the same value for each numerical variable.

In the next part of the introduction we will give a short overview of some popular distance based disclosure risk measures. In the next section we will describe the need of a special treatment of outliers that exist in almost every data set in Official Statistics. We then propose new measures of disclosure risk which give more realistic results when applied to data which include outliers. Finally, we compare all the measures considered in this study with a practical real data example as well as within a large simulation study.

All our proposed measures have been included in R-package *sdcMicro* (see e.g. in [5], [6]).

1.1 Distance Based Disclosure Risk Measures

By using distance based record linkage methods one tries to find the nearest neighbours between observations from two data sets. [7] has shown that these methods outperform probabilistic methods. Such probabilistic methods are often based on the EM-algorithm which is highly influenced by outliers.

Another approach based on cluster analysis is provided by [8] who uses k-means clustering with a high amount of clusters on mixed scaled variables. However, there are much better clustering methods available (see e.g. [9]). k-means should not be applied on mixed scaled variables and, to put it crudely, this approach works as the usual distance based record linkage because it is based on the idea of similar objects and distance metrics.

Another type of measures of disclosure risk - referred to as value disclosure risk - is extensively used, e.g. by [10]. The main goal is to evaluate the gain in explanation of parameters or variables when releasing perturbed data.

[11] uses distance based record linkage and interval disclosure. In the first approach they search for the nearest neighbour from each observation of the masked data value to the original data points. Then they sign those observations for which the nearest neighbor is the corresponding original value. In their second approach they check if the original value falls within an interval centered on the masked value. Then they calculate the length of the intervals based on the standard deviation of the variable (method *SDID*).

In addition to that they define a rank-based interval procedure which is similar to the idea of *rank swapping* (method *RID*). For each variable of the masked data set they define a rank-based interval around each value. The rank-based interval includes p-percent of the total number of observations of the ranked variable.

The proportion of the original values which fall into the calculated interval is used as measure of disclosure risk.

The calculation of an interval is based on a vector k of length p, the dimension of the confidential variables. k indicates how large these intervals for each variable are. In the implementation of *sdcMicro* the elements of k are set to 0.01 by default.

2 Special Treatment of Outliers for Disclosure Risk

Almost all data sets from Official Statistics consists of statistical units whose values in at least one variable are quite different from the main part of the observations. This leads to the fact that these variables are very asymmetric distributed. Such outliers might be enterprises with a very large value for turnover, for example, or persons with extremely high income or even multivariate outliers.

Unfortunately, an intruder may have a big interest in the disclosure of a large enterprise or of an enterprise which has specific characteristics. Since enterprises are often sampled with certainty or have a sampling weight near to 1 the intruder can be very confident that the enterprise he wants to disclose is definitely in the sample. In contrast to that an intruder may not be as interested to disclose statistical units which have the same behaviour than the main part of the observations. For these reasons it is reasonable to define measures of disclosure risk that take the "outlyingness" of an observation into account.

Therefore we assume that outliers should be much more perturbed than non-outliers because they are easier to re-identify even when the distance from the masked observation to its original observation is relatively large.

2.1 "Robustification" of *SDID*

In a first step we robustify method *SDID* because it is obvious that outliers increase the intervals estimated with method *SDID* dramatically since the calculation of the classical standard deviation is based on squared distances between the observations and the arithmetic mean.

However, method *SDID* can be easily robustified by using a robust measure for the standard deviation. We propose to use the MAD instead of the classical standard deviation. The MAD is given by

$$\mathrm{MAD} = 1.4826 * median(|x_i - \tilde{x}|) \quad ,$$

with $\tilde{x}$ being the median and the constant $= 1.4826$ ensures consistency. We will call this robustified method *RSDID*.

3 New Measures of Disclosure Risk

All disclosure risk intervals obtained from methods *SDID*, *RID*, from the methods based on cluster analysis as well as *RSDID* do not depend on the scale of the actual value and therefore, the length of the interval is equal for non-outlying and outlying values. Thus, we now propose new, more realistic measures of disclosure risk which account for the "outlyingness" of each observation.

[11] searches for outliers (after z-transformation of the variables) by calculating Euclidean distances for each observation with respect to the origin. Finally the observations are sorted based on these distances and the five% farthest observations are classified as outliers. Nevertheless, this approach is quite poor for the detection of outliers because of using a non-robust transformation as well as using Euclidean distances in a multivariate space.

The aim is now to measure the distance of each observation to the center of the data in a multivariate space. For a p-dimensional multivariate sample x_i $(i = 1, \ldots, n)$ the Mahalanobis distance is defined as

$$\mathrm{MD}_i = (x_i - t)^T C^{-1} (x_i - t) \quad \text{for} \quad i = 1, \ldots, n \quad , \tag{1}$$

where t is the estimated multivariate location and C the estimated covariance matrix. Usually, t is the multivariate arithmetic mean, and C is the sample covariance matrix.

Multivariate outliers may simply be defined as observations featuring large (squared) Mahalanobis distances. However, this approach has several shortcomings which are visualized in Figure 1. The concept of classical Mahalanobis distances fails completely in this example and does not describe the behaviour neither of the outliers nor of the homogeneous part of the data well. Single extreme observations as well as groups of observations that depart from the main data structure can have a severe influence on this distance measure because both location and covariance are usually estimated in a non-robust manner.

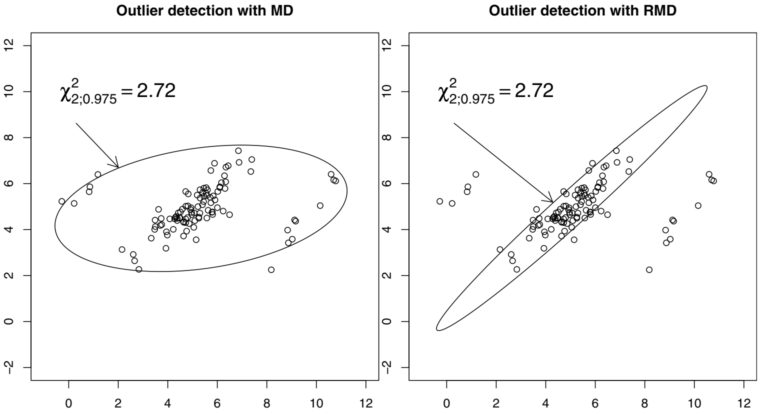


Fig. 1. Illustration of the concept of Mahalanobis distances and robust Mahalanobis distances on a simply 2-dimensional example. LEFT: Tolerance ellipse (95 %) and "outlier detection" using Mahalanobis distances. RIGHT: Tollerance ellipse (95 %) and outlier detection using robust Mahalanobis distances.

The fast minimum covariance determinant (MCD) estimator ([12]) is well known in the literature and has been used to estimate the location and the covariance structure in a robust way. Using this estimator in formula 1 leads to robust Mahalanobis distances (RMD). Figure 1 shows that Mahalanobis distances based on classical measures are not suited for the definition of disclosure risk intervals and that the robust version needs to be chosen. Observations whose RMD_i is greater than $\chi^2_{(0.975,p)}$ may be defined as outliers. This is however only an approximation since squared RMD are only approximately χ^2 distributed (see e.g. in [13]).

The intervals for each data value should now depend on the robust distances, i.e. the intervals may be defined as $k_j \times$ $(\text{RMD}_i)^{1/2}$, $j \in \{1, \ldots, p\}$. Following this approach we obtain a disclosure risk for each observation by checking if any value of an observation falls into the corresponding interval or not. We then calculate the percentage of observations featuring high risk and call this procedure *RMDID1*.

Since we want to consider the "outlyingness" of each observation we simply weight each observation with its RMD or with $(\text{RMD}_i)^{1/2} \cdot k_j$. In this paper we use the latter weight and call this new procedure *RMDID1w*.

The need for an additional approach can simply be met by using a microaggregation procedure for the perturbation of microdata. We assume that we have applied microaggregation with high aggregation level, e.g. 10. All the methods described previously provide a high risk of disclosure if the original value and the microaggregated value are close to each other. But these measures are unrealistic for this simple microaggregation example since 10 observations possess the same value in the microaggregated variable, and data intruder can never be sure which one is the correct link. Especially, if this observation is near the center of the data cloud the previous measures fail to provide a meaningful measure of disclosure risk.

These problems are solved by looking closely at observation that have a relatively high risk of re-identification in *RMDID1* or *RMDID1w*. An observation which is marked as unsafe (with method *RMDID1* or *RMDID1w*) is considered safe if m observations are very close to the masked observation (we call this procedure *RMDID2*). This problem is illustrated in Figure 7 with a simple 2-dimensional example data set that is described below.

We now describe the proposed algorithm as follows:

1. Robust Mahalanobis distances are estimated in order to get a robust multivariate distance for each observation.
2. Intervals are estimated for each observation around every data point of the original data points where the length of the intervals are defined/weighted by squared robust Mahalanobis distances and the parameter k_j. The higher the RMD of an observation the larger the corresponding intervals.
3. Check if the corresponding masked values fall into the intervals around the original values or not. If the value of the corresponding observation lies within such an interval the entire observation is considered unsafe. We obtain a vector indicating which observations are safe or not (→ we are finished already when using method *RMDID1*).

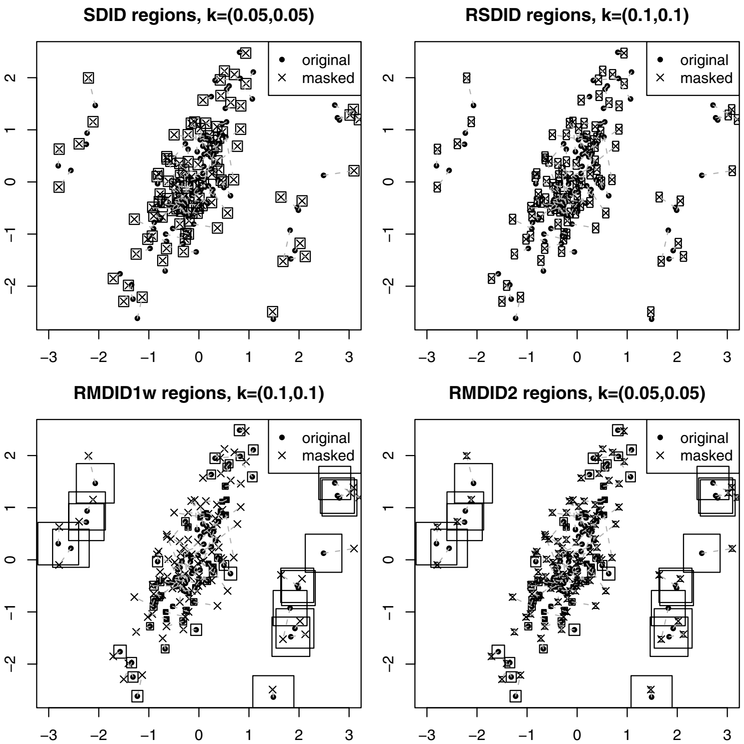


Fig. 2. Original observations and the corresponding masked observations (perturbed by adding additive noise). In the bottom right graphic small additional regions are plotted around the masked values for *RMDID2* procedure.

4. For method *RMDID1w* we calculate the weighted (using RMD) vector of disclosure risk.
5. For method *RMDID2*: whenever an observation is considered unsafe we check if m other observations from the masked data are very close (defined by a parameter $k2$ for the length of the intervals as for *SDID* or *RSDID*) to this observation using Euclidean distances. If more than m points are within these small intervals we conclude that the observation is "safe".

For measures *SDID* and *RSDID* the parameter vector k is a multiplier of the standard deviation. For methods *RMDID1* and *RMDID2* k is a multiplier of the squared RMD. While for standardized data sets the standard deviation is one and the interval around the masked value x_i^{mask} has a length of $2 \cdot k_i$ the RMD weights this interval according to $(\text{RMD}_i)^{1/2}$.

Naturally, most of the intervals corresponding to values in the center of the data are down weighted, and only for those observations which are away of the center of the data cloud the intervals increase.

The second parameter vector $k2$ for method *RMDID2* which evaluates if the masked data has any close neighbours can be set at, for example, 0.05 for each

$i \in 1, \ldots, p$. We look for (automatically) standardized data whose values are within an interval of length $2 \cdot k2_i$ around $x_{i(mask)}$.

Figure 2 points out the idea of weighting the disclosure risk intervals. While for method *SDID* and *RSDID* the rectangular regions around each value are the same as for each observation our proposed methods take the RMD of each observation into account. The difference between the bottom right and the bottom left graphic is that for *RMDID2* rectangular regions around each masked variable are calculated as well. If an observation of the masked variable falls into an interval around the original value it is checked if this observation does have close neighbours, i.e. if the values of m other masked observations are inside a second interval around this masked observation.

While it is not possible to interprete the weighted disclosure risk measures *RMDID1w* and *RMDID2* in a probabilisitic way. However, the proportion of unsafe values on all observations using the unweighted measures can be interpreted as a global, probabilistic measure of disclosure risk.

4 An Example Using the Tarragona Data Set

Please note, when applying the related functions in *sdcMicro* no data standardisation needs to be done since both the center and the scatter of each variable are already considered in our implementation. This means that the standardisation is done automatically.

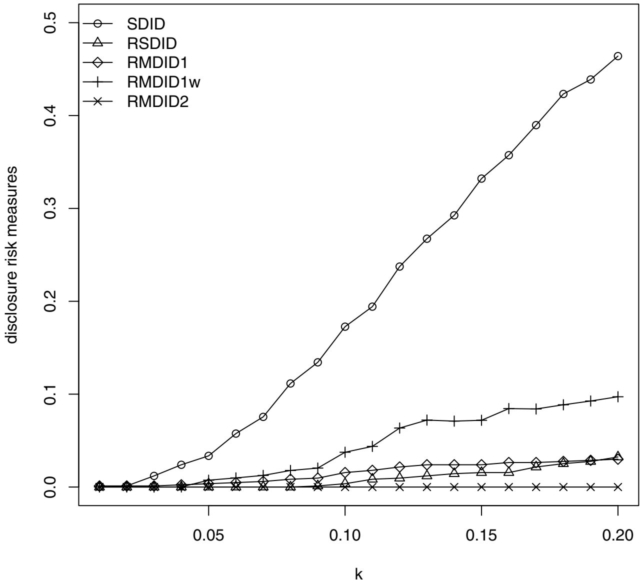


Fig. 3. Influence of parameter k on different distance-based disclosure risk measures

The first result is obtained by using microaggregation method *rmd* from package *sdcMicro*. This algorithm is called RMDM (**R**obust **M**ahalanobis **D**istance based **M**icroaggregation) which was proposed by [5] and has excellent properties (see e.g. the results of the simulation study in [14]). We are interested in how the parameter vector k influences the results of the disclosure risk measures. In this example which results in Figure 3 we set k as a vector as large as the dimension of the Tarragona data set. k varied - with equal values - from 0.01 to 0.2.

One can see that all of the disclosure risk measures account for the increase of the interval length. Naturally, measure *RMDID2* is always zero because microaggregation of the data was conducted and a search for near neighbours was done. Of course, if the parameter m is increased then this measure is quite similar to *RMDID1w*.

5 Simulation Results

Based on the proposed measures the disclosure risk is evaluated in a simulation study with 1000 simulations similar to the approach in [14]. We generate 1000 data sets of dimension 300 × 2 that are multivariate normal distributed with mean vector $\mu = (0, 0)$ and covariance matrix Σ $(\sigma_{ii} = (1, 1), \sigma_{ij} = (0.8, 0.8))$. Furthermore, we include a shifted outlier group with mean $\mu_{out} = c(10, 0)$ and the same variance-covariance structure and calculate all risk measures discussed above for any of the generated data sets.

Please note that we only show medians of the simulation results in order to stay within the limit of pages.

Figure 4 shows the influence of outliers on the *SDID* disclosure risk measure and the deviation of the eigenvalues from the robust estimated covariance matrix between the original and the masked data. However, other measures can be used too (see [15]). This measure of information loss was chosen because the eigenvalues of the covariance matrix may be used to represent the data structure and are input of popular multivariate techniques like (robust) principal component analysis. The robust estimation of the covariance matrix is done using the fast MCD-estimator ([12]). A robust estimation may be prefered since traditional measures may give unrealistic results on inhomogeneous data sets.

It is easy to see that the disclosure risk does not increase when using method *shuffling*. [14] showed that this relates to the meaningless results of *shuffling* when data feature outliers. Increasing the number of outliers in the data results in high amount of information loss also for method *additive noise addition* and *ROMM* (see the description of these methods in [16], [17] and [18]). *Robust shuffling* ([14]) and especially *mdav* (see e.g. [19]) perform better and method *rmd* ([20]) outperforms all the methods.

While the information loss for methods *additive*, *shuffling* and *ROMM* is comparable in Figure 5, the disclosure risk for these methods is relatively low compared to the *shuffling* procedure. But again, all these methods are strongly influenced by outliers and provide worst results regarding information loss. *Robust shuffling* is again performing much better than *shuffling* if outliers are

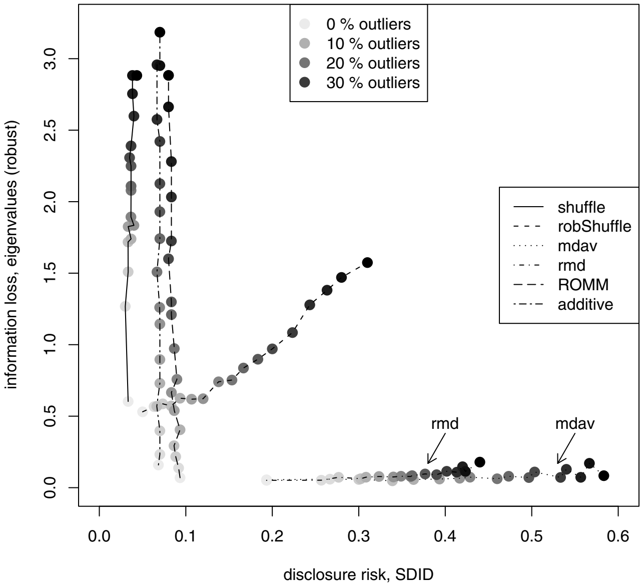


Fig. 4. Effects of shifted outliers (0 till 40 percent in 2.5 percent steps) on some protection methods based on *SDID* measure

included in the data. Again, *rmd* and *mdav* perform best. The "shift" from zero outliers to 2.5 percent outliers that is visible for some methods can be explained because the disclosure risk decreases when the perturbation goes down as soon as outliers appear in the data.

In Figure (6) the *SDID* distance based disclosure risk measure is plotted versus the *RMDID2* measure. It is clearly visible that *SDID* fails and does not account for outliers, i.e. does have the same length of disclosure risk intervals both for non-outliers and outliers. The disclosure risk does not increase, for example, for method *shuffling* since the data structure from the perturbed data is very different from the original one because of the non-robustness of *shuffling* (see also in [14]). In contrast to the other risk measures, *RMDID2* considers the observations as safe when using microaggregation and parameter $m < g$, with g being the aggregation level.

We also want to find out if outlying observations do have a high risk of disclosure. Thus, we divided the data in outliers and non-outliers and visualize the results for the outlier part for every single outlying observation.

Since we evaluate the disclosure risk which is weighted by the RMD for every observation we can simply evaluate which observation possess a high risk of re-identification. For the previous 2-dimensional example we can show which observations are considered unsafe (see Figure 7). The left graphic of Figure 7 shows the unsafe observations discovered by method *SDID*. One can see that some outliers are not considered unsafe although the masked observations is relatively near to the original one. *RMDID2* accounts for this and considers

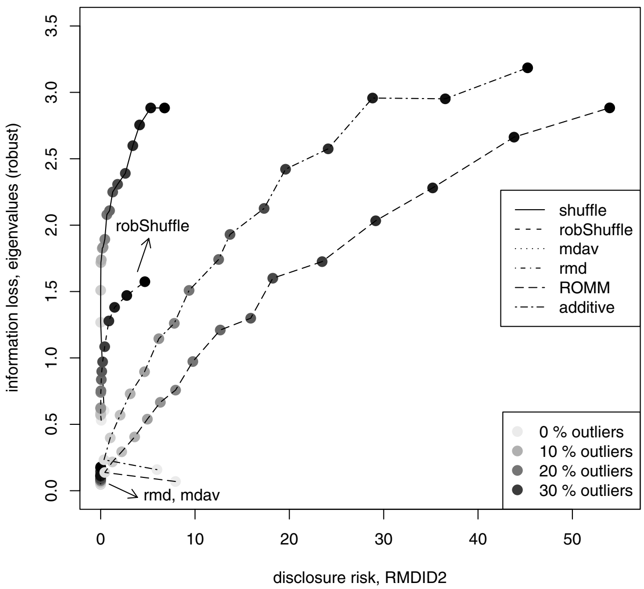


Fig. 5. Effects of outliers on some protection methods based on *RMDID2* measure

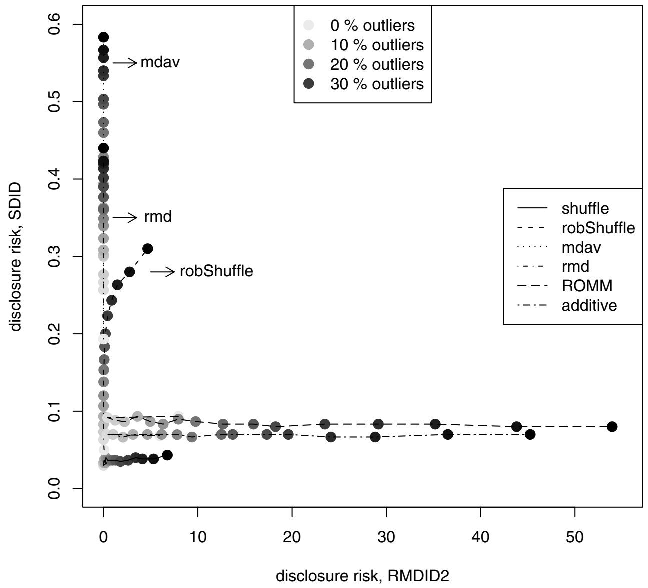


Fig. 6. Comparison of *SDID* and *RMDID2* measure under different contaminations

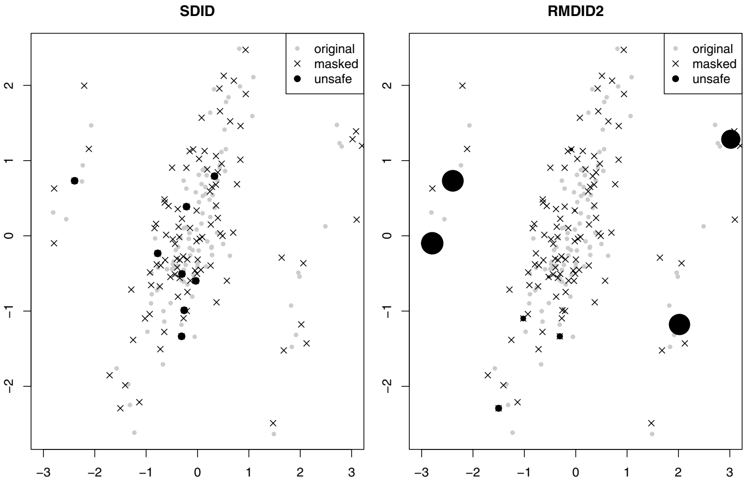


Fig. 7. Disclosure risk evaluated for every observation

some outliers as unsafe as can be seen on the right area of Figure 7. In addition to that each observation does have a different risk. Therefore, one can easily perform additional protection on these observations (e.g. adding noise) as long as the newly proposed measure outlay a risk higher zero or higher a predefined threshold.

6 Conclusion

When considering that data includes outliers (which is the case for virtually any real life data set) we have to tackle two problems considering statistical disclosure control for numerical variables. The first problem is that outliers may disturb the protection methods (or make the generation of adequate synthetic data impossible) with a high loss of information. This problem was partly considered in this study and a deeper insight into this problem was given by [14]. The second problem - the problem which is mainly discussed in this paper - is that outliers must be protected more than the observations which are located near the center of the data cloud, i.e. which are having low robust Mahalanobis distances. We proposed new measures of disclosure risk called *RMDID1*, *RMDID1w* and *RMDID2* that account for these problems and that assign a risk to every observation weighted by the robust Mahalanobis distance of the observation (*RMDID1w* and *RMDID2*). In addition to that we described the problem of microaggregation (but this is also related to all other methods) where an intruder can never be sure which of the aggregated values correspond to the original ones.

The new measures of disclosure risk provide realistic estimations on the risk of re-identification of each observation separately. Therefore, additional protection for high-risk observations may be provided to the masked data resulting in "protected" data with good quality with respect to both high data utility and very low disclosure risk.

All the methods proposed in this paper are freely available on the web and are included in R-package *sdcMicro*.

References

1. Sweeney, L.: k-anonymity: a model for protecting privacy. International Journal on Uncertainty, Fuzziness and Knowledge-based Systems 10(5), 557–570 (2002)
2. Benedetti, R., Franconi, L.: Statistical and technological solutions for controlled data dissemination. In: Pre-Proceedings of New Techniques and Technologies for Statistics, pp. 225–232 (1998)
3. Franconi, L., Polettini, S.: Individual risk estimation in μ-Argus: a review. In: Domingo-Ferrer, J., Torra, V. (eds.) PSD 2004. LNCS, vol. 3050, pp. 262–272. Springer, Heidelberg (2004)
4. Elamir, E., Skinner, C.: Record level measures of disclosure risk for survey microdata. Journal of Official Statistics (submitted, 2006)
5. Templ, M.: sdcMicro: A package for statistical disclosure control in R. In: Bulletin of the International Statistical Institute, 56th Session (2007)
6. Templ, T.: sdcMicro: Statistical Disclosure Control methods for the generation of public- and scientific-use files, R package version 2.4.7 (2008)
7. Domingo-Ferrer, J., Torra, V.: A quantitative comparison of disclosure control methods for microdata. In: Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies, pp. 111–134 (2001)
8. Bacher, J., Brand, R., Bender, S.: Re-identifying register data by survey data using cluster analysis: An empirical study. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems 10(5), 589–608 (2002)
9. Templ, M.: Software development for SDC in R. In: Domingo-Ferrer, J., Franconi, L. (eds.) PSD 2006. LNCS, vol. 4302, pp. 347–359. Springer, Heidelberg (2006)
10. Muralidhar, K., Sarathy, R., Dankekar, R.: Why swap when you can shuffle? a comparison of the proximity swap and data shuffle for numeric data. In: Domingo-Ferrer, J., Franconi, L. (eds.) PSD 2006. LNCS, vol. 4302, pp. 164–176. Springer, Heidelberg (2006)
11. Mateo-Sanz, J., Sebe, F., Domingo-Ferrer, J.: Outlier protection in continuous microdata masking. In: Domingo-Ferrer, J., Torra, V. (eds.) PSD 2004. LNCS, vol. 3050, pp. 201–215. Springer, Heidelberg (2004)
12. Rousseeuw, P.J., Van Driessen, K.: A fast algorithm for the minimum covariance determinant estimator. Technometrics 41, 212–223 (1999)
13. Filzmoser, P.: A multivariate outlier detection method. In: Aivazian, S., Filzmoser, P., Kharin, Y. (eds.) Proceedings of the Seventh International Conference on Computer Data Analysis and Modeling, pp. 18–22. Belarusian State University, Minsk (2004)
14. Templ, M., Meindl, B.: Why shuffle when you can use robust statistics for SDC - a simulation study. In: Domingo-Ferrer, J., Saygın, Y. (eds.) PSD 2008. LNCS, vol. 5262. Springer, Heidelberg (2008)
15. Mateo-Sanz, J., Domingo-Ferrer, J., Sebe, F.: Probabilistic information loss measures in confidentiality protection of continuous microdata. In: Webb, G. (ed.) Data Mining and Knowledge Discovery, vol. 11, pp. 181–193. Springer, Heidelberg (2005)

16. Muralidhar, K., Sarathy, R.: Data shuffling- a new masking approach for numerical data. Management Science 52(2), 658–670 (2006)
17. Brand, R., Giessing, S.: Report on preparation of the data set and improvements on sullivans algorithm. Technical report (2002)
18. Ting, D., Fienberg, S., Trottini, M.: ROMM methodology for microdata release. In: Monographs of official statistics, Work session on statistical data confidentiality, Eurostat, Luxembourg (2005)
19. Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Lenz, R., Longhurst, J., Schulte-Nordholt, E., Seri, G., De Wolf, P.P.: Handbook on statistical disclosure control version 1.01 (2007)
20. Templ, M.: sdcMicro: A new flexible R-package for the generation of anonymised microdata - design issues and new methods. In: Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality. Monographs of Official Statistics (to appear, 2007)

Parallelizing Record Linkage for Disclosure Risk Assessment

Joan Guisado-Gámez[1], Arnau Prat-Pérez[1], Jordi Nin[2],
Victor Muntés-Mulero[1], and Josep Ll. Larriba-Pey[1]

[1]DAMA-UPC, Dept. d'Arquitectura de Computadors
Universitat Politècnica de Catalunya,
Campus Nord, C/Jordi Girona 1-3
08034 Barcelona, Catalonia, Spain
{joan,aprat,vmuntes,larri}@ac.upc.edu
http://www.dama.upc.edu/
[2]IIIA, Artificial Intelligence Research Institute
CSIC, Spanish National Research Council
Campus UAB s/n
08193 Bellaterra Catalonia, Spain
jnin@iiia.csic.es

Abstract. Handling very large volumes of confidential data is becoming a common practice in many organizations such as statistical agencies. This calls for the use of protection methods that have to be validated in terms of the quality they provide. With the use of Record Linkage (RL) it is possible to compute the disclosure risk, which gives a measure of the quality of a data protection method. However, the RL methods proposed in the literature are computationally costly, which poses difficulties when frequent RL processes have to be executed on large data.

Here, we propose a distributed computing technique to improve the performance of a RL process. We show that our technique not only improves the computing time of a RL process significantly, but it is also scalable in a distributed environment. Also, we show that distributed computation can be complemented with SMP based parallelization in each node increasing the final speedup.

Keywords: Record linkage, parallel computing, distributed computing, disclosure risk evaluation.

1 Introduction

The need for data protection methods is larger every day, becoming crucial to anonymize confidential information before releasing it in a private manner. This is true in many situations where data becomes public or semi-public, and a corrupt use of it may lead to the disclosure of such confidential information. A situation where this may arise is when data is released by statistical agencies, where there is a need to preserve the statistical properties of the information while keeping it anonymous.

J. Domingo-Ferrer and Y. Saygın (Eds.): PSD 2008, LNCS 5262, pp. 190–202, 2008.

However, when a protection method is applied, the evaluation of the privacy provided by such method becomes a problem. Re-identification techniques, such as Record Linkage (RL) methods [16,17,18], are the most common techniques for evaluating the quality of a given protection method, *i.e.* the disclosure risk. RL methods model the situation where an intruder sees the protected recordset whereas he has access to records of the original recordset obtained from other sources. The goal of the RL methods used by an intruder is to link the original records with the corresponding records in the protected recordset. As a consequence, the larger the number of records linked by means of these record linkage methods, the larger the disclosure risk of the protection method.

Nowadays, there are two important aspects in the anonymization process. First, the amount of data collected is larger every day due to the availability of larger population databases. Second, the need for faster methods is also more important because it is necessary to provide service to more frequent demands. These two aspects call for the use of parallel applications during the RL process.

In this paper, we propose a distributed strategy that focusses on speeding up the RL processes for large data volumes. We propose this assuming that the recordset fits in the memory of a single computer, deploying a strategy to perform the most expensive part of the whole RL computation on a farm of slave computers with independent memories. The simplicity of the method makes it viable on a set of personal computers connected through a LAN, which makes the whole system feasible for situations where there is a need for speed but the resources for large parallel computers are not available.

Our results show that the use of parallel computing devices on large data sets improves the performance of the RL methods leading the larger sets of data on large numbers of processors to linear speedups of almost P for P processors. The results also show that the overhead of performing a distributed execution of the RL code does not grow with the number of processors, demonstrating the scalability of the strategies proposed. Note also that, having multicore processors make it possible to implement an SMP based parallelization in each node in order to achieve a better performance.

The rest of the paper is organized as follows. We start setting up the problem and describing a memoization technique in Section 2. We describe the technique to distribute the computations in Section 3. We evaluate the parallelization in Section 4. Finally, we explain the related work in Section 5 and conclude in Section 6.

2 Record Linkage

Record Linkage aims at processing a set of recordsets in order to obtain sets of records that belong to the same unique individual. We consider that RL is formed by different phases as shown in Figure 1. First, the data sources are cleaned and pre-processed normalizing attributes in the recordset files individually to allow a simpler comparison with other data in the following steps [9].

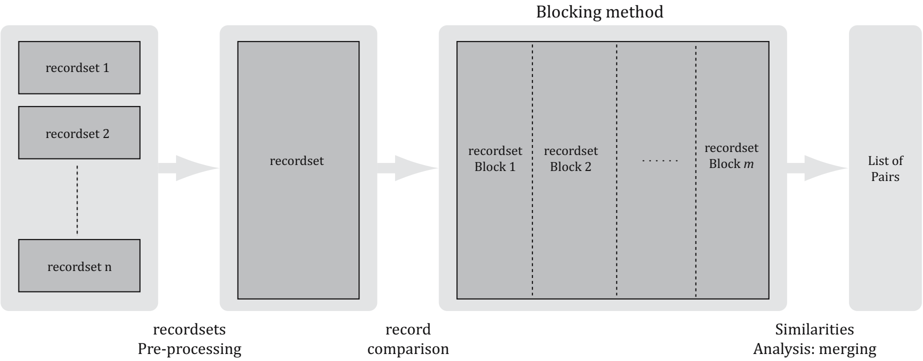


Fig. 1. Record Linkage processing model

Once the pre-processing is done, RL proceeds with the record comparison. The objective of this phase is to obtain pairs of records that possibly belong to the same individual. There are two kinds of RL algorithms for record comparison: those based on probabilistic methods and those based on distance functions [16,17]. During the RL process, records are compared following a strategy that may have several objectives, like reducing the number of comparisons as with the Standard Blocking[2,7] or Sliding Window [10] (also known as Sorted Neighborhood) methods, or finding the largest groups of similar records at the lowest comparison cost as with Reduction using Anchor Record (RAR) [14].

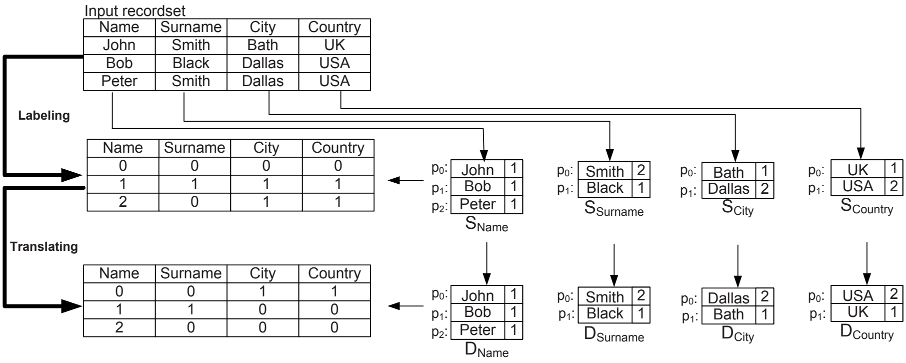
In order to avoid possible errors induced by blocking methods, it is usual to perform several passes using different sorting criteria, like the name or the surname in the case that the entities are human beings.

Finally, it is necessary to analyze the record pairs so that groups of similar records are formed and false positives are discarded if possible, before the eventual expert review process is done.

Record Comparison Process

The comparison phase is the most expensive in the RL process, with quadratic complexity ($O(N^2)$) in the number of records N, as opposed to the linear complexity of the other phases. In this paper we will use the Sliding Window method in order to reduce the record comparison phase complexity to ($O(BN)$), where B is the size of the block used (window). However, it is still the most complex phase in the whole RL process. Therefore, several additional techniques have been proposed in order to further improve performance. Among these, the use of memoization techniques for reducing the number of comparisons has been proven as one of the most effective in terms of performance [6]. In this paper, we assume this proposal as the baseline for our work.

After these considerations and before presenting our technique, we briefly describe the different phases of the Record Comparison process used in this paper. The Record Comparison process is divided into four phases: Comparison Preprocess, Caching, Detection and Merging. Following, we describe these phases.



Input recordset

Name	Surname	City	Country
John	Smith	Bath	UK
Bob	Black	Dallas	USA
Peter	Smith	Dallas	USA

Name	Surname	City	Country
0	0	0	0
1	1	1	1
2	0	1	1

Name	Surname	City	Country
0	0	1	1
1	1	0	0
2	0	0	0

Fig. 2. Labeling and Translating steps

Comparison Preprocess. This phase aims at reducing the comparison cost and making the use of memoization techniques simpler in order to reduce the amount of computations involved in the comparison. We distinguish between two steps in this phase. The first step is **Labeling**, as shown in Figure 2, where the string values in the records are replaced with integer identifiers, and dictionaries are created to match each string with its identifier. This allows to perform exact comparisons between identical strings very efficiently, as well as, preparing the data to make it possible to use memoization techniques with those strings that are not equal.

Although the use of identifiers simplifies the exact comparison of attributes, it is still necessary to compare the non-exact matching values. In our case, this is done by using a two level String Comparison Function based on the Levenshtein distance [15] applied to pairs of tokens coming from the compared string values. Note that the objective of the following steps is to minimize the computational cost of using such approximate comparison function.

As Figure 2 shows, for each comparison attribute A, a list S_A is created at loading time. $S_A = \{p_0, p_1, \cdots, p_{n-1}\}$ where n is the number of unique strings in A. Each p_i is a pair $(v_i, |v_i|)$ where v_i is a string value in A, and $|v_i|$ is its number of occurrences in A. At the same time, each value v in A is replaced by i (*i.e.* its position in the list is used as a string identifier), where $p_i = (v_i, |v_i|)$, $v_i = v$ and $|v_i|$ is the number of occurrences of v in A. Note that the structures on the right hand side of Figure 2 show the dictionaries, while the left hand side part illustrates the data transformation from a string to an identifier.

The second step is **Translating**, also depicted in Figure 2. It consists in sorting S_A obtaining a new structure denoted by D_A. D_A has the same elements than S_A but they have been sorted by the number of times they appear in A. Then, identifiers are reassigned, giving the smallest identifier to the most frequent value. Note that because of the sorting process, it is necessary to translate the identifiers of each value v of A to the new position occupied in D_A.

Memory Allocation. In this phase, a cache for each comparison attribute A is created. Note that, the caches are only created but not populated. These caches are used during the rest of the phases. The caching structures allow storing the results of the comparisons, minimizing the number of actual comparison computations. Usually, the number of unique values in A makes it unfeasible to store the result for all the possible pairs of strings in memory, thus, only a subset of the comparisons is stored. Given that, the cache size depends on the memory available in the system, only m elements in A can be represented. Since D_A is sorted decreasingly by the number of occurrences, the comparisons between the m most frequent strings in A are stored in the cache. Therefore, the comparison between those frequent values are memoized avoiding unnecessary comparisons during the process.

These caches are called Comparison Stores (CSs), and they are proposed and described in detail in [6].

Detection. In this phase the comparisons between records are performed and finally the similarities are detected. In order to find the maximum number of similarities, several passes of a blocking method are performed. Each pass uses a different criteria for sorting the recordset.

During this phase, the Comparison Stores are populated and used intensively. When the comparison of a frequent pair of strings is performed, the result is stored in the CS and it can be used later when the same strings have to be compared. Thus, this result will never be computed again, minimizing the number of comparisons for the frequent values.

Merging. This phase focusses on creating a similarity list removing duplicated similarities that appear as a consequence of doing several passes in the Detection phase. A similarity pair is only inserted in the list if it has not been inserted previously. In order to achieve a good performance, a hash table is used for controlling the similarities that have been already inserted in the list.

3 Record Comparison Process Parallelization

Now, we describe the parallelization of the Record Comparison process. We take the Record Comparison schema presented in Section 2 as the baseline because it is the most recent high performance RL strategy to our knowledge. However, the parallel technique we present in this paper can be used on any Record Comparison Process.

In this paper, we parallelize the Record Comparison process. Given a set of nodes, each one processes a portion of the recordset. Figure 3 shows the cluster-based parallelization that follows a Master/Slave architecture. The Master node is responsible for maintaining the whole input recordset and for sending the data necessary to each Slave node for processing. In fact, this could have two different implementations. First, assuming that all the nodes have all the data, thus, the parallelization comes from the tasks on different data subsets engaged by the different nodes. Second, assuming that only the Master node stores the data, and the Slaves receive the data subsets necessary at each precise moment. We opted for the

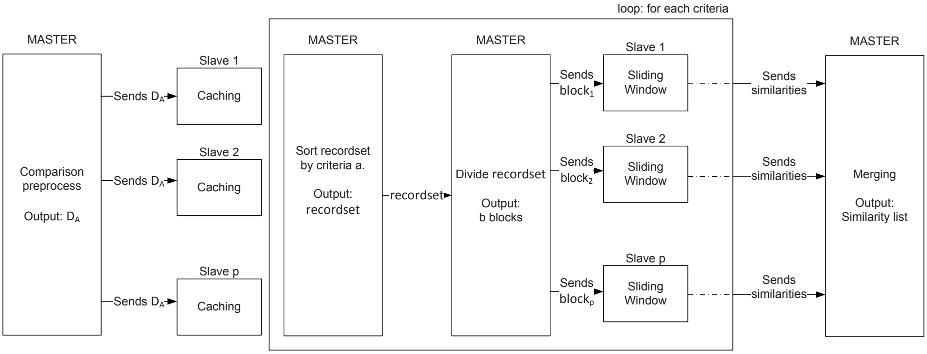


Fig. 3. Parallelization of the Record Comparison Process

second option because it maximizes the amount of space for auxiliary data structures in the Slaves. Thus, by saving record space, we can create larger Comparison Stores, which is significantly beneficial as shown in Section 4.

Assuming that the Master node is the unique node that keeps the whole recordset, it is the only node that can perform the Record Comparison preprocess, and create the D_A lists. As the Slaves need the D_A lists in order to construct their own caches and for the Detection phase, the Master node has to send them to the Slaves. Note that each Slave has its own independent caches (*i.e.* the cached comparisons are not shared among Slaves) so the same comparison may be performed in each Slave, in contrast with the sequential version where the same comparison is performed only once. This part of the preprocess is sequential and it will become the most expensive part of the computations when parallelism is applied, as we show later.

Once all the Slaves have their own D_A structures and caches, they are ready to perform the Detection phase. For each pass in the Detection phase, the Master node has to sort the recordset by the current criteria. Once the recordset is sorted, it is divided into a set of data blocks, in our case $b = \frac{|recordset|}{|Slaves|}$ blocks. The Master node sends each block to a different Slave, and if the Master plays the Slave role, it keeps the last block. At that point, as explained in Section 2, each Slave can start performing the Sliding Window strategy over its block. Once a Slave has finished performing Sliding Window, it has to wait until the Master node sends a new block to be processed. Note that, if the Master plays the Slave role, it will have to finish applying the Sliding Window strategy over its own block before sorting the recordset by the next criteria, dividing it into b blocks and sending them to the Slaves. Therefore, the slower the Master node in performing the Sliding Window, the longer the Slaves will be waiting idle, without working. The convenience of using the Master as a Slave will be discussed in Section 4.

Each Slave at the Detection phase creates a list with the similarities it finds. Because of the several passes that are performed during the Detection phase, some similarities may be duplicated among the lists. Therefore it is necessary that the Slaves send their own list to the Master node. The Master node will merge them creating a similarity list where no duplicated similarities will appear.

4 Experiments

With the objective of evaluating the proposed parallelization, we run a set of experiments that shows the interaction between the size of the recordset with the number of nodes to perform the RL process. To compare the results of our parallelization we will use the Record Comparison process described above which is the best possible sequential algorithm at hand.

The experiments use different numbers of records ranging from 1 million (1M) to 8 million (8M) 128 byte records. The recordsets used in these experiments have been generated using the synthetic record file generator included in the FEBRL toolkit [11], with a 30 percent of duplicates and a maximum of 10 duplicates per record. However, in order to make the whole evaluation as realistic as possible, the frequencies of the names and surnames used to generate the synthetic recordsets with FEBRL have been obtained from the Catalan Statistics Institute [5]. Recordsets are composed of records with eight attributes, out of which four are strings: first name, first surname, second surname, as in the Catalan person-naming system, and address.

In order to perform the experiments we have used a Beowulf [12] cluster with the features described in Table 1.

Table 1. Cluster description

Beowulf cluster	
Number of nodes:	16
Processor of each node:	Intel Core 2 Duo 6600 @2.4Ghz
Memory of each node:	2GB
L2 Cache size of each node:	4096KB
Network:	1Gbps

Comparison of Parallel Strategies

With the objective to minimize the effort to parallelize, we first want to understand the advantages of using the Master as a Slave or not. We will refer to the version where the Master node acts as a Slave, as the Worker-Master version, and the version where it does not, as the Lazy-Master version.

Figure 4 shows the Record Comparison process time in (a), and the cache size of the Master and the Slaves in (b), when different amount of records are considered in both versions. The number of nodes is fixed to 16. Note that, there are 15 Slaves and 1 Master. Note also that, the cache size of the Slaves is equal for both versions, and in the Lazy-Master version the Master node does not have cache because it does not need it.

In Figure 4(a), we can observe that up to a recordset of 4 millions of records, the results are almost the same in both versions. On the contrary, when the recordset exceeds 4 millions, the Worker-Master version is slower. As explained in Section 3 for the Worker-Master version, once the Slaves have finished performing the Sliding Window strategy over their block, they have to wait until the Master

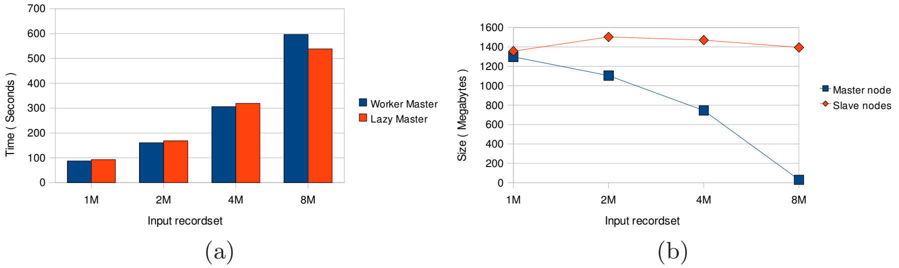


Fig. 4. Execution time for Worker-Master vs. Lazy-Master (a) and Cache sizes for the Master node compared to the Slaves nodes (b)

node finishes also with the Sliding Window over its own block. Afterwards, it sorts the recordset, divides the recordset into new blocks and sends them to the Slaves. In Section 2 the creation of the caches is described as a process that uses the maximum available memory. Since the Master node is the node that contains the input recordset, the larger the recordset size, the smaller the amount of available memory in the Master node. Figure 4 (b) shows that as the number of records grows, the Master cache size diminishes and the difference between the Slaves and Master cache increases. Therefore, the Master node takes longer in performing Sliding Window just because it has to do more comparisons, since they do not fit into its cache and the Slaves have to be waiting more time for the Master to finish.

This experiment supports the choice of using a single node, the Master, for containing the input recordset. It also shows that for large recordsets the Lazy-Master version is better. Since this paper is focused on the management of very large volumes of data, from now on we will work with the Lazy-Master version.

Time Analysis

In this experiment, we want to analyze the Record Comparison process time. This time is dissected as follows: the Detection time, which is the time spent in the Detection phase; the overhead time, which is the time spent in sending the different data over the network; and the Record Comparison preprocessing.

Figure 5 shows the dissection for the parallel version with 2, 4, 8 and 16 nodes. The input size has been fixed at 8 millions of records. Since we have parallelized the Detection phase, its weight over the total time decreases quickly while increasing the number of nodes, so this means that at some point, it is not useful to increase the number of nodes due to Amdahl's law [1]. Note that thanks to our technique, we have reduced the time to link 8 million records from 100 to only about 18 minutes.

It is also interesting to observe the influence of the overhead over the total execution time. The tests show that this overhead is negligible. Figure 6 presents

[1] Amdahl's law is used to find the maximum expected improvement to an overall system when only part of the system is improved.

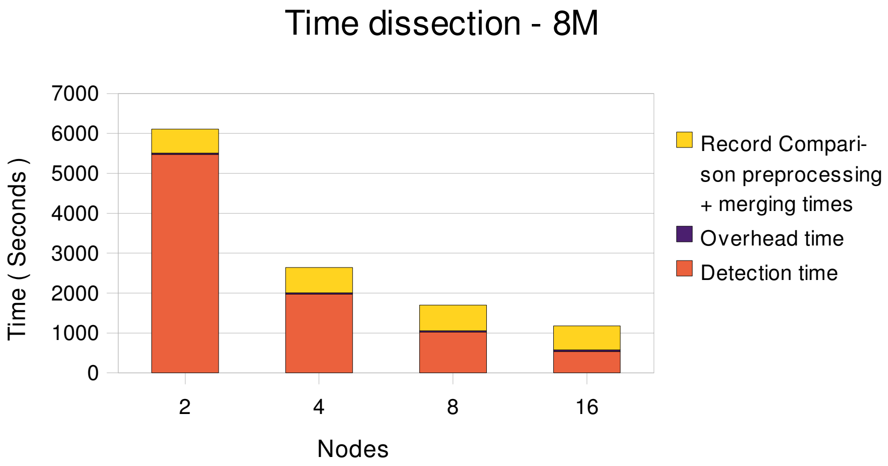


Fig. 5. Dissection of total time for 8M records

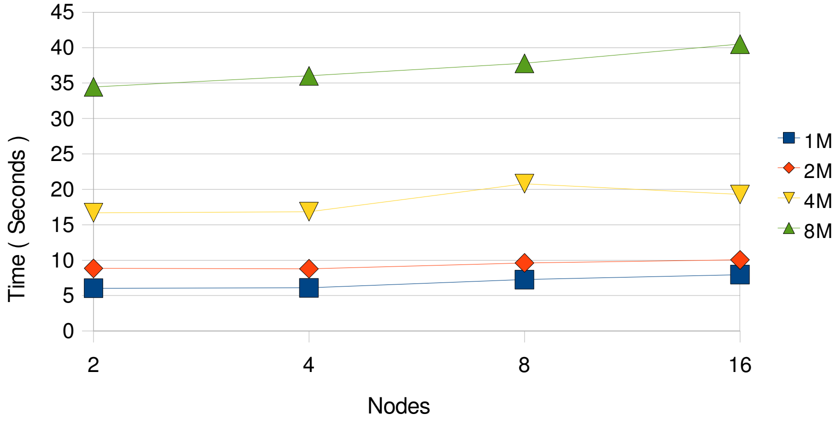


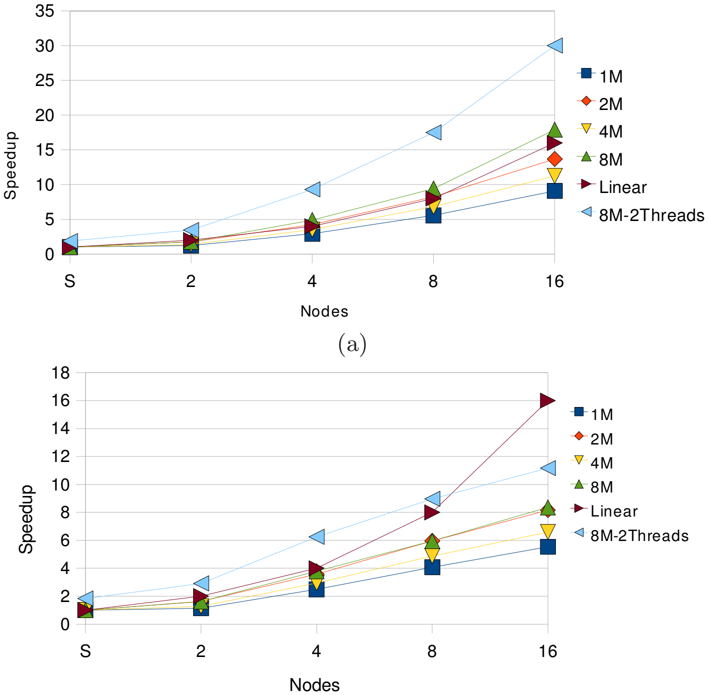
Fig. 6. Overhead produced by the parallelization

the value of the overhead time for the parallel version with 2, 4, 8 and 16 nodes. The input sizes are set at 1, 2, 4 and 8 million of records.

The most interesting observation of these results is the slow linear progression followed for each input recordset. This means that the proposed parallelization is very scalable in the number of nodes.

Speedup Analysis

In this experiment, we want to test the performance obtained with the parallelization proposed in Section 3. We use as baseline for calculating the speedup the sequential version of the Record Comparison process. The executions are run using from 2 to 16 nodes with 1 to 8 million records. We will also see the behavior when the 8M recordset is linked using an SMP parallelization with 2 threads.



(a)

(b)

Fig. 7. Speedup for Detection phase (a) and Record Comparison process (b)

Figure 7(a) shows the speedup obtained in the Detection phase. We can observe that the speedup varies with the size of the input. Generally, the larger the recordset, the better the speedup. Since the differences are clearer in the case of using 16 nodes, we will focus on this scenario. These differences happen because in the sequential version, the larger the recordset, the smaller the caches that fit in memory. On the contrary, in the parallel version, the cache sizes of the Slaves are the same for any input recordset size, because the Slaves are not storing the input recordset. This explains the large distance between the speedup in the 1 million recordset case and the 8 million recordset case. To understand the short distance between 2 million and 4 million recordsets we have to take into account that, as explained in Section 3, since the cached values are no shared among Slaves, more comparisons are performed. Depending on the content of the recordset, the same comparison will be performed in more or less nodes and the speedup will be affected. Although these variations, we can observe that, in average, a significant speedup is obtained, and even better than linear for 8M records. It is really interesting to observe that the speedup for 8M-2threads doubles 8M using 1thread which proves that our distributed technique is complementary with an SMP based parallelization in each node.

We can observe the speedup obtained in the whole Record Comparison Process in Figure 7(b). Note that the speedup obtained is not linear because, as we have seen in the previous experiment, there is a constant time corresponding to the addition of Record Comparison preprocessing and merging time. However, we are able to perform the execution 8 times faster when we manage large recordsets using just 1 thread and 11 times faster when we use 2 threads.

5 Related Work

Re-identification methods are a specific class of data base techniques. These methods are designed to establish relationships among different entities or attributes stored in different data sources. Obtaining the relationships among entities or attributes makes sense in many scenarios such as: Schema matching [1], Data integration [13],Data cleaning [3] and Object integration [8].

There are different classic approaches for the reduction of work during the Record Linkage process, like the Standard Blocking [2,7] and the Sliding Window [10] methods, that intend to reduce the number of record comparisons. On the other hand, methods like RAR is aimed at reducing the number of attribute comparisons [14].

Finally, it is possible to find another approach to reduce the execution time of the RL process by using parallelism, as explained in [4], where it is necessary to know the state of the recordsets processed, either if they are clean or dirty. A good performance is achieved in [4] when the recordsets managed are clean. However, it is unsuitable for very large recordsets, specially when they are dirty, which is the common case. Note that in our approach we are not distinguishing between dirty and clean recordsets. In fact we are always assuming dirty ones since this is the worst case, assuming clean ones would mean do less comparisons among records, and therefore obtaining better times.

6 Conclusions and Future Work

In this paper we have shown that applying distributed strategies to a Record Linkage process is very useful and simple. This shows that organizations with little computing resources may use the PCs in their desktops in a Beowulf configuration to accelerate their RL processes in a cheap and efficient way.

Future work will include proposing more complex algorithms in order to increase the speedup, focusing on the preprocessing also, as the problem to tackle at this point. Another approach will be the analysis of clusters of non-homogeneous computing devices, where instead of dividing the recordset into blocks of equal size, it will be necessary to divide the recordset as a function of the available memory of the nodes, which will help us to obtain a better speedup. It will be also interesting to study the effect of sharing the cached comparisons among the nodes to reduce the amount of comparisons.

Another final approach will be to spread the input recordset out among the nodes, in order to be able to manage an input recordset that does not fit into the memory of a single node. This will imply also the use of parallel sorting techniques in the Detection phase.

Acknowledgments

Partial support by the Spanish MEC (projects ARES – CONSOLIDER INGENIO 2010 CSD2007-00004, eAEGIS – TSI2007-65406-C03-02 and TIN2006-15536-C02-02) and by the Government of Catalunya (grant 2005-SGR-00093 and GRE-00352) is acknowledged. Josep L. Larriba-Pey wants to thank UPC, Government of Catalunya and the Spanish MICINN for his I3 grant.

References

1. Rahm, E., Bernstein, P.A.: A survey of approaches to automatic schema matching. The Very Large Database Journal, 334–350 (2001)
2. Newcombe, H.B.: Record linking: The design of efficient systems for linking records into individuals and family histories. American Journal of Human Genetics (1967)
3. Do, H.H., Rahm, E.: COMA - A system for exible combination of schema matching approaches. In: Proceedings of the 28th Very Large Databases Conference, pp. 610–621 (2002)
4. Kim, H., Lee, D.: Parallel Linkage. In: CIKM, Lisboa, Portugal (2007)
5. http://www.idescat.net
6. Gómez, J., Larriba, J.L., Ribes, J.: Improving Record Linkage Performance. Technical report UPC-DAC-RR-2006-15
7. Jaro, M.A.: Advances in Record Linkage Methodology as Applied to Matching the 1985 Census of Tampa, Florida. Journal of the American Statistical Society, 414–420 (1989)
8. Atencia, M., Schorlemmer, M.: A formal model for situated semantic alignment. In: Proceedings of the 6th International Conference in Agent and Multiagent Systems (2007)
9. Bilenko, M., Basu, S., Sahami, M.: Adaptive Product Normalization: Using Online Learning for Record Linkage in Comparison Sopping. In: Proceedings of the 5th Int'l. Conference on Data Mining 2005, pp. 58–65 (2005)
10. Hernandez, M., Stolfo, S.: The merge/purge problem for large database. In: ACT SGMOD Conf. Proc., pp. 127–138 (1995)
11. Christen, P., Churches, T.: Febrl: Freely extensible biomedical record linkage. Joint Computer Science Technical Report TR-CS-02-05 (2002)
12. Brown, R.G.: Engineering a Beowulf-style Compute Cluster. Duke University Physics Department (2004)
13. Deen, S.M., Amin, R.R., Taylor, M.C.: Data integration in distributed databases. IEEE Transactions on Software Engineering (1987)
14. Sung, S.Y., Li, Z., Peng, S.: A Fast Filtering Scheme for Large Database Cleansing. In: International Conference on Information and Knowledge Management (CIKM), McLean, Virginia,USA (2002)

15. Levenshtein, V.I.: Binary codes capable of correcting deletions, insertions, and reversals. Soviet Physics Doklady, 707–710 (1966)
16. Torra, V., Domingo-Ferrer, J.: Record linkage methods for multidatabase data mining. In: Information Fusion in Data Mining, pp. 101–132. Springer, Heidelberg (2003)
17. Winkler, W.E.: Data cleaning methods. In: Proc. SIGKDD 2003, Washington (2003)
18. Winkler, W.E.: Re-identification methods for masked microdata. In: Domingo-Ferrer, J., Torra, V. (eds.) PSD 2004. LNCS, vol. 3050, pp. 216–230. Springer, Heidelberg (2004)

Extensions of the Re-identification Risk Measures Based on Log-Linear Models

Daniela Ichim

Servizio di Progettazione e Supporto Metodologico nei Processi di
Produzione Statistica
Istituto Nazionale di Statistica
via Cesare Balbo 16, 00184 Roma Italia
ichim@istat.it

Abstract. A global measure of the re-identification risk in microdata files is analyzed. Two extensions of the log-linear models are presented. The first methodology considers the weights in the analysis of contingency tables. The results of several tests performed on real data are presented. In the framework of statistical disclosure control, the second methodology proposes a maximum penalized likelihood approach to the computation of smooth estimates.

Keywords: Statistical disclosure control, microdata, sample uniques, log-linear model, smoothness.

1 Introduction

National statistical institutes (NSIs) often release microdata files from the sample surveys. Such files represent an important source of information for researchers. A major concern for the NSI releasing data on individuals is the need to the protect confidentiality of respondents. Generally this is performed by making assumptions on the tools an intruder might use in order to breach the confidentiality of respondents.

Using an external register, e.g. a file containing information on the population, a possible intruder might try to gather confidential information about (some) individuals by comparing the variables shared by the released microdata file and the external register. Assuming that the shared variables are categorical, in order to identify a sample unit, the intruder would probably compare the combinations of the shared variables. This scenario is often used by the NSIs when releasing individual data stemming from social surveys.

The probability of establishing a correct link between a sample unit and a population unit depends on the frequencies of the combinations of the shared variables in the two files. Obviously, the units that are rare in the sample and population have a greater risk of re-identification. However, the NSI might not have easy and fast access to any external register. Consequently, the NSI generally estimates the risk of re-identification using only the information contained in the sample. An important problem in statistical disclosure control (SDC) is

J. Domingo-Ferrer and Y. Saygın (Eds.): PSD 2008, LNCS 5262, pp. 203–212, 2008.

the estimation of the number of sample uniques that are also population uniques. This global risk measure might be used by the NSIs when deciding the release of a file containing individual records. Individual risk measures are also important, but they are generally used in order to apply statistical disclosure limitation methods.

In this paper the estimation of a global risk measure is discussed. Two extensions of the Poisson-log-linear model are presented. The notations and the general framework are introduced in section 2. In section 3 the problem of the parameter estimation of log-linear models with complex survey data is addressed. Possible methods to deal with the contingency table structure are illustrated in section 4; a penalized maximum likelihood function is proposed for the estimation of sample uniques. Finally, some testing results are presented in section 5.

2 Notations

Consider a simple random sample of size n drawn from a population of size N and denote by $\pi = n/N$ the sampling fraction. Suppose that the re-identification risk in a microdata file must be measured and imagine that the re-identification of units could be performed using an external register by means of some common variables, called key variables. The implicit assumptions of this scenario are discussed, for example, in [8] and [14]. In this scenario the key variables are categorical.

Denote by K the number of cells of the table defined by cross-classifying the key variables $X_1, \ldots, X_m$. Let F_k be the number of population units belonging to the k-th cell. Similarly, let f_k be the number of sample units in the k-th cell.

As reported in [3], a global measure of the re-identification risk is given by the number τ_1 of sample uniques that are also population uniques:

$$\tau_1 = \sum_{k=1}^{K} \mathbb{I}(F_k = 1, f_k = 1)$$

When $F_k, k = 1, \ldots, K$ are independent variables following Poisson distributions with means λ_k and the sample is selected by Poisson sampling with selection probability π, an estimate of τ_1 is given by:

$$\hat{\tau_1} = \sum_{k=1}^{K} \exp(-\mu_k(1-\pi)/\pi), \quad \mu_k = \pi\lambda_k \tag{1}$$

In [13] the full derivation of the above formulas is given.

Generally, $\boldsymbol{\mu} = (\mu_1, \mu_2, \ldots, \mu_K)^{'}$ is estimated using a standard log-linear model

$$\log(\mu_k) = \boldsymbol{x}_k^{'}\boldsymbol{\beta} \tag{2}$$

where $\boldsymbol{x}_k^{'}$ is the specified vector of main effects and interactions of $X_1, \ldots, X_m$. If model (2) is in closed form, the iterative proportional fitting (IPF) algorithm might be used to estimate $\boldsymbol{\beta}$ and $\boldsymbol{\mu}$. Otherwise, the Newton-Raphson method might be used to maximise the likelihood function, see [1].

The standard log-linear approach does not completely consider the data generation process. In the SDC framework, the characteristics of the complex survey should be taken into account, too. Methods to deal with the survey weights and relationships between variables are proposed in the next sections.

3 Survey Weights in Log-Linear Models

The surveys conducted by the NSIs often have a complex structure. From data collection to data processing, each step has a significant impact on data analyses.

An important feature of the sampling surveys is that a weight is associated to each sampled unit. The weights are supposed to account for the sampling and non-response error. From a data analysis point of view, the weights are used to estimate the population characteristics, see [7] for example.

Consider the table derived from the cross-classification of the key variables. Using log-linear models, this contingency table is used to estimate the parameters of the Poisson distributions underlying each cell count, see equation (2). Should we use frequencies or weighted frequencies in the fitting step?

Three approaches to log-linear modeling for complex surveys are discussed: use a) unweighted frequencies, b) weighted frequencies, and c) log-rate models.

Ignoring the weights is certainly the easiest choice. It might even provide good results especially when the weights depend only on the independent variables in model (2). In other words, unbiased estimates will be obtained if (some of) the predictors of the log-linear model are the stratifying variables. However, it should be stressed that both clustering and correlation between weights and the dependent variable might induce biases.

Moreover, in the SDC framework, the model (2) is expressed in terms of key variables. If a stratifying variable is not a key variable, it should not be included in (2). In such situations, biased estimates could be obtained. In real surveys, there might exist stratifying variables that are not key variables. This means that in practical situations it is not always possible to include all the stratifying variables in model (2). Some of this issues were discussed also in [11].

The second approach uses the pseudo maximum likelihood estimation method described in [9]. First the weighted frequencies f_k^w are calculated, $f_k^w = \sum_{i \in J_k} w_i$, where J_k denotes the k-th cell and w_i is the weight of the i-th unit. Then the contingency table is analyzed as if it were an unweighted table. In contrast with the previous approach, the pseudo maximum likelihood estimation can deal with sampling weights, stratification and clustering. The estimates will be unbiased. However, the standard goodness-of-fit tests, like χ^2 or Pearson, can no longer be used because the assumption of independence of observations would be violated when comparing the weighted frequencies f_j^w with the estimated frequencies.

The third alternative, see [2], is to extend the log-linear models to include the weights as an offset variable:

$$\log(\mu_k) = \log(z_k) + \boldsymbol{x}_k^{'}\boldsymbol{\beta} \tag{3}$$

where $z_k = 1/w^k$ is the inverse of the average cell weight $w^k = f_k^w/f_k$.

It is easy to see that model (3) takes into account the population size, leading to the log-rate model described in [6], a model for rates instead of counts.

Using this model, the estimates of the parameters would depend on the weighted frequencies, while the standard errors will depend on the unweighted frequencies. Both parameter estimation and goodness-of-fit tests will be correct. Consequently, the log-rate model should be preferred to the other two choices when the contingency tables derive from complex survey data.

There is a further advantage of the log-rate models. If model (3) is written as:

$$\mu_k = z_k \exp(\boldsymbol{x}_k^{'}\boldsymbol{\beta}),$$

by simply setting $z_k = 0$, the structural zeros may be readily dealt with.

In section 5, the results of some experiments performed using the above models are presented.

Of course, the log-linear model (3) is only a partial solution to the analysis of complex surveys. In the statistical disclosure control framework further developments in the field of the analysis of contingency tables for finite populations are required. For example, the sampling design features might provide useful information for the estimation of τ_1.

4 Smoothness in Log-Linear Models

As discussed in section 2, the contingency table to be analysed using log-linear models derives from the key variables cross-classification. When these variables have many categories, as it often happens in official statistics surveys, the contingency table is sparse. The analysis of sparse tables could give two types of problems. The first one is related to the goodness of fit tests since the χ^2 statistics do not preserve their asymptotic properties. Another problem is due to the possible non-convergence of the algorithms like Newton-Raphson or IPF.

There are two well-known general solutions to the table sparseness problem: table redesign and adding a constant. Collapsing cells and/or omitting variables are compromise strategies, but these approaches could ignore potentially important contributions. Moreover, in the SDC framework, recoding should be applied only to reduce the risk of re-identification. The second solution is the addition of a small number α, called flattening constant, to all or only to the empty cells. Different choices of α have been proposed: $1, 0.5, \sqrt{n}/K$, etc.. A review may be found in [5]. One of the effects of adding a constant is that the sample size is increased and the introduced total count might dominate the cell proportion estimates.

Keeping models as simple as possible would attenuate the sparseness effects. Anyway, it was observed in [10] that, when a simple independence model is used in (2), the estimation of μ_k would be based on information from all the cells having in common even a single characteristic with J_k.

For the estimation of the re-identification risk, a local neighboring approach to the analysis of contingency tables using the log-linear model (2) was presented in [10]. The approach concerns ordinal key variables. The idea is that a sample unique for which its neighboring cells have small values is more likely a population unique than other sample uniques. More precisely, considering a similarity distance d between the levels of the key variables, in [10] the maximisation of the local likelihood function

$$\mathcal{LL}(\boldsymbol{\beta})=\sum_{k'\in N^k}\left[f_{k'}\left[\beta_0+\ldots+\beta_t\left(d(k^{'},k)\right)^t\right]-\exp\left(\beta_0+\ldots+\beta_t\left(d(k^{'},k)\right)^t\right)\right] \tag{4}$$

is proposed, where N^k denotes the considered neighborhood of the k-th cell. In [10] several choices of N^k and d are presented, taking into account their possible multi-dimensionality.

With respect to the standard local polynomial regression framework, see [4], the kernel function used in equation (4),

$$W_0(u)=\begin{cases}1,\ u\in(-1,1)\\ 0,\ otherwise\end{cases}$$

is not a continuous function. This might cause problems to the asymptotic properties of the estimators. Indeed, W_0 is not a proper smoothing function; it ensures only a truncated local polynomial regression. Moreover, if the kernel W_0 is used, all the cells in N^k would equally contribute to a cell parameter estimation.

A second characteristic of $\mathcal{LL}$ is the number of parameters. Even if only the estimated intercept $\hat{\beta}_0$ is used to compute $\hat{\mu}_k$, the other estimates $\hat{\beta}_1,\hat{\beta}_2,\ldots,\hat{\beta}_t$ should be computed, too. Unfortunately, the maximisation of (4) should be performed for each cell, or, at least, for the sample unique cells. This means that the overall number of parameters to estimate is proportional to the number of (sample unique) cells and to the polynomial degree t. The impact this large number of estimations on data analyses and goodness of fit tests is not clear.

Consider again the main idea of the proposal presented in [10]: " if the neighboring cells of a sample unique are small or empty, then it is more likely to have arisen from a small population cell". In the same spirit, the opposite statement should also be true: if the cells in N^k have large values, the k-th cell should not have a small value. In other words, neighboring cells should have values similar in magnitude. That is, in the SDC framework, a smoothness should exist among the cells. This is equivalent to an independence assumption in each reduced contingency table defined by the neighboring cells. The parameters of the log-linear model (2) could be estimated by maximising a penalized likelihood function. The penalty function could be derived from the smoothness constraints, see [12].

For simplicity consider a 2-way contingency table with I rows and J columns. Assume that the cross-classifying variables are ordinal. Denote by $\mathcal{L}(\beta)$ the relevant part of the log-likelihood function of (2), $\mathcal{L} = \sum_{k=1}^{K} [f_k \log(\mu_k) - \mu_k]$. Then, the maximum penalized estimator $\hat{\boldsymbol{\mu}}$ is the value of $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_K)^{'}$ that maximizes

$$\mathcal{PL}(\boldsymbol{\beta}) = \mathcal{L}(\boldsymbol{\beta}) - A \sum_{i=1}^{I-1} \sum_{j=1}^{J-1} \left[\log \left(\frac{\mu_{i,j}\mu_{i+1,j+1}}{\mu_{i,j+1}\mu_{i+1,j}} \right) \right]^2 \qquad (5)$$

The function $\mathcal{PL}$ penalizes for missed independence in the reduced 2x2 tables since $\mathcal{PL}$ takes smaller values when the cross-ratios are much greater (smaller) than 1. It is known that the greater the departure of the cross-ratio from 1 is, the greater the departure from independence is. The generalisation of (5) to multi-dimensional tables is straightforward.

A first advantage of the penalized maximum likelihood approach is that for certain choices of A, the existence, uniqueness and consistency of the estimates are proved in [12]. Indications on practical choices of A are also given. Moreover, goodness of fit tests could also be constructed.

Second, the number of parameters to estimate is greatly reduced. The third advantage of the penalized likelihood approach is that the penalty function could be extended to non-ordinal categorical key variables. Indeed, the independence in the reduced tables can be modeled for nominal variables. However, the properties of the resulting estimators should be derived. Finally, the penalized likelihood function $\mathcal{PL}$ could be readily used together with the log-rate models discussed in the previous section.

5 Experiments

Several preliminary experiments were performed using the Italian census data. The selected variables were: *Province, Gender, Age, Marital status* and *Education*. For each chosen province, a stratified sample was selected from the census data by means of a simple random sampling. The stratification variables were *Gender* and *Age* (14 categories). The weights were computed in order to preserve the population totals in each strata. For each province, three different sampling fractions were used. Moreover, the data of the Italian Labour Force Survey (LFS 2001) was also used. For this sample, a two-stage stratified sampling scheme was used. The stratification was derived at *Province* level using also the dimensions of municipalities. For each sampling strata the weights were computed in order to preserve known population total by *Gender* and *Age* (14 classes). In the experiments reported here the household hierarchy was ignored.

Province, Age (14 categories), *Gender, Marital status* (6 categories) and *Education* (6 categories) were considered as key variables in a first testing step. Then, in a second testing step, only *Age* was no more considered as key variable. This might not be a realistic assumption in practical situations. This test was performed in order to assess the behaviour of the proposed estimators when the key variables are not stratifying variables.

The three versions of the log-linear models (unweighted frequencies, weighted frequencies and log-rate models) were applied. The estimations of the number

Table 1. τ_1 estimation when *Age* is a key variable. SaUn = number of sample uniques, NoW = unweighted model, LR = log-rate model, I = independence model, S = saturated model.

Province	π	SaUn	True	NoW(I)	LR (I)	NoW(S)	LR(S)
Asti	0.015	246	6	0.22	17.31	0.70	23.04
Asti	0.150	316	41	0.58	12.42	4.02	30.43
Asti	0.277	307	73	0.73	7.84	5.30	32.46
Asti LFS	0.005	86	4	0.00	12.71	0.00	18.38
Biella	0.017	215	3	0.18	9.45	1.09	19.11
Biella	0.167	291	38	0.80	10.73	5.22	29.79
Biella	0.308	279	73	1.20	7.33	6.64	24.95
Biella LFS	0.005	127	7	0.00	13.02	0.03	21.27
Cuneo	0.006	201	2	0.00	7.87	0.06	12.23
Cuneo	0.056	288	14	0.00	11.07	0.09	18.93
Cuneo	0.105	270	23	0.01	7.08	0.38	19.17
Cuneo LFS	0.003	108	10	0.00	17.47	0.00	23.66
Ferrara	0.009	204	2	0.01	8.38	0.21	15.27
Ferrara	0.090	259	23	0.02	8.65	1.98	23.66
Ferrara	0.166	259	34	0.04	6.08	2.13	21.35
Ferrara LFS	0.003	122	4	0.00	22.24	0.36	26.08
Frosinone	0.006	217	1	0.00	9.97	0.04	15.15
Frosinone	0.064	301	23	0.01	14.13	0.50	22.79
Frosinone	0.119	275	32	0.01	4.31	1.09	19.54
Frosinone LFS	0.003	82	6	0.00	12.35	0.11	15.35
Latina	0.006	241	4	0.00	6.43	0.02	10.51
Latina	0.064	278	11	0.00	11.18	0.26	20.99
Latina	0.118	307	34	0.00	6.29	1.73	19.24
Latina LFS	0.003	103	7	0.00	12.69	0.04	18.38
Novara	0.009	214	1	0.00	10.71	0.09	15.81
Novara	0.091	308	30	0.02	4.85	0.87	24.24
Novara	0.168	290	44	0.02	4.57	1.10	22.67
Novara LFS	0.003	112	8	0.00	14.20	0.35	20.95
Parma	0.008	222	3	0.00	11.09	0.10	18.13
Parma	0.079	273	19	0.00	5.38	1.81	18.22
Parma	0.146	270	34	0.01	5.00	2.20	17.01
Parma LFS	0.003	113	6	0.00	17.51	0.04	21.51
Ravenna	0.009	237	4	0.01	13.84	0.30	20.48
Ravenna	0.089	292	17	0.01	9.14	1.44	24.03
Ravenna	0.165	306	43	0.04	4.13	2.02	19.56
Ravenna LFS	0.003	116	7	0.00	18.36	0.02	28.12
Rieti	0.021	195	7	0.35	10.64	1.07	20.74
Rieti	0.211	288	71	1.47	12.88	6.54	38.65
Rieti	0.391	301	116	2.98	10.19	16.32	45.92
Rieti LFS	0.007	82	2	0.02	7.46	0.51	13.23
Rimini	0.011	225	1	0.02	12.58	0.45	16.91
Rimini	0.114	299	30	0.05	8.78	2.09	33.62
Rimini	0.212	283	58	0.12	5.34	4.20	29.14
Rimini LFS	0.004	86	3	0.00	11.66	0.00	16.38
Verbano	0.020	193	1	0.29	8.08	0.72	16.96
Verbano	0.195	289	50	0.74	16.24	6.11	44.19
Verbano	0.362	296	113	2.21	8.57	9.11	34.65
Verbano LFS	0.006	107	8	0.00	22.42	0.41	26.60
Vercelli	0.018	225	9	0.28	8.58	1.43	14.80
Vercelli	0.176	279	60	0.89	10.13	6.34	30.93
Vercelli	0.327	283	80	1.04	7.19	7.10	28.52
Vercelli LFS	0.005	111	3	0.00	18.95	0.15	23.14
Viterbo	0.006	203	4	0.02	10.51	0.05	11.76
Viterbo	0.064	305	11	0.26	20.99	0.14	8.21
Viterbo	0.118	315	34	1.73	19.24	0.21	6.34
Viterbo LFS	0.003	91	7	0.04	18.38	0.00	14.41

Table 2. τ_1 estimation when *Age* is not a key variable. SaUn = number of sample uniques, NoW = unweighted model, LR = log-rate model, I = independence model, S = saturated model.

Province	π	SaUn	True	NoW(I)	LR (I)	NoW(S)	LR(S)
Asti	0.015	15	0	0.00	0.10	0.00	0.20
Asti	0.150	24	1	0.00	0.00	0.00	0.09
Asti	0.277	11	0	0.00	0.00	0.00	0.00
Asti LFS	0.005	13	1	0.00	3.73	0.00	3.08
Biella	0.017	27	0	0.00	3.01	0.00	2.24
Biella	0.166	14	1	0.00	0.00	0.00	0.01
Biella	0.309	17	3	0.00	0.00	0.00	0.00
Biella LFS	0.005	9	1	0.00	1.96	0.00	2.38
Ferrara	0.009	29	0	0.00	1.89	0.00	1.45
Ferrara	0.089	16	0	0.00	0.00	0.00	0.00
Ferrara	0.166	13	1	0.00	0.00	0.00	0.00
Ferrara LFS	0.003	13	0	0.00	5.68	0.00	5.41
Latina	0.006	28	0	0.00	1.62	0.00	2.24
Latina	0.064	25	0	0.00	0.00	0.00	0.36
Latina	0.118	12	0	0.00	0.00	0.00	0.00
Latina LFS	0.003	15	0	0.00	2.95	0.00	2.94
Vercelli	0.018	29	0	0.00	3.04	0.00	2.64
Vercelli	0.176	18	0	0.00	0.00	0.00	0.01
Vercelli	0.327	17	1	0.00	0.00	0.00	0.00
Vercelli LFS	0.005	12	2	0.00	2.55	0.00	2.31

Table 3. Initial contingency table

5	4	3	3	5	1	2	5
1	2	3	3	6	4	2	5
5	4	8	4	4	4	11	4
15	8	8	6	5	6	4	3
10	1	11	2	4	4	3	9
8	7	9	3	2	1	2	1
8	2	4	5	7	2	1	1
6	4	3	7	1	1	2	1

of sample uniques that are also population uniques were compared to the true values calculated from the census data. The estimated numbers of sample and population uniques obtained are presented in tables 1 and 2, together with the true values of τ_1. In both tables, only a selection of tested provinces is shown; similar results were obtained in all the other cases. The results of the weighted log-linear model are not shown since they were almost always equal to zero. This might be due to the inflation effect induced by the weights. From tables 1 and 2, it may be observed that better results may be obtained using log-rate models. For the smaller sampling fractions, there is an overestimation tendency, while for greater sampling fractions the log-rate model underestimates. It should be noted that when a stratifying variable is not a key variable, the estimates

Table 4. Results of the maximum likelihood estimation approach

6.5	5.9	5.3	4.8	4.3	3.9	3.5	3.2
6.4	5.8	5.2	4.7	4.3	3.9	3.5	3.1
6.4	5.7	5.2	4.7	4.2	3.8	3.4	3.1
6.3	5.6	5.1	4.6	4.1	3.7	3.4	3.0
6.2	5.6	5.0	4.5	4.1	3.7	3.3	3.0
6.1	5.5	4.9	4.5	4.0	3.6	3.3	3.0
6.0	5.4	4.9	4.4	4.0	3.6	3.2	2.9
5.9	5.3	4.8	4.3	3.9	3.5	3.2	2.9

Table 5. Results of the penalized maximum likelihood estimation approach

5.0	4.0	3.1	3.2	5.0	2.1	1.6	5.0
1.8	1.8	3.1	3.2	6.0	3.9	2.9	4.7
4.9	4.6	8.0	4.4	4.4	4.1	11.1	4.0
15.1	8.2	8.0	6.0	5.0	6.0	4.0	3.4
10.1	3.2	5.6	2.2	4.5	4.0	3.2	8.2
8.4	5.8	10.1	3.2	2.3	1.7	1.7	1.3
8.1	2.2	4.4	5.0	6.8	1.7	1.7	1.2
6.0	4.1	3.2	7.0	1.2	1.3	2.2	1.6

obtained using the unweighted log-linear model are always equal to zero. More testing and simulations will be performed in order to assess the properties of the log-rate models in the SDC framework.

The penalized maximum likelihood approach is illustrated by means of a simple numerical example. Table 3 presents an 8 x 8 contingency table derived from 2 ordinal variables; the cells (5, 2) and (7, 7) have both a value equal to 1. The neighbors of the cell (5, 2) have large values; instead, the neighbors of the cell (7, 7) have small values.

Table 4 shows the results obtained by fitting an independence model using the maximum likelihood approach. The table 5 shows the results of fitting the table 3 by maximizing the penalized likelihood function (5). The parameter A was determined by means of the expectation-maximisation algorithm described in [12]. When smoothness is assumed, tables 4 and 5 indicate that the maximisation of a penalized likelihood function might be a valid methodology for the analysis of contingency tables.

6 Conclusion

Two problems related to the estimation of a global risk measure were addressed. First the analysis of contingency tables derived from complex surveys was discussed. A log-rate model using the weights as an offset variable was presented. Unbiasness and validity of goodness of fit tests are two significant characteristics of these models. Promising results are obtained when real data were fitted using this modeling procedure.

In the statistical disclosure control framework, table smoothness is an important issue since the estimation of any risk of re-identification measure might be performed by borrowing information from the neighboring cells. Moreover, the relationships between the cross-classifying variables might determine the number of sample uniques that are also population uniques. A penalized likelihood approach was proposed to deal with smoothness. The penalty function was expressed in terms of independence constraints. Starting from the presented preliminary results, further experiments will be performed.

Acknowledgments. Istat is not responsible for any views or results presented. The author was supported by the Network of Excellence in the European Statistical System in the field of Statistical Disclosure Control (ESSnet-SDC).

References

1. Agresti, A.: Categorical Data Analysis. Wiley, New York (1990)
2. Clogg, C.C., Eliason, S.R.: Some Common Problems in Log-Linear Analysis. Sociological Methods and Research 16, 8–44 (1987)
3. Elamir, E.A.H.: Analysis of Re-Identification Risk Based on Log-Linear Models. In: Domingo-Ferrer, J., Torra, V. (eds.) PSD 2004. LNCS, vol. 3050, pp. 273–281. Springer, Heidelberg (2004)
4. Fan, J., Gijbels, I.: Local Polynomial Modelling and its Applications. Chapman & Hall, London (1996)
5. Fienberg, S.E., Holland, P.W.: On the Choice of Flattening Constants for Estimating Multinomial Probabilities. Journal of Multivariate Analysis 2, 127–134 (1972)
6. Haberman, S.J.: Analysis of Qualitative Data. New Developments, vol. 2. Academic Press, New York (1979)
7. Lohr, S.L.: Sampling: Design and Analysis. Duxbury Press (1999)
8. Polettini, S.: Some Remarks on the Individual Risk Methodology. Monographs of Official Statistics. In: Work Session on Statistical Data Confidentiality. European Comission (2003)
9. Rao, J.N.K., Thomas, D.R.: The Analysis of Cross-Classified Categorical Data from Complex Surveys. Sociological Methodology 18, 213–269 (1988)
10. Rinott, Y., Shlomo, N.: A Smoothing Model for Sample Disclosure Risk Estimation. In: Tomography, Networks and Beyond. IMS Lecture Notes-Monograph Series Complex Datasets and Inverse Problems, vol. 54, pp. 161–171 (2007)
11. Skinner, C.J., Shlomo, N.: Assessing Identification Risk in Survey Micro-data Using Log Linear Models. Journal of the American Statistical Association, Applications and Case Studies (forthcoming)
12. Simonoff, J.S.: A Penalty Function Approach to Smoothing Large Sparse Contingency Tables. The Annals of Statistics 11, 208–218 (1983)
13. Skinner, C., Holmes, D.: Estimating The Re-Identification Risk per Record in Microdata. J. Official Statistics 14, 361–372 (1998)
14. Willenborg, L., De Waal, T.: Elements of Disclosure Control. Lecture Notes in Statistics, vol. 155. Springer, Berlin (2001)

Use of Auxiliary Information in Risk Estimation

Loredana Di Consiglio[1] and Silvia Polettini[2]

[1] ISTAT
Servizio Progettazione e Supporto Metodologico
nei Processi di Produzione Statistica
Via Cesare Balbo 16, 00184 Roma, Italy
diconsig@istat.it

[2] Dipartimento di Scienze Statistiche
Università degli Studi di Napoli Federico II
Via L. Rodinò 22 – 80128 Napoli, Italy
spolettini@unina.it

Abstract. In the release of microdata files, reidentification of a record implies disclosure of the values of a possibly large set of sensitive variables. When microdata files are released by statistical Agencies, a careful assessment of the associated disclosure risk is therefore required.

In order for an informed decision to be made, maximising accuracy and precision of the risk estimators is crucial. Clearly such characteristics will affect the risk assessment process and Agencies should choose the estimator that performs best. In fact, estimators may perform poorly, especially for those records whose real risk is higher. To improve estimation, we propose to introduce external information, arising from a previous census as is done in the context of small area estimation (see [10]). In [4] we considered SPREE - type estimators that use the association structure observed at a previous census (see [9]); in this paper we consider models that use the structure of a population contingency table while allowing for smooth variation of the latter. To assess the statistical properties of this estimator and compare it with alternative approaches, we show results of a simulation study that is based on a complex sampling scheme, typical of most households surveys in Italy. Comparison is made with a simple SPREE estimator and a Skinner-type estimator [13,6], applied to a complex sampling scheme.

Keywords: Disclosure, generalized linear model, per record risk, SPREE, simulation study.

1 Introduction

In the release of both microdata files and tables, a major concern in disclosure limitation is to avoid record reidentification. Even if data arise from a sample survey, it is always possible that an intruder, using information from other sources and the published data, might link one or more of the released records to one

J. Domingo-Ferrer and Y. Saygın (Eds.): PSD 2008, LNCS 5262, pp. 213–226, 2008.

or more units in the population. Note that reidentifying a record in microdata implies disclosing the values of a possibly large set of sensitive variables.

For microdata, the disclosure scenario allows for the existence of external information that permits record reidentification. This information usually consists of publicly available variables (called key or identifying variables) known for the population and also present in the file to be released. The reidentification risk is defined as the probability that a *correct* link between a specific record in the population and a record in the sample is established. Under this definition risk measures depend on frequency of cells in the contingency table built by cross tabulating the key variables. The problem of estimating disclosure risk has thus parallels with that of small area estimation of counts in cross-classifications. In that context, besides borrowing strength from neighbouring areas or cells, auxiliary information, derived from external sources such as administrative registers or a census, is exploited. We expect that improvements in risk assessment can be achieved by introducing external information, especially when estimates of risk are required for low sample cell frequencies, i.e. sample uniques, doubles, etc.. For population surveys, external information about the association structure useful for risk estimation can be obtained from contingency tables built on previous census data; updated information for at least some margins of the contingency table induced by the disclosure scenario is also available. For instance, current population counts for region, sex and age classes are available from public registers; for other classifications, it might be the case that design based calibrated estimators at the national level give sufficiently accurate figures for the population. For estimation of counts from a contingency table, the so called structure preserving estimator (SPREE,[9]) makes precise use of the above mentioned information. As a generalization of the SPREE considered in [4], in this paper we analyse the generalized linear structural model (GLSM) proposed by Zhang and Chambers [15] to estimate population cell frequencies. These methods are described in Sect. 3.

Both the SPREE and the GLSM might produce estimates for population frequencies that are lower than their sample counterparts. Based on the generalized iterative proportional fitting (GIPF, [5]), we propose a modification of both methods that allows for this logical constraint and produces population estimates that are always higher than the corresponding sample frequencies, while considering the structure of association of the table and its marginal constraints.

In Sect. 5 we then propose simple risk estimators based on the reciprocal of the estimated frequency, including for comparison a Skinner-type estimator for sample uniques that relies on a superpopulation model defined in Sect. 4. In Sect. 7 we analyse the performance of these estimators by simulation. Having a clear idea of the accuracy and precision of the risk estimators is indeed crucial for risk assessment and subsequent data protection. Research has however focused mainly on the definition of models for risk estimation, quantification of uncertainty of estimates having received only minor attention. An approach to tackle this problem is proposed by Rinott and Shlomo [11] in a specific setting.

Our simulation study is described in Sect. 6 and consists of 1,000 pseudo-samples drawn from the population collected at the 2001 Italian population census, using the previous one (carried out in 1991) as a source of auxiliary information. Specifically, the auxiliary information refers to the structure of association in the 1991 census contingency table based on the variables that we use to estimate the risk. We also use available information about the margins of the above mentioned contingency table at current time, consisting in standard design-unbiased estimators of population cell frequencies and in population counts that are usually available from administrative sources.

2 Estimating the Reidentification Risk for Microdata

Consider a disclosure scenario defining q categorical key variables, denoted by $Z_1, \ldots, Z_q$, with $C_1, \ldots, C_q$ categories respectively. This scenario is appropriate for most population surveys, where identification can be based on variables such as place of residence, sex and age. Records with the same key values are identical for reidentification and should have the same risk of disclosure. Cross-classification of the key variables generates a contingency table with a total number of $K = \prod_k C_k$ cells at both the population and the sample level; cell frequencies in the population and sample table, respectively, are denoted by F_k and f_k. Intuitively, rare traits in the population are the ones that could lead to disclosure, but to be exposed to disclosure risk such rare records should also be included in the sample. The problem is therefore discriminating between the sample cells that are structurally small in the population and those that are small in the microdata only because of sampling; direct or indirect inference about the corresponding population size is required for these cells.

A large part of the literature has focused on estimating measures based on the frequency of sample unique cells that are also population unique (see [1,7,12]). These quantities can be used as global risk measures for the microdata file. However it is also important to be able to assess the disclosure risk associated with the release of individual records; if the population contingency table were known, a simple risk measure for each record in cell k of the sample table could be defined using the corresponding population cell size, $r_k = 1/F_k$. As F_k is unknown, the above definition is not usable. A solution is to specify a statistical model for $\boldsymbol{F} = (F_1, \ldots, F_k)$ and derive suitable risk estimates, such as $\mathrm{E}(1/F_k|\boldsymbol{f})$.

The estimation of risk for low count cells is a challenging problem and to obtain more accurate and precise estimators, all the available information should be used. Typical applications consider very large and sparse contingency tables, often with logical constraints inducing structural zeros. Estimating F_k is particularly difficult for high risk cells, having low sample and population sizes. Finite population theory cannot account for all the information about the population structure and would produce unreliable estimates, in particular when the sampling fraction is small. In the next section we present estimation methods that introduce auxiliary information for the estimation of counts in a cross-classification.

3 Small Area Estimators for Cross-Classifications

Small area methods allow for external information by introducing explicit or implicit models for the relationship between the variable of interest and the auxiliary variables; the definition of disclosure risk in terms of cells of a certain contingency table suggests using the structure preserving estimators (SPREE, [9]). For tabular data arising from cross-classification of categorical variables, Purcell and Kish [9] propose that the external information obtained from an administrative or a census source can be exploited to improve the estimation of counts. The *association structure* completely describing the relationship among variables is derived from a supplementary table that is observed at a previous time $t_0 = t - L$. This association structure is then updated using current information at time t on the (partial) association between the variables present in the *allocation structure* m. The *allocation structure* usually consists of margins of the current frequency table; typically, counts on classes defined by sex and age can be obtained by administrative records, so that these are often used to define the allocation structure. Moreover, reliable survey estimates can be obtained when aggregating over geography; these represent an additional information to be used when updating the association among variables.

In [4] we have proposed SPREE based risk estimators. SPREE can exhibit large bias if the association structure was subjected to a significant alteration over time; to permit additional flexibility in the association structure, Zhang and Chambers [15] introduce a class of log-linear structural models for the cross-classification which generalizes SPREE by introducing linear models on the parameters defining the interactions among variables.

In this paper we propose a simple risk estimator based on the GSLM methodology for estimation of population counts F_k.

Section 3.1 contains a brief summary of the SPREE method, whereas Sect. 3.2 presents the GLSM.

3.1 The Structure Preserving Estimator

Let us consider a three-way table; note that any multi-way table can be reduced to three-way by properly re-defining the classification.

Let d denote the geographical or administrative domain, h the classification given by the auxiliary variables (sex and age in our application) and let i be the classification given by the other key variables.

Let X_{dhi} be the association structure, i.e. the table completely observed at previous time $t_0 = t - L$. Finally, define by F_{dhi} the counts of the current contingency table to be estimated and by m the allocation structure, i.e. the updated margins.

In its original formulation, SPREE consists in adjusting the X_{dhi} to agree with the updated information in m, while preserving the relationships among variables present in X_{dhi} as much as possible. The aim is to obtain estimates of the current counts F_{dhi} that minimize the χ^2 distance between X_{dhi} and F_{dhi} with constraints given by m. As mentioned in [9], explicit solutions only exist in trivial cases. In general, Iterative Proportional Fitting (IPF), which consists in iteratively adjusting to marginal constraints until convergence, is applied to obtain an approximate solution, denoted by $\hat{F}_{dhi}^{\text{SPREE}}$, to the above optimization problem.

The IPF on X_{dhi} may produce estimates of the current cell counts that are lower than the observed counts f_{dhi}; to overcome this inconsistency, we propose to apply the generalized iterative proportional fitting (GIPF, [5]) instead of IPF. In general, GIPF allows to obtain solution of a minimisation problem under convex constraints, and can be easily applied with the constraints we have imposed. This strategy clearly differs from simply equating estimates and sample frequencies for the inconsistent cells.

Depending on the information available, SPREE allows different specifications of the allocation structure m. Here we consider the specification used in our application (see Sect. 6), namely a pair of bivariate marginal tables: $m = (\{\hat{F}_{.hi}\}, \{F_{dh.}\})$, where $\hat{F}_{.hi}$ are design based estimates and $F_{dh.}$ come from administrative registers.

The structure preserving estimator is shown (see [9]) to preserve all the interactions of X_{dhi} but those redefined by the allocation structure, so that the higher order interactions of F_{dhi} are set equal to that of X_{dhi}; the bias of $\hat{F}_{dhi}^{\text{SPREE}}$ therefore depends on the extent to which the equality of the interactions holds for the data. For further details on SPREE see [9] and [15]. Note that with respect to the Purcell and Kish estimator, the introduction of the additional constraints $\hat{F}_{dhi}^{\text{SPREE}} > f_{dhi}, d = 1 \ldots D, h = 1 \ldots H, i = 1 \ldots I$ in the allocation step is expected to induce slight modifications in the association structure.

3.2 The Generalized Linear Structural Model

Under the same setting and for the same estimation problem of Sect. 3.1, [15] show that the SPREE is a special case of a generalized linear structural model for the domain proportions. As before we consider the case that updated margins $m = (\{\hat{F}_{.hi}\}, \{F_{dh.}\})$ are available for the current population table.

Using the notation introduced in the previous section, consider the within-domain proportions θ_{dhi}^{F} and θ_{dhi}^{X}, relative to the target population table and the auxiliary table at time t_0, respectively:

$$\theta_{dhi}^{F} = \frac{F_{dhi}}{F_{dh.}}, \qquad \sum_i \theta_{dhi}^{F} = 1,$$

θ_{dhi}^{X} being defined similarly.

Define now the saturated log-linear representation of the population counts:

$$log(F_{dhi}) = log(\theta_{dhi}^{F}) + log(F_{dh}) = \alpha_0^F + \alpha_d^F + \alpha_h^F + \alpha_i^F + \alpha_{dh}^F + \alpha_{di}^F + \alpha_{hi}^F + \alpha_{dhi}^F$$

and of the auxiliary complete table:

$$log(X_{dhi}) = log(\theta_{dhi}^{X}) + log(X_{dh}) = \alpha_0^X + \alpha_d^X + \alpha_h^X + \alpha_i^X + \alpha_{dh}^X + \alpha_{di}^X + \alpha_{hi}^X + \alpha_{dhi}^X \ .$$

Let

$$\mu^F_{dhi} = log(\theta^F_{dhi}) - \frac{1}{I}\sum_i log(\theta^F_{dhi}) = \alpha^F_i + \alpha^F_{di} + \alpha^F_{hi} + \alpha^F_{dhi} \;, \tag{1}$$

$$\mu^X_{dhi} = log(\theta^X_{dhi}) - \frac{1}{I}\sum_i log(\theta^X_{dhi}) = \alpha^X_i + \alpha^X_{di} + \alpha^X_{hi} + \alpha^X_{dhi} \;.$$

A generalized linear model is introduced to link the two structural models above to inform F_{dhi} through the known complete table X_{dhi}. With a three-way table, two different saturated models may be proposed. The first one assumes a linear structure with constant regression coefficient among strata:

$$\mu^F_{dhi} = \lambda_{hi} + \beta\, \mu^X_{dhi} \tag{2}$$

where $\sum_i \lambda_{hi} = 0$. This model is a proportional interaction model, and SPREE is a special case of the latter for $\beta = 1$; see [15] for a more detailed description of the implications of the models.

If the regression coefficients are allowed to vary among strata the model is

$$\mu^F_{dhi} = \lambda_{hi} + \beta_h\, \mu^X_{dhi} \tag{3}$$

and is equivalent to a stratified proportional interaction model

$$\begin{aligned} \alpha^F_{i;h} &= \lambda_{hi} + \beta_h\, \alpha^X_{i;h} \;, \\ \alpha^F_{di;h} &= \beta_h\, \alpha^X_{di;h} \;. \end{aligned}$$

Models (2) and (3) represent two examples of the so called generalized linear structural model. Note that both refer to population quantities. For this reason unbiased estimates of cell proportions θ^F_{dhi} are introduced in (1). The presence of a complex sampling scheme is accounted for by selecting appropriate design based estimators.

Following standard techniques for generalized linear models, the model parameters are estimated by iterative weighted least squares; here the weighting matrix is designed to include the covariance matrix of the direct estimators above; see [15] for details.

The procedure outlined produces first-step estimates of the population counts, that we denote by $\tilde{F}_{dhi}$. Note that under SPREE the first-step estimate is just $\tilde{F}_{dhi} = X_{dhi}$, i.e. the previous table without adjusting for the observed data.

Recalling that updated margins $m = (\{\hat{F}_{.hi}\}, \{F_{dh.}\})$ are available, exactly as described for the SPREE methods, the first-step estimates can be adjusted to match with m by IPF to obtain the final estimates $\hat{F}^{\text{GLSM}}_{dhi}$. Here again use of IPF could produce estimated population cell counts that are lower than the observed sample counts; for this reason we propose to modify the second estimating step by introducing the addition constraints $\hat{F}^{\text{GLSM}}_{dhi} > f_{dhi}, d = 1 \ldots D, h = 1 \ldots H, i = 1 \ldots I$; the first step estimates are therefore adjusted to auxiliary marginal tables by means of GIPF to ensure consistency with the sample frequencies.

4 Risk Estimation Based on Loglinear Models

To gather information on small cells from larger ones, not relying on external information, superpopulation models can be introduced. Skinner and Holmes [13] use the structure of the table through a loglinear model which is fitted to the data. They focus on sample uniques and model the F_{dhi}, $d = 1 \dots D, h = 1 \dots H, i = 1 \dots I$, as Poisson r.v. with mean π_{dhi}. The mechanism generating $\boldsymbol{f}$ is assumed to be a Bernoulli sampling with known probability p, so that $f_{dhi}|\lambda_{dhi} \sim$ Poisson$(p\pi_{dhi})$; finally a loglinear model for the expected sample frequencies $p\pi_{dhi}$ (where p only produces an offset term) is defined, from which the π_ks are estimated. The superpopulation model implies that $F_{dhi} - f_{dhi}|f_{dhi} \sim$ Poisson$((1-p)\pi_{dhi})$ and this relation allows one to estimate the risk as the expected value of $1/F_{dhi}$ given sample uniqueness

$$\mathrm{E}\left(1/F_{dhi}|f_{dhi} = 1\right) = \frac{1 - \exp\{-(1-p)\hat{\pi}_{dhi}\}}{(1-p)\hat{\pi}_{dhi}}, \tag{4}$$

by plugging in the loglinear estimates. The model defined in [13] has greater generality, as it also specifies a lognormal distribution for the loglinear parameters π_{dhi}, $\log(\pi_{dhi}) = \mu_{dhi} + \epsilon_{dhi}$, $\epsilon_{dhi} \sim N(0, \sigma^2)$, $\exp(\mu_{dhi})$ being the expected frequencies in the loglinear model assumed for the population, to account for overdispersion due to lack of fit. Practical application of the method however has dropped this assumption, partly because of possible negativity of empirical Bayes estimates of the lognormal variance.

The approach relies on a single, good fitting loglinear model. Forster and Webb [8] propose a related, fully Bayesian approach incorporating model uncertainty through model averaging. Skinner and Shlomo [14] propose a model search strategy explicitly targeted to risk estimation under the framework defined in [13]. Because of the associated computational cost, our simulation study does not apply the latter procedure; note however that the application described in the paper considers a three-way representation of the contingency tables, and the number of models to search would not be large in that case.

5 Risk Estimators

Having observed that the estimand is $r_{dhi} = 1/F_{dhi}$ for nonempty sample cells, our first proposal simply estimates r_k by

$$\hat{r}_{dhi}^{\mathrm{SPREE}} = 1/\hat{F}_{dhi}^{\mathrm{SPREE}} \quad . \tag{5}$$

For the generalized linear structural model we use simple estimators analogous to (5); model (2) is not analysed in this paper, see Sect. 7 for further details. We restrict attention to the stratified model (3) to define

$$\hat{r}_{dhi}^{\mathrm{GLSM}} = 1/\hat{F}_{dhi}^{\mathrm{GLSM}} \quad . \tag{6}$$

Finally, we consider the risk estimator described in Sect. 4, namely

$$\hat{r}_{dhi}^{\mathrm{SK}} = \frac{1 - \exp\{-(1-p)\hat{\pi}_{dhi}\}}{(1-p)\hat{\pi}_{dhi}} \quad . \tag{7}$$

6 Simulation Plan

We performed a simulation study consisting of 1,000 synthetic samples drawn from a known real population, namely the population registered at the 2001 Italian Census for 6 Italian regions (Val d'Aosta, Piemonte, Toscana, Umbria, Campania, Molise). Samples were drawn using the strategy of Labour Force Survey (LFS), as detailed in [4]. Note that the LFS sampling design is used for most Italian social surveys. The availability of the target population from which our samples were extracted allows us to assess the performance of the estimators, as clearly the estimand is known and equals $1/F_k$ for each cell k.

The six regions above were selected in light of their different geographical position (North, Center and South), the differences they exhibit in the distribution of the key variables, their variability in the number of inhabitants (Val d'Aosta and Molise are small regions where we expect higher risks of disclosure) and finally the substantial variation of their sampling rates. The latter characteristic results from sample size being planned to guarantee a target precision level of LFS estimates. The LFS is based on a complex sample design with stratification of municipalities. In each sample municipality, a systematic sample of households is selected; each member of sampled households is included in the LFS sample.

In year 2001 the population of the six regions amounted to over 15 millions; the effective sample size in terms of individuals results in over 80,000 records.

LFS estimates use sampling design weights obtained by a calibration process that controls over known totals of sex and ages (see [3]). Although the actual calibration process is more complex, for simplicity we have calibrated only on sex by age at regional level.

The key variables selected are region of residence (6 classes as described above), sex, age (in 20 classes), marital status (in 4 classes), education (in 5 classes). As the cell of the cross-tabulation is not a planned domain for LFS, we expect that especially the cells with smaller population counts, i.e. higher risk of identification if selected in the sample, will be present in a small subset of the universe of all samples.

For the 1,000 simulated samples the average percentage of sample unique cells was about 9% (overall, between 8 and 10%), with over 45% of empty cells (overall, between 44 and 47%).

The estimators (5) and (6) described in this paper employ the association structure at a previous time. Complete information on it is available from the census conducted in year 1991. The temporal lag is large, but we can study the performance of the method almost in its worst condition since we expect that the stability in the association structure decreases with time. The estimators also make use of available information on the margins of the above mentioned contingency table at current time. In the terminology of Sect. 3.1, the allocation structure has been defined as $m = (\{\hat{F}_{.hi}\}, \{F_{dh.}\})$. In our application $F_{dh.}$ represents the 2001 census counts of the marginal table defined by sex by age (classes indicated with h) by region (classes indicated with d). In practice however these counts would come from updated administrative sources. On the other hand, the counts $F_{.hi}$ of the marginal cross-table defined by education by marital status (classes denoted with i) by sex by age (classes denoted with h) are unknown; in our application we resort to calibration estimates $\hat{F}_{.hi}$. Increasing the number or detail of the key variables necessarily affects the precision of these estimators; variable age was indeed recoded to five year classes to limit the variability of direct estimators. The expected effect of increasing the number of cells is in terms of increasing in turn the variability of our risk estimates. We plan to study this in detail to draw more precise conclusions. The practical importance of this limitation is dictated by the data release strategy: the key variables should enter exactly as they appear in the released file. In fact, age is preferably released in one year classes, a classification that might be too fine to allow precise direct estimation of margins. In this case the problem of getting sufficiently precise estimates of marginal counts might be addressed by further modelling within the proportional fitting procedure, an issue that we plan to address in the future.

7 Results

We discuss and compare the estimators listed in Sect. 5. We discarded the estimator based on model (2) for two reasons. First, the associated computational burden was heavy; second, a preliminary inspection of the population parameters (that are known in our simulation) revealed a certain variability among strata defined by age by sex, so that model (3) was judged more appropriate.

As regards estimator (7), its formulation requires uniform selection probabilities, an assumption that in our sampling design is met at a geographical detail finer than the the regional one to which we refer. To adjust to our framework, for each region d the mean sampling fraction p_d computed over the corresponding municipalities was specified in model (4). A simple analytic expression for the risk is available for sample uniques only; for simplicity results are restricted to this case. Finally, we fitted to the sample data the loglinear model having as sufficient statistics the margins present in the allocation structure m.

Evaluation of the performance of risk estimators relies on bias and relative root mean square error. There are two sources of variation in this assessment: the sample cell size and the population cell size. We consider a conditional assessment, by analysing the above measures for low, fixed sample frequency ($f_k = 1, 2$). The assessment is clearly restricted to samples where the cell has been observed, as the risk is, of course, not defined (and not of interest) when sample cell is empty. By consequence, for the smallest cells, particularly for regions with lower sampling rates, the performance criteria have sometimes been evaluated on a very small number of samples. In this case, conclusions must be drawn with due care but can still be useful to outline the expected pattern.

Figure 1 reports a graphical assessment of the performance of the three estimators over sample uniques for all the available samples. To avoid plotting all the replications of sample uniques across all simulations, for each cell we plot a summary of the estimates over our 1,000 samples, showing the minimum and maximum (grey dots) and the median (black dots) of the estimates over the eligible samples. Figure 1 shows that in general all the estimators permit to

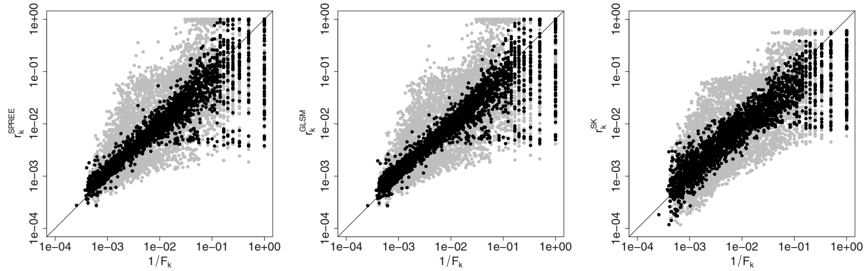


Fig. 1. Performance of the estimators for sample uniques over all the simulated samples. Left to right: SPREE-type estimator $\hat{r}_{dhi}^{\text{SPREE}}$; GLSM-type estimator $\hat{r}_{dhi}^{\text{GLSM}}$ with stratification; Skinner-type estimator $\hat{r}_{dhi}^{\text{SK}}$. Per cell minimum and maximum (grey dots) and median (black dots) of the estimates over all available samples are plotted.

distinguish sample uniques between risky and safe, the first two exhibiting a very similar beahviour, with the GLSM estimator being slightly less variable, and apparently preferable to the third estimator. This is not surprising, as $\hat{r}_{dhi}^{\text{SK}}$ does not make use of auxiliary information.

Table 1 shows a conditional assessment of bias for sample unique cells. Overall, the Skinner-type estimator exhibits underestimation of the true risk, the other estimators showing a less marked bias on average. This is however the result of an assessment over *different population cell sizes*, that is, different risk levels. Figure 2 in the Appendix indicates more clearly the underestimating pattern of all the estimators for very low population frequencies. Such underestimation of extremely high risks seems more severe and persistent for $\hat{r}_{dhi}^{\text{SK}}$. With respect to the Skinner-type estimator, the uppermost panels show a higher variability, a result of structural models gathering information from an auxiliary population contingency table, a feature that is not shared by estimator (7).

Table 1. Conditional assessment of bias for sample unique cells; mean number of samples over which the assessment was conducted: 176.5

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
$\hat{r}_{dhi}^{\text{SPREE}}$	−0.99630	−0.00047	0.00045	−0.00871	0.00283	0.85710
$\hat{r}_{dhi}^{\text{GLSM}}$	−0.99520	−0.00097	0.00037	−0.01876	0.00205	0.73310
$\hat{r}_{dhi}^{\text{SK}}$	−0.99190	−0.00738	0.00007	−0.04404	0.00178	0.30840

As shown in Tab. 2, the bias of the first two estimators is already reduced for sample doubles. Conditioning also on population cell frequency (figure not shown here) we noticed the same underestimation pattern as before for the highest risk cells ($F_{dhi} \leq 3$), that remain difficult to estimate. In this case however the bias is lower than that observed for sample uniques, always below 0.2 in absolute value.

To limit the computational burden, our study only evaluates the variability of the estimators across the simulated samples; in particular we consider the

Table 2. Conditional assessment of bias for sample doubles; mean number of samples over which the assessment was conducted: 119.60

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
$\hat{r}_{dhi}^{\text{SPREE}}$	−0.43710	−0.00120	0.00016	0.00233	0.00089	0.37700
$\hat{r}_{dhi}^{\text{GLSM}}$	−0.44220	−0.00144	0.00012	−0.00058	0.00074	0.29810

relative root MSE (RRMSE). In practical applications, the variability of the estimators could be assessed by bootstrap; model-based bootstrap as suggested in [2] (sect. C6) seems particularly suited to the framework of GLSM-type risk estimators.

The figures for the RRMSE (Table 3) indicate that the GSLM based estimator with stratum specific structure coefficients is preferable to the other estimators.

Table 3. Conditional assessment of RRMSE for sample uniques and doubles; mean number of samples over which the assessment was conducted: 176.5 and 119.6

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
$f_{dhi} = 1$						
$\hat{r}_{dhi}^{\text{SPREE}}$	0.00000	0.36750	0.62480	0.85280	0.98570	17.09000
$\hat{r}_{dhi}^{\text{GLSM}}$	0.00000	0.37080	0.61280	0.82170	0.95200	11.04000
$\hat{r}_{dhi}^{\text{SK}}$	0.01756	0.44750	0.63860	0.91830	1.04500	7.39100
$f_{dhi} = 2$						
$\hat{r}_{dhi}^{\text{SPREE}}$	0.00000	0.31840	0.46990	0.58330	0.70550	6.20600
$\hat{r}_{dhi}^{\text{GLSM}}$	0.00000	0.31420	0.45910	0.56080	0.68330	4.68100

8 Comments

In this paper we presented a comparative analysis of some simple risk estimators based on a linear structural method that is designed to estimate frequency in a population contingency table using auxiliary information. We considered the structure preserving estimator (SPREE) and a generalization of the latter, namely the generalized linear structural model (GLSM) estimator. In both cases we have modified the estimation process so as to ensure that the observed cell frequency f_{dhi} does not exceed the corresponding estimated population frequency. Once the frequency has been estimated, we simply derived the risk estimate as the reciprocal of the estimated population count. We also considered for comparison a Skinner-type estimator based on a Poisson superpopulation model with loglinear modelling of the observed counts.

Results shown are restricted to the most challenging case of cells with low sample frequencies. The simulation experiment conducted clearly indicates that all the estimators tends to underestimate very high risk records. The Skinner type estimator, not relying on detailed information on the population table at a previous time nor on any other external information, shows a more evident tendency to a negative bias; the two other estimators benefit from external information, even though in the smallest regions, characterized by the highest sampling fractions, some cells are not well captured by the model. With respect to the SPREE, the GLSM estimator is preferable in terms of MSE.

Whereas the estimator (7) only requires the observed data and information about the sampling fraction, the structural estimators (5) and (6) come with some computational and administrative burden, as they require an estimation process (especially the GLSM estimator) and management of census data. The process of building the appropriate table is an important step that requires at least some insights about the available information and the classes in which estimates with sufficient precision can be obtained from the sample at hand. The population table from which the association structure is borrowed must be properly organized; besides that, margins must be computed from the available sources such as administrative archives and the sample on release. Finally, in order for the variables collected at a census to be compatible with the key variables available in the survey microdata, treatments, such as recoding, are usually necessary, as sometimes the definitions may vary. This process is nontrivial and might be computationally demanding, depending on the size of the population. An advantage is that the census table has to be collected and organised only once in several years. Indeed the same association structure is modelled at subsequent releases, the only change being the update of margins.

In the context outlined, the GLSM based estimator (6) emerges as a more accurate and precise estimator. This comes at the costs mentioned above. Loglinear model estimation required for the estimator (7) relies on maximum likelihood; although the procedure is well known and readily available in standard software, when large tables are analysed, the associated computational costs may also be high.

Acknowledgments. The empirical study was conducted under a joint research agreement between Istat and Università di Napoli Federico II (Protocollo di ricerca per la collaborazione alla ricerca e verifica sperimentale relativa ai metodi statistici utili per la valutazione del rischio di violazione della riservatezza).

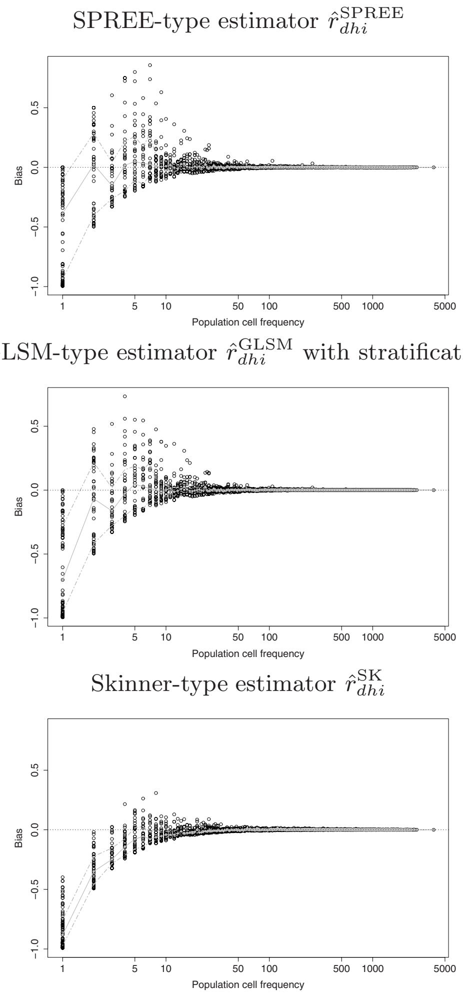
The authors wish to thank two anonymous referees for comments that greatly improved the paper.

References

1. Chen, G., Keller-McNulty, S.: Estimation of identification disclosure risk in microdata. Journal of Official Statistics 14, 79–95 (1998)
2. EURAREA Consortium. Project Reference vol. 1 (2004), https://www.statistics.gov.uk/eurarea
3. Deville, J.C., Särndal, C.E.: Calibration estimators in survey sampling. Journal of the American Statistical Association 87, 367–382 (1992)

4. Di Consiglio, L., Polettini, S.: Improving individual risk estimators. In: Domingo-Ferrer, J., Franconi, L. (eds.) PSD 2006. LNCS, vol. 4302, pp. 243–256. Springer, Heidelberg (2006)
5. Dykstra, R.L.: An iterative procedure for obtaining i-projections onto the intersection of convex sets. The Annals of Probability 13, 975–984 (1985)
6. Elamir, E.A.H., Skinner, C.J.: Record level measures of disclosure risk for survey microdata. Journal of Official Statistics 22(3), 525–539 (2006)
7. Fienberg, S.E., Makov, U.E.: Confidentiality, uniqueness, and disclosure limitation for categorical data. Journal of Official Statistics 14, 385–397 (1998)
8. Forster, J.J., Webb, E.L.: Bayesian disclosure risk assessment: predicting small frequencies in contingency tables. Journal of the Royal Statistical Society: Series C 56(5), 551–570 (2007)
9. Purcell, N.J., Kish, L.: Postcensal estimates for local areas (small domains). International Statistical Review 48, 3–18 (1980)
10. Rao, J.N.K.: Small area estimation. John Wiley & Sons, Hoboken (2003)
11. Rinott, Y., Shlomo, N.: Variances and confidence intervals for sample disclosure risk measures. In: Proceedings of the 56th Session of the ISI, Lisbon, August 22-29, 2007 (2007)
12. Skinner, C.J., Elliot, M.J.: A measure of disclosure risk for microdata. Journal of the Royal Statistical Society, Series B 64, 855–867 (2002)
13. Skinner, C.J., Holmes, D.J.: Estimating the re-identification risk per record in microdata. Journal of Official Statistics 14, 361–372 (1998)
14. Skinner, C.J., Shlomo, N.: Assessing identification risk in survey micro-data using log linear models. Technical Report 14, S3RI Methodology Working Papers Series (2006), http://eprints.soton.ac.uk/41842/01/s3ri-workingpaper-m06-14.pdf
15. Zhang, L., Chambers, R.L.: Small area estimates for cross-classifications. Journal of the Royal Statistical Society, Series B 66(2), 479–496 (2004)

Appendix



SPREE-type estimator $\hat{r}_{dhi}^{\mathrm{SPREE}}$

GLSM-type estimator $\hat{r}_{dhi}^{\mathrm{GLSM}}$ with stratification

Skinner-type estimator $\hat{r}_{dhi}^{\mathrm{SK}}$

Fig. 2. Plot of bias vs population cell frequencies for sample unique cells. Grey broken lines represent 25% and 75% percentiles of bias for given population cell size, grey dots joined by a solid line represent medians of the same quantities.

Accounting for Intruder Uncertainty Due to Sampling When Estimating Identification Disclosure Risks in Partially Synthetic Data

Jörg Drechsler[1] and Jerome P. Reiter[2]

[1] Institute for Employment Research, 90478 Nuremberg Germany
[2] Duke University, Durham NC 27708, USA

Abstract. Partially synthetic data comprise the units originally surveyed with some collected values, such as sensitive values at high risk of disclosure or values of key identifiers, replaced with multiple draws from statistical models. Because the original records remain on the file, intruders may be able to link those records to external databases, even though values are synthesized. We illustrate how statistical agencies can evaluate the risks of identification disclosures before releasing such data. We compute risk measures when intruders know who is in the sample and when the intruders do not know who is in the sample. We use classification and regression trees to synthesize data from the U.S. Current Population Survey.

Keywords: CART, Disclosure, Risk, Synthetic data.

1 Introduction

Several national statistical agencies disseminate multiply-imputed, partially synthetic data to the public. These comprise the units originally surveyed with only some collected values replaced with multiple imputations [1,2]. For example, in the Survey of Consumer Finances, the U.S. Federal Reserve Board replaces monetary values at high disclosure risk with multiple imputations, releasing a mixture of imputed values and the not replaced, collected values [3]. The U.S. Bureau of the Census protects data in the Survey of Income and Program Participation [4] and in longitudinal business databases [5,6] by replacing all values of sensitive variables with multiple imputations, leaving non-sensitive variables at their actual values. They also have created synthesized origin-destination matrices, i.e. where people live and work, available to the public as maps via the web (On The Map, http://lehdmap.did.census.gov/). They plan to protect the identities of people in group quarters (e.g., prisons, shelters) in the American Communities Survey by replacing quasi-identifiers for records at high disclosure risk with imputations. Partially synthetic, public use data are being developed for the Longitudinal Business Database, the Longitudinal Employer-Household Dynamics survey, and the American Communities Survey veterans and full sample data. Other examples of partially synthetic data are in [7,8,9,10].

J. Domingo-Ferrer and Y. Saygın (Eds.): PSD 2008, LNCS 5262, pp. 227–238, 2008.

Because the original records remain on the file, intruders may be able to link those records to external databases, even though values are synthesized. It is prudent for agencies to assess the risks of such identification disclosures before releasing the file. When they are too high, additional synthesis or some other action is needed before release. In this article, we illustrate how to compute risks of identification disclosure for partially synthetic data using a subset of the U. S. Current Population Survey. We show how to incorporate intruders' uncertainty about which records are in the sample and how to assess different synthesis strategies. We also illustrate an application of classification and regression tree methodology for generating partially synthetic data.

2 Review of Partially Synthetic Data

The agency constructs partially synthetic datasets based on the s records in the observed data, D_{obs}, in a two-part process. First, the agency selects the values from the observed data that will be replaced with imputations. Second, the agency imputes new values to replace those selected values. Let $Y_{\text{rep},i}$ be all the imputed (replaced) values in the ith synthetic dataset, and let Y_{nrep} be all unchanged (not replaced) values. The values in Y_{nrep} are the same in all synthetic datasets. Each synthetic dataset, D_i, is then comprised of $(Y_{\text{rep},i}, Y_{\text{nrep}})$. Imputations are made independently for $i = 1, \ldots, m$ times to yield m different synthetic datasets. These synthetic datasets are released to the public.

When using parametric imputation models, the $Y_{\text{rep},i}$ should be generated from the Bayesian posterior predictive distribution of $(Y_{\text{rep},i}|D_{\text{obs}})$, or some approximation to it such as the sequential regression imputation methods [11]. In this article, we generate the $Y_{\text{rep},i}$ from a series of regression tree (CART) models. These models are described in Section 4.1.

Inferences about some scalar estimand, say Q, are obtained by combining results from the D_i. Specifically, suppose that the data analyst estimates Q with some point estimator q and estimates the variance of q with some estimator v. For $i = 1, \ldots, m$, let q_i and v_i be respectively the values of q and v in D_i. It is assumed that the analyst determines q_i and v_i as if D_i was in fact a random sample collected with the original sampling design. The following quantities are needed for inferences for scalar Q:

$$\bar{q}_m = \sum_{i=1}^{m} q_i/m \tag{1}$$

$$b_m = \sum_{i=1}^{m} (q_i - \bar{q}_m)^2/(m-1) \tag{2}$$

$$\bar{v}_m = \sum_{i=1}^{m} v_i/m \ . \tag{3}$$

The analyst then can use $\bar{q}_m$ to estimate Q and

$$T_{\text{p}} = b_m/m + \bar{v}_m \tag{4}$$

to estimate the variance of $\bar{q}_m$. When s is large, inferences for scalar Q can be based on t-distributions with degrees of freedom $\nu_{\mathrm{p}} = (m-1)(1+r_m^{-1})^2$, where $r_m = (m^{-1}b_m/\bar{v}_m)$. Derivations of these methods are presented in [2]. Extensions for multivariate Q are presented in [12].

3 Identification Disclosure Risk Measures for Partial Synthesis

To evaluate disclosure risks, we compute probabilities of identification by following the approach in [13]. Related approaches for non-synthetic data are in [14,15,16,17]. Roughly, in this approach we mimic the behavior of an ill-intentioned user of the released data who possesses the true values of the quasi-identifiers for selected target records (or even the entire population). The intruder has a vector of information, $\mathbf{t}$, on a particular target unit in the population which may or may not correspond to a unit in the m partially synthetic datasets, $\mathbf{D} = \{D_1, \ldots, D_m\}$. Let t_0 be the unique identifier (e.g., full name and address of a survey respondent) of the target, and let d_{j0} be the (not released) unique identifier for record j in $\mathbf{D}$, where $j = 1, \ldots, s$. Let M be any information released about the simulation models.

The intruder's goal is to match unit j in $\mathbf{D}$ to the target when $d_{j0} = t_0$, and not to match when $d_{j0} \neq t_0$ for any $j \in \mathbf{D}$. Let J be a random variable that equals j when $d_{j0} = t_0$ for $j \in \mathbf{D}$ and equals $s+1$ when $d_{j0} = t_0$ for some $j \notin \mathbf{D}$. The intruder thus seeks to calculate the $Pr(J = j|\mathbf{t}, \mathbf{D}, M)$ for $j = 1, \ldots, s+1$. He or she then would decide whether or not any of the identification probabilities for $j = 1, \ldots, s$ are large enough to declare an identification. Let Y_{rep} be all original values of the variables that were synthesized. Because the intruder does not know the actual values in Y_{rep}, he or she should integrate over its possible values when computing the match probabilities. Hence, for each record in $\mathbf{D}$, we compute

$$Pr(J = j|\mathbf{t}, \mathbf{D}, M) = \int Pr(J = j|\mathbf{t}, \mathbf{D}, Y_{\mathrm{rep}}, M) Pr(Y_{\mathrm{rep}}|\mathbf{t}, \mathbf{D}, M) dY_{\mathrm{rep}} \ . \quad (5)$$

This construction suggests a Monte Carlo approach to estimating each $Pr(J = j|\mathbf{t}, \mathbf{D}, M)$. First, sample a value of Y_{rep} from $Pr(Y_{\mathrm{rep}}|\mathbf{t}, \mathbf{D}, M)$. Let Y^{new} represent one set of simulated values. Second, compute $Pr(J = j|\mathbf{t}, \mathbf{D}, Y_{\mathrm{rep}} = Y^{\mathrm{new}}, M)$ using exact or, for continuous synthesized variables, distance-based matching assuming Y^{new} are collected values. This two-step process is iterated R times, where ideally R is large, and (5) is estimated as the average of the resultant R values of $Pr(J = j|\mathbf{t}, \mathbf{D}, Y_{\mathrm{rep}} = Y^{\mathrm{new}}, M)$. When M has no information, the intruder can treat the simulated values in each $Y_{\mathrm{rep},i}$ as plausible draws of Y_{rep}.

To illustrate, suppose that age, race, and sex are the only quasi-identifiers in a survey of households. The agency releases $m > 1$ partially synthetic datasets with all values of race and age synthesized and sex not changed. We suppose that the agency does not release any information about the imputation model but does reveal which variables are synthesized. Suppose that an intruder seeks to identify a white male aged 45, *and he knows that this target is in the sample.* In each D_i, the intruder would search for all records matching the target on age, race, and sex. Let $N_{\mathbf{t},i}$ be the number of matching records in $\mathbf{D}_i$, where $i = 1, \ldots, m$. When no one with all of those characteristics is in $\mathbf{D}_i$, set $N_{\mathbf{t},i}$ equal to the number of males in D_{obs}, i.e., match on all non-simulated quasi-identifiers. For $j = 1, \ldots, s$,

$$Pr(J = j|\mathbf{t}, \mathbf{D}, M) = (1/m) \sum_i (1/N_{\mathbf{t},i})(Y_{ij}^{\text{new}} = \mathbf{t}) \ , \tag{6}$$

where $(Y_{ij}^{\text{new}} = \mathbf{t}) = 1$ when record j is among the $N_{\mathbf{t},i}$ matches in D_i and equals zero otherwise. We note that $Pr(J = s + 1|\mathbf{t}, \mathbf{D}, M) = 0$ because the intruder knows this target is in the sample.

Now suppose that the intruder *does not know that this target is in the sample.* For $j = 1, \ldots, s$, we have to replace $N_{\mathbf{t},i}$ in (6) with $F_{\mathbf{t}}$, the number of records in the population that match the target on age, race, and sex. When the intruder and the agency do not know $F_{\mathbf{t}}$, it can be estimated using the approach in [17], which assumes that the population counts follow an all-two-way-interactions log-linear model. The agency can determine the estimated counts, $\hat{F}_{\mathbf{t}}$, by fitting this log-linear model with D_{obs}. Alternatively, since D_{obs} is in general not available to intruders, the agency can fit a log-linear model with each D_i, resulting in the estimates $\hat{F}_{\mathbf{t},i}$ for $i = 1, \ldots, m$. We note that $Pr(J = s + 1|\mathbf{t}, \mathbf{D}, M) = 1 - \sum_{j=1}^{s} Pr(J = j|\mathbf{t}, \mathbf{D}, M)$.

For some target records, the value of $N_{\mathbf{t},i}$ might exceed $F_{\mathbf{t}}$ (or $\hat{F}_{\mathbf{t}}$ if it is used). It should not exceed $\hat{F}_{\mathbf{t},i}$, since $\hat{F}_{\mathbf{t},i}$ is required to be at least as large as $N_{\mathbf{t},i}$. For such cases, we presume that the intruder sets $Pr(J = s + 1|\mathbf{t}, \mathbf{D}, M) = 0$ and picks one of the matching records at random. To account for this case, we can re-write (6) for $j = 1, \ldots, s$ as

$$Pr(J = j|\mathbf{t}, \mathbf{D}, M) = (1/m) \sum_i \min\left(1/F_{\mathbf{t}}, 1/N_{\mathbf{t},i}\right) (Y_{ij}^{\text{new}} = \mathbf{t}) \ . \tag{7}$$

As suggested in [16], we quantify disclosure risks with summaries of the identification probabilities in (6) and (7). It is reasonable to assume that the intruder selects as a match for $\mathbf{t}$ the record j with the highest value of $Pr(J = j|\mathbf{t}, \mathbf{D}, M)$, if a unique maximum exists. We consider three disclosure risk measures. To calculate these measures, we need some further definitions. Let $\mathbf{T} = \{\mathbf{t}_1, \ldots, \mathbf{t}_{|\mathbf{T}|}\}$ be the set of the intruder's targets. Let c_j be the number of records in the released data with the highest match probability for the target $\mathbf{t}_j$; let $I_j = 1$ if the true match is among the c_j units and $I_j = 0$ otherwise. Let $K_j = 1$ when $c_j I_j = 1$ and $K_j = 0$ otherwise. The *expected match risk* is defined as $\sum_{j \in \mathbf{T}} (1/c_j) I_j$. When $I_j = 1$ and $c_j > 1$, the contribution of unit j to the expected match risk reflects the intruder randomly guessing at the correct match from the c_j candidates. The *true match risk* equals $\sum_{j \in \mathbf{T}} K_j$. Finally, we introduce the *true match rate* equal to $\sum_{j \in \mathbf{T}} K_j / \sum_{j \in \mathbf{T}} (c_j = 1)$, which is the percentage of true matches for the targets that have a unique match in $\mathbf{D}$.

Table 1. Description of variables used in the empirical studies

Variable	Label	Range
Sex	X	male, female
Race	R	white, black, American Indian, Asian
Marital status	M	7 categories, coded 1–7
Highest attained education level	E	16 categories, coded 31–46
Age (years)	G	0 – 90
Child support payments ($)	C	0, 1 – 23,917
Social security payments ($)	S	0, 1 – 50,000
Household alimony payments ($)	A	0, 1 – 54,008
Household property taxes ($)	P	0, 1 – 99,997
Household income ($)	I	-21,011 – 768,742

4 Empirical Evaluation

We simulate partial synthesis for a subset of public release data from the March 2000 U.S. Current Population Survey. The data comprise ten variables measured on $N = 51,016$ heads of households. The variables, displayed in Table 1, were selected and provided by statisticians at the U.S. Bureau of the Census. Similar data are used in [18] to illustrate and evaluate releasing fully synthetic data.

Marginally, there are ample numbers of people in each sex, race, marital status, and education category. Many cross-classifications have few people, especially those involving minorities with $M \notin \{1, 7\}$. There are 521 records with unique combinations of age, race, marital status, and sex. There are 284 combinations of the four variables that have only two records in the dataset. There are 2064 empty cells in the four-way contingency table.

We treat the N records as a population and take a random sample of $n = 10,000$ for D_{obs}. We consider age, race, marital status, and sex to be quasi-identifiers that intruders may know precisely. Cross-classification of these four variables in the sample yields 473 sample uniques, 241 duplicates and 2909 empty cells in the four-way contingency table. Intruders might have access to other variables on the file, such as property taxes. Thus, the computations in this section serve to illustrate our suggested disclosure risk measures rather than to evaluate the actual disclosure risks for this specific dataset (which is already in the public domain).

We generate synthetic datasets for each of two scenarios: replace all values of age, marital status, and race without changing sex; and, replace all values of marital status and race without changing age and sex. The synthetic data are generated using regression trees, as we now describe.

4.1 CART Synthesis Models

CART models are a flexible tool for estimating the conditional distribution of a univariate outcome given multivariate predictors. Essentially, the CART model partitions the predictor space so that subsets of units formed by the partitions have relatively homogeneous outcomes. The partitions are found by recursive binary splits of the predictors. The series of splits can be effectively represented by a tree structure, with leaves corresponding to the subsets of units.

CART models also can be used to generate partially synthetic data [19]. To synthesize all values of age, marital status, and race, we proceed as follows. First, using D_{obs} we fit the tree of age on all other variables except race and marital status. Label this tree $\mathcal{Y}_{(G)}$. We require a minimum of five records in each leaf of the tree and do not prune it; see [19] for discussion of pruning and minimum leaf size. Let L_{Gw} be the wth leaf in $\mathcal{Y}_{(G)}$, and let $Y_{(G)}^{L_{Gw}}$ be the $n_{L_{Gw}}$ values of $Y_{(G)}$ in leaf L_{Gw}. In each L_{Gw} in the tree, we generate a new set of values by drawing from $Y_{(G)}^{L_{Gw}}$ using the Bayesian bootstrap [20]. These sampled values are the replacement imputations for the $n_{L_{Gw}}$ units that belong to L_{Gw}. Repeating the Bayesian bootstrap in each leaf of the age tree results in the ith set of synthetic ages, $Y_{(G)\text{rep},i}$.

To avoid releasing only values of the observed ages in each leaf, we could take an additional step suggested in [19]. In each leaf, we could estimate the density of the bootstrapped values using a Gaussian kernel density estimator with support over the smallest to the largest value of $Y_{(G)}$. Then, for each unit, we would sample randomly from the estimated density in that unit's leaf using an inverse-cdf method. The sampled values rounded to the nearest integer would be the $Y_{(G)rep,i}$. We do not take this extra step here.

Imputations are next made for marital status. Using D_{obs}, we fit the tree, $\mathcal{Y}_{(M)}$, with all variables except race as predictors. To maintain consistency with $Y_{(G)\text{rep},i}$, units' leaves in $\mathcal{Y}_{(M)}$ are located using $Y_{(G)\text{rep},i}$. Occasionally, some units may have combinations of values that do not belong to one of the leaves of $\mathcal{Y}_{(M)}$. For these units, we search up the tree until we find a node that contains the combination, then treat that node as if it were the unit's leaf. Once each unit's leaf is located, values of $Y_{(M)\text{rep},i}$ are generated using the Bayesian bootstrap. Imputing races follows the same process: we fit the tree $\mathcal{Y}_{(R)}$ using all variables as predictors, place each unit in the leaves of $\mathcal{Y}_{(R)}$ based on their synthesized values of age and marital status, and sample new races using the Bayesian bootstrap.

The process is repeated independently $m = 5$ times. These m datasets would be released to the public. All CART models are fit in S-Plus using the "tree" function. It takes about five minutes to generate five synthetic datasets with all three variables. The sequential order of imputation is $G - M - R$; see [19] for a discussion of the ordering of the trees. The synthesis of only marital status and race is similar except that the process begins with marital status. Although we use the CART method only to generate categorical data, it is straightforward to apply the method to generate continuous variables [19].

4.2 Data Utility

Evaluating disclosure risk is, of course, only part of the story. We could create completely worthless data and have very low disclosure risks. Hence, it is important to examine data usefulness when evaluating disclosure risks.

Table 2. Point estimates and standard errors for observed data, synthetic data with age not replaced, and synthetic data with age replaced

	Observed Data	True Age	Synth. Age
Estimand	q_{obs} (SE)	$\bar{q}_5$ ($\sqrt{T_{\text{p}}}$)	$\bar{q}_5$ ($\sqrt{T_{\text{p}}}$)
Avg. education for married black females	39.5 (.21)	39.6 (.21)	39.7 (.20)
Coefficient in regression of $\sqrt{C}$ on:			
Intercept	-94.5 (27)	-94.5 (27)	-95.3 (27)
Female	12.5 (5.4)	12.4 (5.4)	12.2 (5.4)
Non-white	-1.72 (4.7)	-0.34 (4.9)	-0.53 (4.8)
Education	3.44 (0.6)	3.44 (.60)	3.46 (0.6)
Number of youths in house	1.33 (1.6)	1.34 (1.6)	1.37 (1.6)
Coefficient in regression of $\sqrt{S}$ on:			
Intercept	81.0 (4.5)	78.1 (4.6)	79.4 (4.9)
Female	-11.1 (1.1)	-11.1 (1.1)	-10.6 (1.1)
Black	-7.0 (1.6)	-6.3 (1.9)	-5.3 (1.8)
American Indian	-8.2 (4.7)	-8.9 (7.1)	-10.8 (5.5)
Asian	0.1 (3.3)	-3.1 (3.8)	2.3 (3.7)
Widowed	5.0 (1.2)	4.7 (1.2)	4.3 (1.2)
Divorced	-3.0 (1.7)	-0.3 (1.8)	0.3 (1.8)
Single	-1.4 (2.1)	2.0 (2.1)	3.5 (2.2)
High school	3.6 (1.1)	3.8 (1.1)	3.8 (1.1)
Some college	5.2 (1.3)	5.1 (1.3)	5.7 (1.3)
College degree	8.3 (1.7)	8.3 (1.7)	8.1 (1.7)
Advanced degree	10.1 (2.1)	9.8 (2.2)	9.8 (2.2)
Age	0.22 (.06)	0.25 (.06)	0.23 (.07)
Coefficient in regression of $\log(I)$ on			
Intercept	4.80 (.10)	4.78 (.10)	4.82 (.15)
Black	-0.14 (.03)	-0.16 (.03)	-0.12 (.03)
American Indian	-0.20 (.07)	-0.21 (.09)	-0.12 (.08)
Asian	-0.01 (.05)	0.04 (.06)	0.01 (.05)
Female	0.02 (.02)	0.01 (.03)	-0.002 (.03)
Married in armed forces	-0.04 (.10)	-0.30 (.15)	-0.19 (.11)
Widowed	-0.07 (.06)	-0.17 (.07)	-0.30 (.08)
Divorced	-0.11 (.04)	-0.14 (.05)	-0.13 (.04)
Separated	-0.28 (.09)	-0.13 (.11)	-0.24 (.10)
Single	-0.15 (.04)	-0.11 (.04)	-0.12 (.04)
Education	0.113 (.003)	0.113 (.003)	0.114 (.003)
Household size > 1	0.54 (.03)	0.54 (.03)	0.52 (.03)
Females married in armed forces	-0.49 (.14)	-0.22 (.16)	-0.39 (.14)
Widowed females	-0.27 (.07)	-0.15 (.07)	-.07 (.08)
Divorced females	-0.34 (.05)	-0.31 (.06)	-0.33 (.06)
Separated females	-0.45 (.11)	-0.48 (.13)	-0.41 (.12)
Single females	-0.35 (.05)	-0.37 (.05)	-0.33 (.05)
Age	0.043 (.003)	0.043 (.003)	.041 (.003)
Age^2 ×1000	-0.42 (.03)	-0.42 (.03)	-0.41 (.03)
Property tax ×10000	0.27 (.03)	0.29 (.03)	0.29 (.03)

Child support regression fit using records with $C > 0$. Social security regression fit using records with $S > 0$ and $G > 54$. Income regression fit using records with $I > 0$.

Table 2 provides some evidence of the usefulness of the five synthetic datasets. It displays the point estimates and standard errors for several quantities based on the observed and partially synthetic data. Synthetic estimates are computed from the $m = 5$ datasets using the methods described in Section 2. The synthetic data point estimates are generally within two standard errors of the observed data point estimates. The biggest differences are for quantities associated with small sub-groups, such as married in the armed forces. We believe that the results in Table 2 are evidence of good quality, especially since the regressions involved subsets of data, transformations of variables, and interaction effects. We note that these results were obtained without any tuning other than to decide on the minimum number of records for each leaf and the order of synthesis. We also note that the results for synthesizing or not synthesizing age are similar.

4.3 Disclosure Risk

We consider four scenarios with different assumptions about the information available to the intruder. Across all scenarios, we assume the intruder knows the sex, age, race and marital status of some target records, for example from an external database.

- Scenario I: the intruder knows the identifiers for 10,000 randomly specified units in the population but does not know who is in the survey.
- Scenario II: the intruder knows the identifiers for 10,000 randomly specified units in the population and knows who is in the survey.
- Scenario III: the intruder knows the identifiers for all $N = 51,016$ units in the population but does not know who is in the survey.
- Scenario IV: the intruder knows the identifiers for all $N = 51,016$ units in the population and knows who is in the survey.

For Scenarios I and II, 1,968 of the intruder's target records are included in D_{obs}. For Scenario I, we estimate each $\hat{F}_{\mathbf{t},i}$ by fitting the all-two-way-interactions log-linear model on each D_i. An intruder might do this if he is unsure whether or not his $10,000$ records are representative of the population. It is prudent for the agency to assess the disclosure risk using estimated counts based on D_{obs} as well. For Scenario III, the intruder presumably would use the known values of $F_{\mathbf{t}}$. For interest, we report the results for the first and third scenarios using both estimated and true population counts.

For Scenarios I and III, we consider three intruder strategies. The first is that the intruder matches to the released data no matter what the value of $Pr(J = s + 1|\mathbf{t}, \mathbf{D}, M)$. That is, the intruder ignores the chance that a record is not in the sample. The second is that the intruder matches to the released data only when $Pr(J = s + 1|\mathbf{t}, \mathbf{D}, M) < \gamma$, where $0 < \gamma < 1$. The third is that the intruder does not match whenever $Pr(J = s + 1|\mathbf{t}, \mathbf{D}, M)$ is the maximum probability for the target.

We compare the risks when only race and marital status are synthesized to the risks when age, race, and marital status are synthesized.

Table 3. Disclosure risks when only marital status and race are synthesized and intruder matches regardless of the value of $Pr(J = s + 1|\mathbf{t}, \mathbf{D}, M)$

	Scen. I		Scen. II	Scen. III		Scen. IV
	$F_{\mathbf{t}}$	$\hat{F}_{\mathbf{t},i}$		$F_{\mathbf{t}}$	$\hat{F}_{\mathbf{t},i}$	
Expected match risk	72.3	71.6	74.7	367.8	361.1	365.1
True match risk	26	40	37	131	201	172
Number of single matches	1,942	3,445	593	9,769	17,555	2,905
True match rate (%)	1.34	1.16	6.24	1.34	1.14	5.92

Synthesis of Race and Marital Status Only. Table 3 displays the risk measures when age is left unchanged and the intruder matches regardless of the value of $Pr(J = s + 1|\mathbf{t}, \mathbf{D}, M)$. In all scenarios, the great majority of declared matches are incorrect, as evident by the low true match rates. True match rates are highest when the intruder knows who is in the sample, as might be expected. Given $\mathbf{T}$, the expected match risk measures are very similar for an intruder with response knowledge and an intruder not knowing who participated in the survey. The true match risk measures are higher when using the $\hat{F}_{\mathbf{t},i}$ instead of $F_{\mathbf{t}}$. This is because the number of matches with $c_j = 1$ is higher when matching with $\hat{F}_{\mathbf{t},i}$ instead of $F_{\mathbf{t}}$, as evident in the third row of the table.

Naturally, the numbers of expected and true matches increase when the intruder has information for the whole population rather than only for a sample. Quite simply, there are more targets to match. The expected and true risk measures when only around 2000 records are in $\mathbf{T} \cap D_{\text{obs}}$ are roughly 1/5 the magnitudes when all 10000 records in D are in $\mathbf{T} \cap D_{\text{obs}}$.

The results in Table 3 presume that the intruder always considers the record j with maximum $Pr(J = j|\mathbf{t}, \mathbf{D}, M)$, where $j = 1, \ldots, s$ a match no matter how small this maximum is. With this strategy, the number of true matches is swamped by the number of false matches. For targets with $J = s + 1$ as the maximum match probability, the intruder might not match if he deems $Pr(J = s + 1|\mathbf{t}, \mathbf{D}, M)$ to be too high, say exceeding a threshold γ. Large values of γ result in a higher number of true and false matches. Small values of γ reduce the chance of false matches but miss out on some true matches. Table 4 presents the risk measures for Scenario I and III using $\gamma = 0.5$. As expected, there is a reduction in both the number of true matches and the total number of single matches. In fact, in Scenario I the intruder detects very few correct matches. However, in both scenarios the true match rate increases from around 1% to at least 8%.

The intruder also might choose not to match for targets with $Pr(J = s + 1|\mathbf{t}, \mathbf{D}, M) \geq Pr(J = j|\mathbf{t}, \mathbf{D}, M)$ for $j = 1, \ldots, s$. Applying this strategy, the intruder obtains 2 true matches (with a match rate of 50%) in Scenario I and 6 true matches (with a match rate of 20%) in Scenario III.

Synthesis of Age, Race, and Marital Status. The agency may decide that the disclosure risks are too high when synthesizing only race and marital status.

Table 4. Disclosure risks for Scenario I and III when only marital status and race are synthesized and the intruder matches if $Pr(J = s+1|\mathbf{t}, \mathbf{D}, M) \leq 0.5$

	Scen. I		Scen. III	
	$F_{\mathbf{t}}$	$\hat{F}_{\mathbf{t},i}$	$F_{\mathbf{t}}$	$\hat{F}_{\mathbf{t},i}$
Expected match risk	3	1	9.5	6
True match risk	3	1	9	6
Number of single matches	17	11	102	64
True match rate (%)	17.65	9.09	8.82	9.37

Table 5. Disclosure risks when age, marital status, and race are synthesized and intruder matches regardless of the value of $Pr(J = s+1|\mathbf{t}, \mathbf{D}, M)$

	Scen. I		Scen. II	Scen. III		Scen. IV
	$F_{\mathbf{t}}$	$\hat{F}_{\mathbf{t},i}$		$F_{\mathbf{t}}$	$\hat{F}_{\mathbf{t},i}$	
Expected match risk	3.5	3.0	4.2	14.0	14.7	16.0
True match risk	2	3	3	4	12	12
Number of single matches	2,651	6,879	1,252	13,641	34,972	6,359
True match rate (%)	0.075	0.044	0.240	0.029	0.034	0.189

Table 6. Disclosure risks for Scenario I and III when age, marital status, and race are synthesized and the intruder matches if $Pr(J = s+1|\mathbf{t}, \mathbf{D}, M) \leq 0.5$

	Scen. I		Scen. III	
	$F_{\mathbf{t}}$	$\hat{F}_{\mathbf{t},i}$	$F_{\mathbf{t}}$	$\hat{F}_{\mathbf{t},i}$
Expected match risk	0	0	0	0
True match risk	0	0	0	0
Number of single matches	6	6	48	41
True match rate (%)	0	0	0	0

Table 5 displays the results if age is also synthesized, assuming that the intruder matches no matter what. The risks decrease significantly. The true match rate drops well below 1% for all scenarios. Table 6 displays the risks when the intruder matches only if $Pr(J = s+1|\mathbf{t}, \mathbf{D}, M) < 0.5$. The intruder cannot detect any correct matches.

Synthesizing age appears to reduce the disclosure risks substantially for this dataset. Given the similarity in the data utility of the two approaches, we suspect that many agencies would opt to synthesize age.

5 Concluding Remarks

The simulation results suggest several conclusions about disclosure risks in partially synthetic data. These include:

1. Knowing which targets are in the sample increases the true match rate compared to not knowing which targets are in the sample, so that disclosure risks increase.
2. Intruders who match to the synthetic data regardless of the value of $Pr(J = s+1|\mathbf{t}, \mathbf{D}, M)$ can find more true matches at the expense of a higher false match rate than intruders who would not match when $Pr(J = s+1|\mathbf{t}, \mathbf{D}, M)$ is large.
3. There are differences in the risk measures when using estimated population counts versus true population counts. However, they tend to be small and arguably not worth worrying about.
4. Synthesizing variables that are primary contributors to the disclosure risks, in particular age, can reduce disclosure risks substantially.

In general, it is difficult for the agency to know what information is owned by intruders. We recommend that the agency evaluate disclosure risks under conservative but realistic assumptions of intruder knowledge. For example, to begin, the agency can assume that intruders know exactly who is in the sample and have correct values of all quasi-identifiers. The agency then can back off these assumptions, for example assuming that intruders do not know who is in the sample or that intruders do now know some quasi-identifiers. By computing risk and utility under a variety of assumptions, the agency can decide if the disclosure risks are adequately low for the proposed microdata release.

Acknowledgments. This research was supported by grants from the U.S. National Science Foundation (NSF-ITR-0427889 and NSF-MMS-0751671) and the IAB in Germany.

References

1. Little, R.J.A.: Statistical analysis of masked data. J. Off. Stat. 9, 407–426 (1993)
2. Reiter, J.P.: Inference for partially synthetic, public use microdata sets. Surv. Methodol. 29, 181–189 (2003)
3. Kennickell, A.B.: Multiple imputation and disclosure protection: the case of the 1995 Survey of Consumer Finances. In: Record Linkage Techniques, pp. 248–267. National Academy Press, Washington (1997)
4. Abowd, J.M., Stinson, M., Benedetto, G.: Final report to the Social Security Administration on the SIPP/SSA/IRS public use file project. Technical report, U.S. Census Bureau Longitudinal Employer-Household Dynamics Program (2006)
5. Abowd, J.M., Woodcock, S.D.: Disclosure limitation in longitudinal linked data. In: Doyle, P., Lane, J., Zayatz, L., Theeuwes, J. (eds.) Confidentiality, Disclosure, and Data Access: Theory and Practical Applications for Statistical Agencies, pp. 215–277. North-Holland, Amsterdam (2001)
6. Abowd, J.M., Woodcock, S.D.: Multiply-imputing confidential characteristics and file links in longitudinal linked data. In: Domingo-Ferrer, J., Torra, V. (eds.) PSD 2004. LNCS, vol. 3050, pp. 290–297. Springer, Heidelberg (2004)
7. Reiter, J.P.: Simultaneous use of multiple imputation for missing data and disclosure limitation. Surv. Methodol. 30, 235–242 (2004)

8. Little, R.J.A., Liu, F., Raghunathan, T.E.: Statistical disclosure techniques based on multiple imputation. In: Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives, pp. 141–152. John Wiley & Sons, New York (2004)
9. Mitra, R., Reiter, J.P.: Adjusting survey weights when altering identifying design variables via synthetic data. In: Domingo-Ferrer, J., Franconi, L. (eds.) PSD 2006. LNCS, vol. 4302, pp. 177–188. Springer, Heidelberg (2006)
10. Drechsler, J., Bender, S., Rässler, S.: Comparing fully and partially synthetic datasets for statistical disclosure control in the German IAB Establishment Panel. Joint Eurostat UNECE Worksession on Statistical Data Confidentiality, Manchester, WP. 11 (2007)
11. Raghunathan, T.E., Lepkowski, J.M., van Hoewyk, J., Solenberger, P.: A multivariate technique for multiply imputing missing values using a series of regression models. Surv. Methodol. 27, 85–96 (2001)
12. Reiter, J.P.: Significance tests for multi-component estimands from multiply-imputed, synthetic microdata. J. Stat. Plan. Inf. 131, 365–377 (2005)
13. Reiter, J.P., Mitra, R.: Estimating risks of identification disclosure in partially synthetic data. J. Priv. Conf. (to appear)
14. Duncan, G.T., Lambert, D.: The Risk of disclosure for microdata. Journal of Business and Economic Statistics 7, 207–217 (1989)
15. Fienberg, S.E., Makov, U.E., Sanil, A.P.: A Bayesian approach to data disclosure: Optimal intruder behavior for continuous data. J. Off. Stat. 13, 75–89 (1997)
16. Reiter, J.P.: Estimating identification risks in microdata. J. Amer. Stat. Assoc. 100, 1103–1113 (2005)
17. Elamir, E.A.H., Skinner, C.J.: Record level measures of disclosure risk for survey microdata. J. Off. Stat. 22, 525–529 (2006)
18. Reiter, J.P.: Releasing multiply-imputed, synthetic public use microdata: An illustration and empirical study. J. Roy. Stat. Soc. A 168, 531–544 (2005)
19. Reiter, J.P.: Using CART to generate partially synthetic, public use microdata. J. Off. Stat. 21, 441–462 (2005)
20. Rubin, D.B.: The Bayesian bootstrap. Ann. Stat. 9, 130–134 (1981)

How Protective Are Synthetic Data?

John M. Abowd[1] and Lars Vilhuber[1]

School of Industrial and Labor Relations, Cornell University

Abstract. This short paper provides a synthesis of the statistical disclosure limitation and computer science data privacy approaches to measuring the confidentiality protections provided by fully synthetic data. Since all elements of the data records in the release file derived from fully synthetic data are sampled from an appropriate probability distribution, they do not represent "real data," but there is still a disclosure risk. In SDL this risk is summarized by the inferential disclosure probability. In privacy-protected database queries, this risk is measured by the differential privacy ratio. The two are closely related. This result (not new) is demonstrated and examples are provided from recent work.

1 Introduction

When Rubin (1993) introduced the idea of fully synthetic data , there was considerable appeal to releasing data that represented "no actual individual's" responses, and skepticism regarding its feasibility. Subsequent research has adequately demonstrated the feasibility. However, the basic question "How much protection does the synthetic data methodology provide?" remained largely unanswered. The reason is basic: statistical disclosure limitation (SDL) did not provide an adequate framework to answer the question. In the intervening 15 years, a well-developed methodology emerged in the computer science (CS) literature on privacy in databases that allows a synthesis of the techniques used in disclosure limitation and privacy-preserving data mining. The key to this synthesis is the recognition that the privacy measures proposed by the computer scientists and the statistical disclosure limitation methods share a common fundamental– the conditional distribution of the release data, given the underlying confidential data. This short paper provides a roadmap and some examples for moving between the SDL and CS concepts that relate to measuring the protection afforded by synthetic data and the resulting analytical validity of the release data.

2 Definitions

Let X represent a confidential database organized as n rows and k columns of a database table. For clarity in this exposition, assume that only discrete variables may be released and that all variables have been coded as binary outcomes (*e.g.*, yes-no answers). Although one of the great conceptual advantages of fully synthetic data is the possibility of combining continuous and discrete variables,

J. Domingo-Ferrer and Y. Saygın (Eds.): PSD 2008, LNCS 5262, pp. 239–246, 2008.

there is no loss of generality in the assumption that the release data consist of contingency tables because all interactions up to k-way are allowed and there are no restrictions on the underlying probabilities. As we will see below, there are practical restrictions on the direct application of these techniques to databases where k is large. We are not going to discuss sampling as a disclosure limitation technique; consequently, we will assume that n is the population and $n_i = 1$ is a population unique. That is, there is one, and only one row of X in which i^{th} column has a 1.

Let $\boldsymbol{\pi}$ be the $(k \times 1)$ vector of probabilities associated with the complete table, where all elements of $\boldsymbol{\pi}$ are strictly positive. Assume that the contingency table is summarized by a vector of counts $\boldsymbol{n}$ that is also $(k \times 1)$ with $n = \sum_{i=1}^{k} n_i$. Without loss of generality, assume that the confidential data are distributed Multinomial, $\boldsymbol{n} \sim \mathrm{M}(n, \boldsymbol{\pi})$. Summarize all prior information about the parameters by assuming that they are drawn from a Dirichlet distribution, $\boldsymbol{\pi} \sim \mathrm{D}(\boldsymbol{\alpha})$, where $\boldsymbol{\alpha}$ is the $(k \times 1)$ vector of prior sample sizes with $\alpha_0 = \sum_{i=1}^{k} \alpha_i$. Then, the posterior predictive distribution of the confidential data can be constructed by noting that $\boldsymbol{\pi} \sim \mathrm{D}(\boldsymbol{\alpha} + \boldsymbol{n})$ *a posteriori.*

Let $\tilde{X}$ denote a single synthetic data set based on X. Suppose that $\tilde{X}$ is $(m \times k)$. The synthetic data can be constructed by first sampling $\tilde{\boldsymbol{\pi}} \sim \mathrm{D}(\boldsymbol{\alpha} + \boldsymbol{n})$, then constructing the rows of $\tilde{X}$ from counts sampled from $\mathrm{M}(m, \tilde{\boldsymbol{\pi}})$. Because of the way $\tilde{X}$ is constructed, we can represent the conditional distribution of $\tilde{X}$ given X using

$$\Pr[\boldsymbol{m}|\boldsymbol{n}, MD] = \mathrm{E}_{\boldsymbol{\pi}|\boldsymbol{n}}[\mathrm{M}(m, \boldsymbol{\pi})\,|MD] \tag{1}$$

where we have noted explicitly that the conditional distribution depends upon the Multinomial-Dirichlet (MD).

The argument leading up to the construction of $\Pr[\boldsymbol{m}|\boldsymbol{n}, MD]$ above is a complete Bayesian analysis, and equation (1) defines the posterior predictive distribution of $\tilde{X}$ given X. But the Bayesian analysis is not essential to the synthetic data construction. Any transition function $\Pr[\boldsymbol{m}|\boldsymbol{n}]$ that defines a proper conditional distribution for the synthetic counts given the confidential counts can be used to synthesize data. Dwork *et al.* (2006) define a synthesizer for the same confidential database problem by sampling k i.i.d. random variables from the Laplace (double exponential) distribution $\mathrm{Lap}(0, 2/\epsilon)$, where the reason for defining the scale parameter in the form shown will be made clear below. Let $\boldsymbol{y}$ be the $(k \times 1)$ vector of Laplacian random variables. Define the synthetic counts as $\boldsymbol{m} = \boldsymbol{n} + \boldsymbol{y}$. Using the properties of the Laplace distribution, they construct an alternative conditional distribution

$$\Pr[\boldsymbol{m}|\boldsymbol{n}, Lap] = \Pr[\boldsymbol{n} + \boldsymbol{y}|\boldsymbol{n}, \epsilon]\,. \tag{2}$$

The above discussion has been in terms of conditional distributions. A generic random sanitizer is defined as any function $\tilde{X} \leftarrow \mathrm{San}(X, Y)$ that maps the confidential data X and random noise Y of specified dimensionality into a sanitized copy of the database, denoted $\tilde{X}$ here to emphasize its relation to synthetic data. Because of the way we constructed $\boldsymbol{n}$ from X, there is a completely equivalent sanitizer $\boldsymbol{m} \leftarrow \mathrm{San}\left(\boldsymbol{n}, \boldsymbol{y}\right)$. Hence, any sanitizer can be used to construct a conditional distribution $\Pr\left[\boldsymbol{m}|\boldsymbol{n}, San\right]$. Thus, a discussion of sanitizers is equivalent to a discussion of the conditional distribution constructed from those sanitizers, and in the remainder of this paper, we will focus on conditional distributions, without loss of generality.

3 Statistical Disclosure Limitation and Differential Privacy

Consider a generic conditional distribution $\Pr\left[\boldsymbol{m}|\boldsymbol{n}\right]$, and represent the conditional probabilities in a matrix $\Upsilon\ (k \times k)$. SDL methods focus on the rows of Υ. For example, if $\Upsilon = I$, then the release data are identical to the confidential data. If

$$\max\left(\mathrm{diag}\left(\Upsilon\right)\right) < 1 - \delta;$$

then, the release data differ from confidential data in every dimension by at least δ. That is, for all $i = 1, \dots k$

$$\Pr\left[m_i \neq n_i|\boldsymbol{n}, San\right] > \delta$$

and the SDL is defined to have infused at least $\delta-$percent uncertainty into every tabulation. Acceptable levels of δ are usually an inverse function of n_i. Furthermore, the actual values of δ are usually kept secret.

By contrast, the computer science data privacy literature concerns itself with the columns of Υ. To understand this formally, consider two copies of X, say $X^{(1)}$ and $X^{(2)}$ that differ in a single row such that $\left|\boldsymbol{n}^{(1)} - \boldsymbol{n}^{(2)}\right| = 2$. While this condition looks obscure, it amounts to assuming that the two copies of the database differ on a single attribute of a single row; hence, some n_i changes from 0 to 1 while exactly one other n_j changes from 1 to 0. Dwork *et al.* (2006) define $\epsilon-$differential privacy as the requirement that

$$\max\left|\ln\left(\frac{\Pr\left[\boldsymbol{m}|\boldsymbol{n}^{(1)}\right]}{\Pr\left[\boldsymbol{m}|\boldsymbol{n}^{(2)}\right]}\right)\right| \leq \varepsilon \tag{3}$$

where the max is taken over $\forall \boldsymbol{n}^{(1)}, \boldsymbol{n}^{(2)}$ where $\left|\boldsymbol{n}^{(1)} - \boldsymbol{n}^{(2)}\right| = 2$ and all columns of Υ respecting the convention that the larger element is placed in the numerator.[1] Thus, the computation of the ratios of elements of each column of Υ considers only those combinations for the numerator and denominator that can be reached by change of a single row of X. As an enhancement, Machanavajjhala *et al.* (2008) define $\left(\epsilon, \delta\right)-$probabilistic differential privacy as the requirement that equation (3) hold with probability $1 - \delta$ for $\forall \boldsymbol{n}^{(1)}, \boldsymbol{n}^{(2)}$ where $\left|\boldsymbol{n}^{(1)} - \boldsymbol{n}^{(2)}\right| = 2$, where the

[1] Dwork *et al.* (2006) actually call this $\epsilon-$indistinguishability. Dwork (2006) standardizes the terminology to $\epsilon-$differential privacy.

probabilities are calculated with respect to the joint distribution of $(\boldsymbol{m}, \boldsymbol{n})$, given $\boldsymbol{\alpha}$. They interpret probabilistic differential privacy as $\epsilon-$differential privacy that fails with probability δ, a rare event.

Conventional SDL methods and differential privacy definitions are related by the concept of an inferential disclosure. An inferential disclosure occurs when the attacker can infer the value of a variable for a row in the confidential data by comparing the release data to the information available without the release data (the attacker's information set, or prior). The attacker's prior knowledge is summarized by the ratio

$$\frac{\Pr\left[\boldsymbol{n} = \boldsymbol{n}^{(1)}\right]}{\Pr\left[\boldsymbol{n} = \boldsymbol{n}^{(2)}\right]}$$

which measures the extent to which the attacker can ascertain the difference between $\boldsymbol{n}^{(1)}$ and $\boldsymbol{n}^{(2)}$ without using the release data. The attacker's gain in information from having access to the synthetic release data $\boldsymbol{m} = \tilde{\boldsymbol{m}}$ is given by the posterior odds ratio

$$\frac{\frac{\Pr\left[\boldsymbol{n}=\boldsymbol{n}^{(1)}|\tilde{\boldsymbol{m}}\right]}{\Pr\left[\boldsymbol{n}=\boldsymbol{n}^{(2)}|\tilde{\boldsymbol{m}}\right]}}{\frac{\Pr\left[\boldsymbol{n}=\boldsymbol{n}^{(1)}\right]}{\Pr\left[\boldsymbol{n}=\boldsymbol{n}^{(2)}\right]}}. \tag{4}$$

If the posterior odds ratio is large, then the release data contain a great deal of information about the row associated with the change from $\boldsymbol{n}^{(1)}$ and $\boldsymbol{n}^{(2)}$. At the limit, if this ratio is infinite, an inferential disclosure is certain. But it turns out that

$$\frac{\frac{\Pr\left[\boldsymbol{n}=\boldsymbol{n}^{(1)}|\tilde{\boldsymbol{m}}\right]}{\Pr\left[\boldsymbol{n}=\boldsymbol{n}^{(2)}|\tilde{\boldsymbol{m}}\right]}}{\frac{\Pr\left[\boldsymbol{n}=\boldsymbol{n}^{(1)}\right]}{\Pr\left[\boldsymbol{n}=\boldsymbol{n}^{(2)}\right]}} = \frac{\Pr\left[\boldsymbol{m} = \tilde{\boldsymbol{m}}|\boldsymbol{n}^{(1)}\right]}{\Pr\left[\boldsymbol{m} = \tilde{\boldsymbol{m}}|\boldsymbol{n}^{(2)}\right]}$$

Hence, $\epsilon-$differential privacy limits the maximum gain in information (posterior odds) for an attacker who knows all properties of the disclosure limitation procedure $(\Pr[\boldsymbol{m}|\boldsymbol{n}])$, and all rows of X save one, to

$$\max\left[\frac{\Pr\left[\boldsymbol{m} = \tilde{\boldsymbol{m}}|\boldsymbol{n}^{(1)}\right]}{\Pr\left[\boldsymbol{m} = \tilde{\boldsymbol{m}}|\boldsymbol{n}^{(2)}\right]}\right]$$

where the max is taken over $\forall \boldsymbol{n}^{(1)}, \boldsymbol{n}^{(2)}$ where $\left|\boldsymbol{n}^{(1)} - \boldsymbol{n}^{(2)}\right| = 2$ and all columns of Υ. Furthermore, $(\epsilon, \delta) -$ probabilistic differential privacy limits the maximum gain in information for an attacker with this information with probability $1 - \delta$.

We can now answer the question posed in the title. Fully synthetic data, the type we have discussed in this paper, are protective of the confidential data to the extent that they limit inferences of the type defined by equation (4). Hence, synthetic data that display $\epsilon-$differential privacy are guaranteed to be protective against an attacker with full information about the data protection process (knowledge of $\boldsymbol{\alpha}$ and n for $\Pr[\boldsymbol{m}|\boldsymbol{n}, MD]$; knowledge of ϵ but not n

for $\Pr[\boldsymbol{m}|\boldsymbol{n}, Lap]$; knowledge of $\Pr[\boldsymbol{m}|\boldsymbol{n}, San]$, in general) and knowledge of all but one row of X. Similarly, synthetic data that display $(\epsilon, \delta)-$ probabilistic differential privacy are protective against the same attacker with probability $1-\delta$.

Thus, synthetic data that have one of these differential privacy properties protect against an attacker with an enormous information set, certainly containing more information than conventional SDL procedures assume. But, what of synthetic data procedures that do not satisfy differential privacy? A sanitizer that doesn't satisfy either $\epsilon-$differential privacy or $(\epsilon, \delta)-$ probabilistic differential privacy displays infinite differential privacy ($\epsilon \to \infty$) for some kinds of attacks. Virtually every SDL procedure in regular use–suppression, coarsening, swapping, shuffling, sampling, and most noise-infusion techniques–fails to satisfy differential privacy. For this reason, the users of these methods normally safeguard the parameters and conditioning information required to calculate $\Pr[\boldsymbol{m}|\boldsymbol{n}, San]$. However, applying a differential privacy audit to synthesizers and sanitizers in regular use can be very instructive about their strengths and limitations, as we hope the examples below will demonstrate.

4 Applications

4.1 The Multinomial-Dirichlet Synthesizer

Figure 1 displays $\Pr[\boldsymbol{m}|\boldsymbol{n}, MD]$ a Multinomial-Dirichlet synthesizer that has $(2, 0.0006)-$probabilistic differential privacy. The synthesizer displays the entire sample space for $n = 5, k = 2, \alpha_0 = 1.0, \alpha_1 = \alpha_2 = 0.5$. There is no suppression in the output; hence, every combination of actual data (rows) can produce any possible outcome (columns). This synthesizer displays finite differential privacy, as can be seen in Figure 2. It is the eight cells that have values in excess of 2 that cause the failure of strict $\epsilon-$differential privacy, and those cells have a combined probability of 0.0006.

The properties displayed in Figure 1 are generic features of Multinomial-Dirichlet synthesizers that satisfy finite differential privacy. Notice that the cells that have the largest log posterior odds ratios are those in which the synthesizer delivers "unusual" outcomes–outcomes that are far from the sample data. The

	m_1		0	1	2	3	4	5
n_1	n_2	m_2	5	4	3	2	1	0
0	5		0.647228	0.294194	0.053490	0.004863	0.000221	0.000004
1	4		0.237305	0.395508	0.263672	0.087891	0.014648	0.000977
2	3		0.067544	0.241227	0.344610	0.246150	0.087911	0.012559
3	2		0.012559	0.087911	0.246150	0.344610	0.241227	0.067544
4	1		0.000977	0.014648	0.087891	0.263672	0.395508	0.237305
5	0		0.000004	0.000221	0.004863	0.053490	0.294194	0.647228

Fig. 1. Multinomial-Dirichlet synthesizer with (2,0.0006)-prob. differential privacy

			m_1	0	1	2	3	4	5
$n_1^{(1)}$	$n_2^{(1)}$	$n_1^{(2)}$	$n_2^{(2)}$ m_2	5	4	3	2	1	0
0	5	1	4	1.003353	0.295930	1.595212	2.894495	4.193778	5.493061
1	4	2	3	1.256572	0.494432	0.267708	1.029848	1.791988	2.554128
2	3	3	2	1.682361	1.009417	0.336472	0.336472	1.009417	1.682361
3	2	4	1	2.554128	1.791988	1.029848	0.267708	0.494432	1.256572
4	1	5	0	5.493061	4.193778	2.894495	1.595212	0.295930	1.003353

Fig. 2. Differential privacy values (log posterior odds ratios) for MD synthesizer

natural tendency is to set the synthesizer so that it suppresses these outcomes, but that technique creates zeros in the rows of the transition matrix and, hence, infinite differential privacy. For these cases, probabilistic differential privacy allows the log posterior odds ratios to be large for exactly the low-probability outcomes of the synthesizer.

4.2 The Laplace Sanitizer

Figure 3 displays $\Pr[\boldsymbol{m}|\boldsymbol{n}, Lap]$ for the same (5×2) data matrix with the parameters of the Laplace distribution chosen to guarantee $2-$differential privacy, as in the example above. In order to make the comparison with the MD synthesizer interesting, We have assumed that the total size of the database, $n = 5$ is known. Hence, the appropriate distribution for the noise is $\text{Lap}\,(0, 2/\epsilon)$ with $\epsilon = 2$ (see Dwork *et al.*, page 8), but there is only one query being protected, not two, since the total number of rows in the database is known. Figure 4 confirms that the transition matrix guarantees $2-$differential privacy.

The Laplace sanitizer displayed in Figure 3 is also typical. It displays larger probabilities for the rare events than the MD synthesizer because it never allows the log odds ratio to exceed 2. But, it is also more peaked around the high-probability transitions, which is a feature of the double exponential noise used in the sanitizer.

	m_1	0	1	2	3	4	5
n_1	n_2 m_2	5	4	3	2	1	0
0	5	0.816060	0.159046	0.021525	0.002913	0.000394	0.000062
1	4	0.183940	0.632121	0.159046	0.021525	0.002913	0.000456
2	3	0.024894	0.159046	0.632121	0.159046	0.021525	0.003369
3	2	0.003369	0.021525	0.159046	0.632121	0.159046	0.024894
4	1	0.000456	0.002913	0.021525	0.159046	0.632121	0.183940
5	0	0.000062	0.000394	0.002913	0.021525	0.159046	0.816060

Fig. 3. Laplace synthesizer with 2-differential privacy

			m_1	0	1	2	3	4	5
$n_1^{(1)}$	$n_2^{(1)}$	$n_1^{(2)}$	$n_2^{(2)}$ m_2	5	4	3	2	1	0
0	5	1	4	1.489880	1.379885	2.000000	2.000000	2.000000	2.000000
1	4	2	3	2.000000	1.379885	1.379885	2.000000	2.000000	2.000000
2	3	3	2	2.000000	2.000000	1.379885	1.379885	2.000000	2.000000
3	2	4	1	2.000000	2.000000	2.000000	1.379885	1.379885	2.000000
4	1	5	0	2.000000	2.000000	2.000000	2.000000	1.379885	1.489880

Fig. 4. Differential privacy values (log posterior odds ratios) for Laplace sanitizer

5 Discussion

This short article is just meant to illustrate what is required to answer the question "How protective are synthetic data?" and to provide some generic examples for simple problems. The two articles upon which we have primarily relied contain many more details of both procedures. In particular Machanavajjhala *et al.* (2008) show that the real challenge for the MD synthesizer is to handle problems where the number of columns in the database is huge. Their example, an origin-destination commuting pattern database, has 8.2 million rows. Both the MD synthesizer and the Laplace sanitizer deliver poor analytical validity in this example unless the domain is coarsened. The MD synthesizer gives poor results without coarsening because the minimum prior sample size that must be spread across the 8.2 million possible origins is usually much larger than the number of sample individuals. The Laplace synthesizer also adds noise to each origin and, while the properties of the Laplace noise do not depend upon the number of potential origins (8.2 million), if the release data are provided for each origin, the total amount of noise in the release data is comparable to the M-D synthesizer.

Coarsening the domain can be difficult since all feasible outcomes must have positive transition probabilities for every row of the input database in order to preserve either type of differential privacy. Machanavajjhala *et al.* (2008) address this problem by combining distance-based coarsening with a probabilistic pruning algorithm. When used in combination, the analytical properties of the data can be preserved with a $(4, 0.0001)-$probabilistic differential privacy (Machanavajjhala *et al.*, 2008, page 9).

Dwork *et al.* (2006) consider an equally difficult problem–all possible tables from a census of population. Barak *et al.* (2008) show how to guarantee $\epsilon-$differential privacy by coarsening this problem via a restatement in the Fourier basis, where far fewer free coefficients are required to guarantee privacy.

There are many unsolved problems in the application of formal privacy models and SDL to fully synthetic data. This article illustrates the common ground in the two methodologies and points out ways to implement the procedures in complex data models.

Acknowledgements

We gratefully acknowledge the support of National Science Foundation grants SES-0339191, SES-0427889 and CNS-0627680. We are also indebted to Fredrik Andersson, Cynthia Dwork, Johannes Gehrke, Dan Kifer, Ashwin Machanavajjhala, Kobbi Nissim, Jerry Reiter, and Adam Smith for valuable ongoing discussions of the issues discussed herein.

References

Rubin, D.B.: Discussion of statistical disclosure limitation. Journal of Official Statistics 9, 461–468 (1993)

Barak, B., Chaudhuri, K., Dwork, C., Kale, S., McSherry, F., Talwar, K.: Privacy, accuracy, and consistency, too: A holistic solution to contingency table release. In: PODS 2007 (2007)

Dwork, C.: Differential privacy. In: Bugliesi, M., Preneel, B., Sassone, V., Wegener, I. (eds.) ICALP 2006. LNCS, vol. 4052, pp. 1–12. Springer, Heidelberg (2006)

Dwork, C., McSherry, F., Nissim, K., Smith, A.: Calibrating noise to sensitivity in private data analysis. In: Halevi, S., Rabin, T. (eds.) TCC 2006. LNCS, vol. 3876, pp. 265–284. Springer, Heidelberg (2006)

Machanavajjhala, A., Kifer, D., Abowd, J., Gehrke, J., Vilhuber, L.: Privacy: Theory meets practice on the map. In: International Conference on Data Engineering, ICDE 2008 (in press, 2008)

Auditing Categorical SUM, MAX and MIN Queries

Francesco M. Malvestuto

Sapienza University of Rome, Computer Science Dept.,
Via Salaria 113, 00198 Roma, Italy
malvestuto@di.uniroma1.it

Abstract. Auditing consists in logging answered queries and checking, each time that a new query is submitted, that no sensitive information is disclosed by combining responses to answered queries with the response to the current query. Such a method for controlling data disclosure naturally raises the following inference problem: Given a set Q of answered queries and a query q, is the information asked by q determined by responses to queries in Q? We solve this inference problem for sum-queries (of real type and nonnegative real type), max-queries and min-queries and provide tests running in polynomial time. To achieve this, we introduce an inference model which, unlike previous inference models, is sound and complete in that the answer to the question of the inference problem comes out to be affirmative if and only if the information asked by q coincides exactly with the value of q determined by Q.

Keywords: Aggregate function, sum-query, max-query, min-query, null values.

1 Introduction

A *statistical database* is a collection of individual data about which queries concerning values of aggregate functions for certain subsets of the individual data may be answered without revealing confidential information contained in the individual data. An example is a database allowing SUM, MAX and MIN queries. For instance, we may have a file of employees with fields NAME, GENDER, DEGREE, SALARY supporting queries of the form ‘give me the sum of salaries of all employees whose gender and degree satisfy a certain selection criterion”. What measures suffice to protect the confidentiality of the salary information? This suggests an obvious security problem: how to prevent the exact or approximate disclosure of confidential data from the answers to aggregate queries. This is the *statistical disclosure problem* and many different approaches have been proposed for dealing with this problem. Examples include perturbation of the database itself, perturbation of query answers and query restriction. Yet another approach is *to audit* the queries in order to determine when enough information has been given out so that compromise becomes possible and we focus on this approach. In the standard statement of the auditing problem, it is assumed that: (1) there is one confidential field, say A (SALARY, in the example above) and (2) the set of individual records with the A-fields removed is publicly available. Then, for each individual record r, a variable x_r is introduced for the value of A in r, and each answered query is translated into an equation such as $\sum_{r \in S} x_r = a$,

J. Domingo-Ferrer and Y. Saygın (Eds.): PSD 2008, LNCS 5262, pp. 247–256, 2008.

$\mathop{Max}\limits_{r\in S} x_r = a$ or $\mathop{Min}\limits_{r\in S} x_r = a$ where S is the set of records that satisfy the selection criterion of the query and a is the corresponding value of the aggregate function. At this point, the problem is to check whether or not some variable x_r has the same value for every solution of the resultant equation system.

The standard formulation has two limitations: (*i*) the number of variables to introduce is equal to the size of the database which might contain a huge number of records, and (*ii*) even if the value of A in a record r is not determined by the equation system, it could be disclosed by a user coalition having a prior knowledge of the values of A in a set of records S^*; this happens with SUM queries if the sum-expression $\sum\limits_{r\in S^*\cup\{r\}} x_r$ has a unique feasible value, with MAX queries if the max-expression $\mathop{Max}\limits_{r\in S^*\cup\{r\}} x_r$ has a unique feasible value and this is greater than the maximun of A in S^*, and with MIN queries if the min-expression $\mathop{Min}\limits_{r\in S^*\cup\{r\}} x_r$ has a unique feasible value and this is less than the minimum of A in S^*.

In order to overcome the limitations above, some authors [11] suggested an alternative approach to the auditing problem for SUM queries, in which assumption (2) of the standard formulation is relaxed (so that the problem comes out to be independent of the size of the database). To achieve this, the notions of "elementary categories" and "categories" are introduced. An *elementary category* or a *cell* is a logically possible combination of values of category attributes (GENDER and DEGREE, in the example above); for instance, (GENDER = Male, DIVISION = Gynecology) is not a cell. A *category* is a nonempty set of cells.

Initially, for each cell ω, the set of records $R(\omega)$ whose category descriptions match ω is determined; if $R(\omega) = \emptyset$, then ω is called a *null cell.* Analogously, a category C is called a *null category* if it consists of null cells only. Moreover, a nonnull category C is recognized as being a *sensitive target* if the distribution of A over the record set $R(\mathrm{C}) = \cup_{\omega\in \mathrm{C}} R(\omega)$ satisfies some pre-fixed sensitivity criterion (e.g., threshold criterion or concentration criterion et cetera). During the interrogation of the statistical database, the selection criterion of every query is evaluated to a (possibly null) category, which is called the *characteristic category* of the query [13]. For instance, in the example above, the cell set is reported in the following table

GENDER	DEGREE
Male	High-school
Male	Bachelor
Male	Ph.D.
Female	High-school
Female	Bachelor
Female	Ph.D.

If the selection criterion of a query is GENDER = Female & DEGREE • High-school, then the characteristic category of the query is

GENDER	DEGREE
Female	Bachelor
Female	Ph.D.

Given a SUM query with characteristic category C, say $sum(\mathsf{C}, A)$, the response to the query is the sum of values of A over $R(\mathsf{C})$, which is denoted by SUM(C, A). In order to model the amount of information conveyed by the responses to answered queries, an equation system is constructed as follows. For each cell ω, a variable x_ω is introduced to model the information released about the value at ω, and each answered query $sum(\mathsf{C}, A)$ is translated into an equation such as $\sum_{\omega \in \mathsf{C}} x_\omega = a$, where a = SUM(C, A). The problem is then to check for each sensitive target whether or not the corresponding value is uniquely determined by the equation system.

In this paper we extend this framework to MAX and MIN queries such as $max(\mathsf{C}, A)$ and $min(\mathsf{C}, A)$, for which again we can build up a system of equations such as $\underset{\omega \in \mathsf{C}}{Max}\, x_\omega = a$ or $\underset{\omega \in \mathsf{C}}{Min}\, x_\omega = a$ using the additivity of the MAX and MIN aggregate functions in order to decide whether or not a sensitive target is protected or unprotected after answering a set of queries. It should be noted that, if x_ω is a variable featured by the equation system, then the cell ω need not be nonnull; therefore, x_ω takes on not only numeric values (corresponding to the case that ω is a nonnull cell) but also a conventional non-numeric value for the case that is a null cell. This conventional value is called *null value* in database systems. We introduce a special symbol for the null value, which we call *nil* and denote by $\odot$. Accordingly, the responses to queries $sum(\mathsf{C}, A)$, $max(\mathsf{C}, A)$ and $min(\mathsf{C}, A)$ are defined as follows:

$$\text{SUM}(\mathsf{C}, A) = \begin{cases} \sum_{r \in R(\mathsf{C})} r_A & \text{if } \mathsf{C} \text{ is a nonnull category} \\ \odot & \text{else} \end{cases}$$

$$\text{MAX}(\mathsf{C}, A) = \begin{cases} \underset{r \in R(\mathsf{C})}{Max}\ r_A & \text{if } \mathsf{C} \text{ is a nonnnull category} \\ \odot & \text{else} \end{cases}$$

$$\text{MIN}(\mathsf{C}, A) = \begin{cases} \underset{r \in R(\mathsf{C})}{Min}\ r_A & \text{if } \mathsf{C} \text{ is a nonnnull category} \\ \odot & \text{else} \end{cases}$$

At this point, in order to restore the additivity of the three aggregation functions, we add the following algebraic rules:

$$a+\odot = \odot+a = a \qquad \max(a, \odot) = \max(\odot, a) = a \qquad \min(a, \odot) = \min(\odot, a) = a$$

which entail that nil behaves like zero with respect to +, like –∞ with respect to max and +∞ with respect to min. Thus, if C_1 and C_2 are two disjoint categories, then we always have

$$\begin{aligned} &\text{SUM}(\mathsf{C}_1 \cup \mathsf{C}_2, A) = \text{SUM}(\mathsf{C}_1, A) + \text{SUM}(\mathsf{C}_2, A) \\ &\text{MAX}(\mathsf{C}_1 \cup \mathsf{C}_2, A) = \max\,(\text{MAX}(\mathsf{C}_1, A), \text{MAX}(\mathsf{C}_2, A)) \\ &\text{MIN}(\mathsf{C}_1 \cup \mathsf{C}_2, A) = \min\,(\text{MIN}(\mathsf{C}_1, A), \text{MIN}(\mathsf{C}_2, A)) \end{aligned}$$

Using the algebraic rules above, we state a general criterion for the value of query q to be determined by answers to queries in Q. Finally, we explicitly solve the inference problems for sum-queries of real type, for sum-queries of nonnegative-real type, for max-queries and for min-queries by providing tests that run in polynomial time.

Related Work. (*The inference problem*) Previous inference models for sum-queries [3, 4, 5, 9, 10, 11, 12, 13, 14] and for max-queries and min-queries [3, 7] are not sound because they do not allow for null values. Moreover, for max-queries and min-queries, in [3, 7] the authors make use of an information model that requires the knowledge not only of responses to answered queries but also of the underlying database, and in [3] the author assumes that the values of the aggregation variable are all distinct. (*Null values*). To the best of our knowledge, null values have never been dealt with in a systematic way even if they make an appearance in [2].

The paper is organized as follows. Section 2 contains the formal statement of our inference problem. In Section 3 we introduce an equation system, called the "base system", which models the information content of responses to answered queries, and state a general inference criterion. In section 4, we prove that the inference problem can be solved in polynomial time for sum-queries of real type or of nonnegative real type. In section 5, we prove that the inference problem can be solved in polynomial time for max-queries and the same holds for min-queries. Section 6 contains a closing remark. Note that, for shortness, all proofs are omitted.

2 The Inference Problem

A query q on a database R will be denoted by $f(\mathsf{C}, A)$, where f stands for *sum*, *max* or *min* and C is the characteristic category of q. By $q(R)$ we denote the *value* of q on R, that is, the response to q if q is answered. Note that $q(R) = \odot$ if and only if C is a null category. A *system of queries* on R is a set of queries $Q = \{q_1, \ldots, q_n\}$ all with the same aggregation function and with the same aggregation variable. By $Q(R)$ we denote the function that maps each query q_i in Q to its value $q_i(R)$ $(1 \leq i \leq n)$. We may view $Q(R)$ as an abstract representation (or a coding) of R and, typically, the function that maps R to $Q(R)$ is not invertible because there exist (infinitely) many databases R' for which $Q(R') = Q(R)$. If $Q(R') = Q(R)$, we say that R and R' are *equivalent modulo* Q. Thus, $Q(R)$ is a unique representation of the class of databases equivalent to R modulo Q. Finally, let q be any query with same aggregation function and with the same aggregation variable as Q and assume we are given $Q(R)$ but not $q(R)$. Since $Q(R)$ is a unique representation of the class of databases equivalent to R modulo Q, knowing $Q(R)$ always allows to determine the set of values of q on databases that are equivalent to R modulo Q, each of which can be taken to be a *feasible value* of q *given* $Q(R)$. Of course, $q(R)$ is itself a feasible value of q given $Q(R)$. Finally, we say that the value of q on R is *determined* by $Q(R)$ if there is exactly one feasible value of q given $Q(R)$, which then must equal $q(R)$.

3 The Information Model

Let Q be a system of queries on a database R. We now introduce an equation system to model the amount of information conveyed by $Q(R)$.

3.1 The Base System

Let $Q = \{q_1, \ldots, q_N\}$, where $q_i = f(\mathrm{C}_i, A)$ $(1 \le i \le N)$, and let $\Omega = \{\omega_1, \ldots, \omega_M\}$ be the set of cells. First of all, we reduce Q and Ω as follows. For each i, $1 \le i \le N$, if $q_i(R) = \odot$ (that is, C_i is a null category) then we delete q_i from Q and delete each $\omega_j \in \mathrm{C}$ from Ω. Let $Q = \{q_1, \ldots, q_n\}$, $n \le N$, and $\Omega = \{\omega_1, \ldots, \omega_m\}$, $m \le M$, be the result of the reduction. For any category C, the *support* of C is the (possibly empty) set $J = \{j\colon \omega_j \in \mathrm{C}\}$. Accordingly, if J_i is the support of C_i $(1 \le i \le n)$, then one has $\mathrm{C}_i = \{\omega_j : j \in J_i\}$. Let D be the value-set of A and let $D_\odot = D \cup \{\odot\}$. For each j $(1 \le j \le m)$, we introduce a $D_\odot$-valued variable x_j which stands for the value of the query $f(\{\omega_j\}, A)$ on R. By the additivity of the aggregate function f, the variables x_j are subject to a system of n equations with constant terms $q_1(R), \ldots, q_n(R)$, which we write:

$$\Theta_{j \in J_i} x_j = q_i(R) \qquad\qquad (1 \le i \le n)$$

and call the *base system* of $Q(R)$. Here, $\Theta_{j \in J_i} x_j$ means $\sum_{j \in J_i} x_j$ or $\mathit{Max}_{j \in J_i}\, x_j$ or $\mathit{Min}_{j \in J_i}\, x_j$ depending on whether f is *sum*, *max* or *min* respectively.

Fact 1. A $D_\odot$-valued m-tuple $\mathbf{a} = (a_1, \ldots, a_m)$ is a solution of the base system of $Q(R)$ if and only if there exists a database R' equivalent to R modulo Q such that a_j is the value of the query $f(\{\omega_j\}, A)$ on R' $(1 \le j \le m)$.

Consider a query $q = f(\mathrm{C}, A)$ and let J be the support of C. If $J = \emptyset$, then C is a null category and $q(R)$ is equal to nil. Henceforth, we always limit our considerations to the case that $J \neq \emptyset$. With C we associate the expression $\Theta_{j \in J} x_j$ and we call an element v of $D_\odot$ a *feasible value* of this expression if there exists a solution $\mathbf{a} = (a_1, \ldots, a_m)$ of the base system for which

$$\Theta_{j \in J} a_j = v.$$

Moreover, the expression associated with C is an *invariant* of the base system of $Q(R)$ if it has exactly one feasible value, which will be referred to as *the value* of the invariant. Note that, by Fact 1, the feasible values of the expression associated with C are all and the only values of the query q on databases equivalent to R modulo Q. Therefore, the following holds.

3.2 Inference Criterion

Let $Q \cup \{q\}$ be a system of queries on a database R. We shall state an algebraic criterion for the value of q on R to be determined by $Q(R)$. Let $q = f(\mathrm{C}, A)$ and let J be the support of C.

Theorem 1. *Let $Q \cup \{q\}$ be a system of queries on database R. The value of q is determined by $Q(R)$ if and only if the expression associated with the characteristic category of q is an invariant of the reduced base system of $Q(R)$.*

In the next two sections, we shall apply Theorem 1 to sum-queries and max-queries to derive polynomial tests for recognizing queries whose values are determined by $Q(R)$.

4 Sum-Queries

In this section we deal with sum-queries of real type and of nonnegative real type. In both cases we provide a polynomial test to decide whether the value of the query q is determined by $Q(R)$. We discuss the two cases separately in the following two subsections.

4.1 Sum-Queries of Real Type

Consider the following equation system that is obtained from the base system of $Q(R)$ by restricting the range of each variable to $\mathfrak{R}$:

$$\sum_{j \in J_i} x_j = q_i(R) \qquad (1 \le i \le n), \quad \text{with } x_j \in \mathfrak{R}. \tag{1}$$

Note that a sum-expression that is an invariant of the base system is also an invariant of system (1); however, the converse need not hold if the sum-expression is a zero-invariant of system (1). The following states a necessary and sufficient condition for a sum-expression to be an invariant of the base system.

Theorem 2. *Let $Q \cup \{q\}$ be a system of sum-queries on database R, and let J be the support of the characteristic category of q. The value of q is determined by $Q(R)$ if and only if*

(a) *the sum-expression $\sum_{j \in J} x_j$ is an invariant of system* (1)*, and*

(b) *if it is a zero-invariant of system* (1)*, then*

(b1) *there exists no solution* **a** *of system* (1) *with $a_j = 0$ for $j \in J$, or*

(b2) *there exists i such that J_i is a subset of J.*

If this is the case, then the value of q is given by the value of the invariant $\sum_{j \in J} x_j$ of system (1).

Remark 2. Condition (b) distinguishes between the case that null values are allowed and the case that they are not [12].

From a computational point of view, conditions (a) and (b1) can be tested in polynomial time using standard methods of linear algebra [12]; moreover, it is easy to see that condition (b2) can also be tested in polynomial time.

Example 1. Let R be a statistical database whose records have two category fields `GENDER`, `DEGREE` and one data field `SCORE`. The value-sets of `GENDER` and `DEGREE` are {`Male`, `Female`} and {`High-school`, `Bachelor`, `Ph.D.`}, respectively. The cell set is reported in the Introduction. Assume that the data field `SCORE` is an $\mathfrak{R}$-valued variable. Consider the system $Q = \{q_1, q_2, q_3\}$ of sum queries, where $q_i = sum(\mathrm{C}_i, \texttt{SCORE})$, $1 \le i \le 3$, and

C_1 = {(Male, High-school), (Male, Bachelor), (Female, High-school), (Female, Bachelor)}
C_2 = {(Male, High-school), (Male, Ph.D.), (Female, High-school), (Female, Ph.D.)}
C_3 = {(Male, Ph.D.), (Female, Ph.D.)}.

The supports of C_1, C_2 and C_3 are $J_1 = \{1, 2, 4, 5\}$, $J_2 = \{1, 3, 4, 6\}$ and $J_3 = \{3, 6\}$, respectively. Assume that $q_1(R) = 2$ and $q_2(R) = q_3(R) = 1$. Thus, system (1) reads

$$\begin{cases} x_1 + x_2 + x_4 + x_5 = 2 \\ x_1 + x_3 + x_4 + x_6 = 1 \\ \qquad x_3 + x_6 = 1 \end{cases}$$

where x_1 , …, x_6 are $\Re$-valued variables.

Consider now the sum-query $q = sum(\texttt{C}, \texttt{SCORE})$ where

C = {(Male, High-school), (Male, Bachelor), (Female, Bachelor)}

The support of C is $J = \{1, 2, 5\}$ and the sum-expression associated with C is $x_1 + x_2 + x_5$. The general solution of system (1) is $(\alpha, \beta, \gamma, -\alpha, 2-\beta, 1-\gamma)$, where α, β and γ are arbitrary real numbers. Therefore, the feasible values of q are $\alpha+2$ where α is any element of $\Re$. By Theorem 2, the value of q is not determined by $Q(R)$. ■

4.2 Sum-Queries of Nonnegative Real Type

Consider the following equation system that is obtained from the base system of $Q(R)$ by restricting the range of each variable to $\Re^+$:

$$\sum_{j \in J_i} x_j = q_i(R) \qquad (1 \le i \le n), \quad \text{with } x_j \in \Re^+ \quad . \tag{2}$$

Again, a sum-expression that is an invariant of the (reduced) base system is also an invariant of system (2); however, the converse need not hold if the sum-expression is a zero-invariant of system (2).

Theorem 3. *Let $Q \cup \{q\}$ be a system of sum-queries on database R, and let J be the support of the characteristic category of q. The value of q is determined by $Q(R)$ if and only if*

(a) *the sum-expression $\sum_{j \in J} x_j$ is an invariant of system* (2), *and*

(b) *if $\sum_{j \in J} x_j$ is a zero-invariant of system* (2), *then there exists i such that J_i is a subset of J.*

If this is the case, then $q(R)$ is given by the value of the invariant $\sum_{j \in J} x_j$ of system (2).

Remark 3. Condition (b) distinguishes between the case that null values are allowed and the case that they are not [11, 12].

From a computational point of view, note that the invariance as well as the zero-invariance of a sum-expression can be tested in polynomial time [11, 12]; therefore, conditions (a) and (b) can be tested in polynomial time.

Example 2. Consider the same database R, the same query system Q and same query q as in Example 1, but assume that the data field `SCORE` is a variable of nonnegative type. Then, the general solution of system (2) is $(0, \beta, \gamma, 0, 2-\beta, 1-\gamma)$ with $0 \leq \beta \leq 2$ and $0 \leq \gamma \leq 1$. Therefore, the sum-expression $x_1 + x_2 + x_5$ associated with the characteristic category of q is an invariant with value 2. By Theorem 2, the value of q is determined by $Q(R)$ and $q(R) = 2$. ■

5 Max-Queries

In this section we consider max-queries and state a polynomial test, which *mutatis mutandis* applies to min-queries too.

Let $\leq$ denote the linear ordering of $\Re$. The operation max induces an extension of $\leq$ to $\Re_{\odot}$, denoted by $\leq_{max}$, defined as follows:

$$a \leq_{max} b \text{ if } max(a, b) = b$$

Note that for every two real numbers a and b one has $a \leq_{max} b$ if and only if $a \leq b$, and for every element a of $\Re_{\odot}$ one has $\odot \leq_{max} a$.

The base system reads:

$$\underset{j \in J_i}{Max}\, x_j = \quad q_i(r) \;\; (1 \leq i \leq n) \tag{3}$$

First of all, note that the i-th equation requires that $x_j \leq_{max} q_i(R)$ for each $j \in J_i$ in that for every feasible value v of x_j one has $v \leq_{max} q_i(R)$. More in general, for each j let $I_j = \{i: j \in J_i\}$ and let u_j be the real number defined as follows

$$u_j = \underset{i \in I_j}{Min}\, q_i(R)$$

Since $x_j \leq_{max} q_i(R)$ for each $j \in J_i$ one also has $x_j \leq_{max} u_j$ for all j. Moreover, it is easy to see that the m-tuple $\mathbf{u} = (u_1, \ldots, u_m)$ is a solution of the base system. In order to obtain an invariance test for a max-expression such as $\underset{j \in J}{Max}\; x_j$ we need some further notions.

Let $upper(J) = \underset{j \in J}{Max}\, u_j$. Since $\mathbf{u} = (u_1, \ldots, u_m)$ is a solution of the base system and $x_j \leq_{max} u_j$ for all j, one has that $upper(J)$ is a feasible value of the max-expression $\underset{j \in J}{Max}\, x_j$ and that $\underset{j \in J}{Max}\, x_j \leq_{max} upper(J)$. Let $top(J) = \{j \in J\text{: } u_j = upper(J)\}$ and $top(J_i) = \{j \in J_i\text{: } u_j = q_i(R)\}$. Trivially, $upper(J) = \underset{j \in top(J)}{Max}\; u_j$; therefore, if

$top(J) \cap top(J_i) \neq \emptyset$, then $upper(J) = q_i(R)$ so that $\underset{j \in J}{Max}\, x_j \le_{max} upper(J) = q_i(R)$; if in addition one has $top(J_i) \subseteq top(J)$, then $q_i(R) \le_{max} \underset{j \in top(J)}{Max}\, x_j \le_{max} \underset{j \in J}{Max}\, x_j \le_{max} q_i(R)$ and, hence, $\underset{j \in J}{Max}\, x_j = q_i(R)$. So, the following holds.

Theorem 4. *Let $Q \cup \{q\}$ be a system of max-queries on database R, and let J be the support of the characteristic category of q. The value of q is determined by $Q(R)$ if and only if there exists i such that $top(J_i) \subseteq top(J)$. If this is the case, then $q(R) = q_i(R)$.*

Of course, the inference criterion given by Theorem 4 can be tested in polynomial time.

Example 3. Consider the same set $Q \cup \{q\}$ as in Example 1 with $q_1(R) = 2$ and $q_2(R) = q_3(R) = 1$, but now the four queries are all max-queries. Recall that $J_1 = \{1, 2, 4, 5\}$, $J_2 = \{1, 3, 4, 6\}$, $J_3 = \{3, 6\}$ and $J = \{1, 2, 5\}$. The base system of $Q(R)$ reads

$$\begin{cases} \max(x_1, x_2, x_4, x_5) = 2 \\ \max(x_1, x_3, x_4, x_6) = 1 \\ \max(x_3, x_6) = 1 \end{cases}$$

Here, one has $u_1 = u_3 = u_4 = u_6 = 1$, $u_2 = u_5 = 2$ and $top(J_1) = \{2, 5\}$, $top(J_2) = J_2$, $top(J_3) = J_3$. Moreover, $upper(J) = 2$ and $top(J) = \{2, 5\}$. Since $top(J_1) = top(J)$, by Theorem 4 the value of q is determined by $Q(R)$ and $q(R) = q_1(R)$ (= 2). ■

Example 4. Consider the same system $Q \cup \{q\}$ as in Example 3 but with $q_1(R) = q_2(R) = q_3(R) = 0$. The base system of $Q(R)$ reads

$$\begin{cases} \max(x_1, x_2, x_4, x_5) = 0 \\ \max(x_1, x_3, x_4, x_6) = 0 \\ \max(x_3, x_6) = 0 \end{cases}$$

Here, one has $u_1 = u_2 = u_3 = u_4 = u_5 = u_6 = 0$; moreover, $top(J_i) = J_i$ $(1 \le i \le 3)$. Since, $upper(J) = 0$ and $top(J) = J$, one has that $top(J_i) \subseteq top(J)$ for no i so that, by Theorem 4, the value of q is not determined by $Q(R)$. Actually, the feasible values of q range from $\odot$ to 0. ■

6 Closing Note

Given the list of sensitive targets for the pair (f, A), where f is either *sum* or *max* or *min* and A is a confidential data field, and a current query $q = f(\mathrm{C}, A)$, auditing consists in checking that the value of no sensitive target can be inferred from responses to previously answered queries and the response to q. We have solved this inference problem by auditing only answered queries of the same type as q and considering only max-queries, min-queries and sum-queries of real type and of nonnegative real type. For sum-queries of nonnegative integral type and, hence, for count-queries, the inference problem encounters the same computational difficulties as linear integer programming problems. A direction for future research is the inference problem by relaxing the assumption that answered queries to audit are of the same type as the current query.

References

1. Bishop, Y.M., Fienberg, S.E., Holland, P.W.: Discrete Multivariate Analysis: Theory and Practice. MIT Press, Cambridge (1975)
2. Chen, M.C., McNamee, L., Melkanoff, M.A.: A model of summary data and its applications in statistical databases. In: Rafanelli, M., Klensin, J.C., Svensson, P. (eds.) Proc. IV Int. Working Conf. on Statistical & Scientific Database Management. LNCS, vol. 339, pp. 354–372. Springer, Heidelberg (1989)
3. Chin, F.: Security problems on inference control for SUM, MAX, and MIN queries. J. ACM 33, 451–464 (1986)
4. Chin, F.Y., Özsoyoglu, G.: Statistical database design. ACM Trans. Database Syst. 6, 113–139 (1981)
5. Chin, F.Y., Özsoyoglu, G.: Auditing and inference control in statistical databases. IEEE Trans. Knowl. Data Eng. 8, 574–582 (1982)
6. Farkas, C., Jajodia, S.: The inference problem: a survey. Proc. of the XXIII Int. Conf. on Very Large Databases, 36–45 (1997)
7. Kleinberg, J.M., Papadimitriou, C.H., Raghavan, P.: Auditing Boolean attributes. J. Comput. System Sci. 66, 244–253 (2003)
8. Lenz, H.J., Shoshani, A.: Summarizability in OLAP and statistical databases. In: Proc IX Int. Conf. on Scientific and Statistical Database Management, pp. 132–143 (1997)
9. Malvestuto, F.M.: The derivation problem of summary data. In: Proc. ACM SIGMOD Int. Conf. on Management of Data, pp. 82–89 (1988)
10. Malvestuto, F.M.: A universal-scheme approach to statistical databases containing homogeneous summary tables. ACM Trans. Database Syst. 18, 678–708 (1993)
11. Malvestuto, F.M., Mezzini, M., Moscarini, M.: Auditing sum-queries to make a statistical database secure. ACM Trans. Inf. Syst. Security 9, 31–60 (2006)
12. Malvestuto, F.M., Mezzini, M., Moscarini, M.: An analytical approach to the inference of summary data of additive type. Theoret. Comput. Sci. 385, 264–285 (2007)
13. Malvestuto, F.M., Moscarini, M.: Aggregate evaluability in statistical databases. In: Proc. of the XI Int. Conf. on Very Large Databases, pp. 279–286 (1989)
14. Malvestuto, F.M., Moscarini, M.: Query evaluability in statistical databases. IEEE Trans. Knowl. Data Eng. 2, 425–430 (1990)
15. Shoshani, A.: OLAP and statistical databases: similarities and differences. In: Proc. XVI ACM Symp. on Principles of Database Systems, pp. 185–196 (1997)

Reasoning under Uncertainty in On-Line Auditing

Gerardo Canfora and Bice Cavallo

Department of Engineering, University of Sannio,
Benevento, Italy

Abstract. We propose a Bayesian approach to reasoning under uncertainty in on-line auditing of Statistical Databases. A Bayesian network addresses disclosures based on probabilistic inferences that can be drawn from released data. In particular, we deal with on-line max and min auditing. Moreover, we show how our model is able to deal with the implicit delivery of information that derives from denying the answer to a query and to manage user prior-knowledge.

1 Introduction

A Statistical Database (SDB) is a database system that enables its users to retrieve only aggregate statistics (e.g., mean, max, min and count) for a subset of the entities represented in the database. Consider, for example, a company database containing salaries of employees. A user may want to determine the max or a min salary of the employees in a subset of records in the database. He/she cannot, however, be allowed to glean the salary of any one employee in particular.

In our paper, we propose a Bayesian network (BN) as a disclosure control tool in SDB, based on probabilistic inferences that can be drawn from released data.

Several methods for protecting privacy in SDBs have been suggested in the literature; see reference [1] for a survey. These methods can be classified under four general approaches: conceptual, data perturbation, output perturbation and query restriction. We focus on the query restriction approach, which prevents malicious inferences by denying some unsafe queries. In particular, we deal with the on-line auditing problem [4]-[7]-[8]-[9]-[10]. With on-line auditing, queries are answered one by one in sequence and the auditor has to determine whether the SDB is compromised when answering a new query. In references [2] and [3], we have introduced a Bayesian approach for on-line max and min auditing and we have shown, by means examples, how the model is able to capture user prior-knowledge. In this paper, we optimize the previous model and add the following original contribution:

1. we consider the case in which denial leaks information;
2. we model the case in which the probability distribution of the sensitive field is known;

J. Domingo-Ferrer and Y. Saygın (Eds.): PSD 2008, LNCS 5262, pp. 257–269, 2008.

3. we provide the results of a preliminary set of experimental trials aimed at assessing the scalability of the approach in terms of response time, size of the conditional probability table, and probability of denial.

The paper is organized as follows: Section 2 places our work in the context of previous research; Section 3 introduces notions and definitions useful in the sequel of the work; Section 4 presents the Bayesian approach for on-line max and min auditing; Section 5 discusses the experiments; Section 6 provides conclusion and future work.

2 Related Work

On-line auditing is first studied in references [5] and [13]; these query monitoring approaches completely ignore the answers to the queries and monitor the logs of all the queries.

Reference [4] considers the online sum, max, and mixed sum/max auditing problems. Both the online sum and the online max problem are shown to have efficient auditing algorithms. However, the mixed sum/max problem is shown to be NP-hard. Reference [8] considers the auditing problem for sum queries where the private attribute values are boolean.

Reference [9] focuses on sum-queries with response variable of nonnegative real type, and proposes a compact representation of answered sum-queries, called an information model in "normal form", which allows the query system to decide whether the value of a new sum-query can, or cannot, be safely released.

Reference [7] studies the problem of simulatable auditing; the authors propose an approach that considers the implicit delivery of information that derives from denying the answer to a query. They demonstrate that max queries can be audited in a simulatable paradigm under the classical definition of privacy where a breach occurs if a sensitive value is fully compromised. Moreover, max auditing under a probabilistic definition of privacy is considered in the case that the sensitive values are taken uniformly at random from the set of duplicate free points in a real interval. The same limitations are present in the on-line max and min auditing both under the classical definition of privacy and under the probabilistic one [10].

In references [2] and [3], a Bayesian approach for on-line max and min auditing is introduced and the "no duplicates" assumption is removed.

3 Preliminaries

We assume that:

- T is a table with n records;
- $K = \{1, 2, ..., n\}$;
- X and Y are two fields of T such that the elements of X represented by x_i, with $i \in K$, are distinct among them (each x_i identifies uniquely a subject) and the elements of Y, represented by y_i, are real numbers;

- the sensitive field Y has r distinct values ($r \leq n$);
- the private information takes the form of an association, $(x_i, y_i) \subseteq X \times Y$, that is a pair of values in the same tuple;
- a *l-query* q is a subset of K, that is $q = \{i_1, ...i_l\} \subseteq K$. Let us assume that $i_j < i_{j+1} \quad \forall j \in \{1, ..., l-1\}$. In this paper, we use the terms *query* and *l-query* interchangeably;
- the answer corresponding to a max query q is equal to $max\{y_{i_j} | i_j \in q\}$;
- the answer corresponding to a min query q is equal to $min\{y_{i_j} | i_j \in q\}$;
- m is the answer to a max or a min query;
- $l = |q| > 1$, because if $q = \{j\}$, clearly, y_j is breached irrespective of the value of m and the association (x_j, m) is disclosed.

References [8] and [10] define, for each element y_j, with $j \in K$, the upper bound μ_j as follows:

Definition 1. *$\forall y_j$, $\mu_j = min\{m_k | j \in q_k$ with q_k a max query and m_k the answer$\}$ is the minimum over the answers to the max queries containing j.*

In other words, μ_j is the best possible upper bound for y_j that can be obtained from the answers to the max queries. Similarly, the lower bound λ_j is defined as follows:

Definition 2. *$\forall y_j$, $\lambda_j = max\{m_k | j \in q_k$ with q_k a min query and m_k the answer$\}$ is the maximum over the answers to the min queries containing j.*

Moreover, we consider the following definition of probabilistic compromise [2],[3]:

Definition 3. *A privacy breach occurs if and only if a private association is disclosed with probability greater or equal to a given tolerance probability tol. If a private association is disclosed with tol $= 1$, then the SDB is fully compromised.*

4 Bayesian Approach for On-Line Max and Min Auditing

Given a set of max and min queries $\{q_1, q_2, ..., q_{t-1}\}$, the corresponding answers $\{m_1, m_2, ..., m_{t-1}\}$ and the current query q_t, the auditor has to decide if to deny q_t or to answer otherwise. Obviously if an auditor always denies, the privacy is never breached, but the user has no utility from the SDB. In this section we present a Bayesian approach to support auditor decisions. In our approach, a query is denied not only if the privacy is breached (see Definition 3), because an user can learns something also from a denial.

This section is organized as follows: Section 4.1 presents the probabilistic approach; Section 4.2 discusses how to represent in an efficient way a single max or min query, by means a BN; Section 4.3 presents the overall Bayesian approach for on-line max and min auditing; Section 4.4 shows how the model deals with the implicit delivery of information deriving from denial; Section 4.5 shows how the model is able to capture additional user knowledge about the probability distribution of the sensitive field.

4.1 Probabilistic Approach

In this section, we present a probabilistic approach to deal with max queries; the min case is analogous. Let $q = \{i_1, ..., i_l\}$ be a l-query and $m = max\{y_{i_1}, ..., y_{i_l}\}$ be the corresponding answer, if the auditor gives the answer m then the user knows that:

$$y_{i_j} \leq m \quad \forall i_j \in q \tag{1}$$

$$\exists \overline{k} \in \{1, ..., l\} | y_{i_{\overline{k}}} = m. \tag{2}$$

Moreover, we assume that the user has no knowledge about the probability distribution of the sensitive field. Because of this, we have the following prior probabilities:

$$P(y_{i_j} < m) = P(y_{i_j} = m) = \frac{1}{2} \quad \forall i_j \in q. \tag{3}$$

The following propositions compute the posterior probability that y_j is equal to m, for each $j \in q$, and determine the probabilistic dependencies among the sensitive values in q.

Proposition 1. *Let $q = \{i_1, ..., i_l\} \subseteq K$, then, $\forall j \in q$, the following posterior probabilities hold:*

$$P(y_j = m | max\{y_{i_1}, ..., y_{i_l}\} = m) = \frac{2^{l-1}}{2^l - 1} \tag{4}$$

$$P(y_j < m | max\{y_{i_1}, ..., y_{i_l}\} = m) = \frac{2^{l-1} - 1}{2^l - 1}. \tag{5}$$

Example 1. Let $q = \{1, 3\}$ be a max query with answer $m = 8$, if the auditor provides the answer then the user knows that it is verified one of the following cases: $y_1 = 8$ and $y_3 = 8$; $y_1 < 8$ and $y_3 = 8$; $y_1 = 8$ and $y_3 < 8$. Therefore, the user knows: $P(y_j = 8 | m = 8) = \frac{2}{3}$ and $P(y_j < 8 | m = 8) = \frac{1}{3}, \quad \forall j \in q$.

Proposition 2. *Given $q = \{i_1, ..., i_l\}$ such that $m_q = max\{y_{i_1}, ..., y_{i_l}\} = m$, given $q' \subset q$, with $l' = |q'| > 0$ such that $m_{q'} = max\{y_s | s \in q'\} = m$, then, $\forall j \in q \setminus q'$:*

$$P(y_j = m | m_q = m, m_{q'} = m) = P(y_j < m | m_q = m, m_{q'} = m) = \frac{1}{2}.$$

Example 2. Given $q = \{1, 2, 3, 4, 5\}$ and $m_q = max\{y_1, ..., y_5\} = m$, then if the user knows that $m_{q'} = max\{y_4, y_5\} = m$ then $P(y_j = m | m_q = m, m_{q'} = m) = \frac{1}{2}$, for $j = 1, 2, 3$.

Proposition 3. *Given $q = \{i_1, ..., i_l\}$ such that $m_q = max\{y_{i_1}, ..., y_{i_l}\} = m$, given $q'' \subset q$, with $l'' = |q''| > 0$ such that $m_{q''} = max\{y_s | s \in q''\} < m$, then, $\forall j \in q \setminus q''$:*

$$P(y_j = m | m_q = m, m_{q''} < m) = \frac{2^{(l-l'')-1}}{2^{l-l''} - 1}$$

$$P(y_j < m | m_q = m, m_{q''} < m) = \frac{2^{(l-l'')-1} - 1}{2^{l-l''} - 1}.$$

Example 3. Let $q = \{1, 2, 3, 4, 5\}$, if the user knows that $m_q = max\{y_1, ..., y_5\} = m$ and $m_{q''} = max\{y_2, y_3, y_4\} < m$, then $P(y_j = m | m_q = m, m_{q''} < m) = \frac{2}{3}$, for $j = 1, 5$. Instead, if $m_{q''} = max\{y_1, y_2, y_3, y_4\} < m$ then $P(y_5 = m | m_q = m, m_{q''} < m) = 1$.

4.2 Bayesian Networks and Temporal Transformation

In this section, we present a Bayesian network (BN) able to represent, in efficient way, user uncertain knowledge after a max or min query; this BN computes all the probabilities and dependencies among variables described in Section 4.1.

A BN is a probabilistic graphical model that represents a set of variables and their probabilistic dependencies [12]. A BN, also called belief net, is a directed acyclic graph (DAG) which consists of nodes to represent variables and arcs to represent dependencies between variables. Arcs, or links, also represent causal influences among the variables. The strength of an influence between variables is represented by the conditional probabilities which are summarized in a conditional probability table (CPT). If there is an arc from node A to another node B, A is called a parent of B, and B is a child of A. The set of parent nodes of a node X_i is denoted $parents(X_i)$. The size of the CPT of a node X_i depends on the number s of its states, the number n of $parents(X_i)$, and the number s_j of parent states, in the following way:

$$size(CPT) = s \cdot \prod_{j=1}^{n} s_j. \tag{6}$$

For every possible combination of parent states, there is an entry listed in the CPT. Notice that for a large number of parents the CPT will expand drastically. If node X_i has no parents, its local probability distribution is said to be *unconditional*, otherwise it is *conditional*. If the value of a node is observed, then the node is said to be an *evidence* node.

Independence of causal influence (ICI) [14] among local parent-child or cause-effect relationship allows for further factoring. ICI has been used to reduce the complexity of knowledge acquisition. The size of conditional distribution that encodes the max (or min) operator can be reduced when the n-ary max (resp. min) operator is decomposed into a set of binary max (resp. min) operators. Two well known approaches to the decomposition are: parent divorcing [11] and temporal transformation [6]. Parent divorcing constructs a binary tree in which each node encodes a binary operator. Temporal transformation constructs a linear decomposition tree in which each node encodes a binary operator. In this section, we present a temporal transformation to encode a max query, the min case is analogous. Consider the following example:

Example 4. Let $q = \{1, 2, 3\}$ be a max query, then the 3-query is decomposed into a set of binary max queries by means of a temporal transformation as shown in Fig. 1.

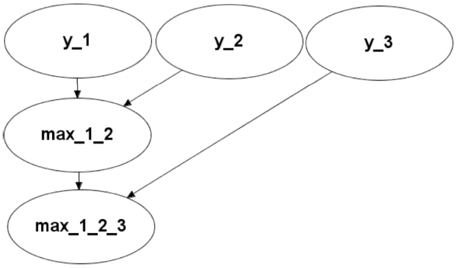


Fig. 1. Temporal transformation for a max 3-query

Given a max query $q = \{i_1, ..., i_l\}$ with answer m, for each y_j with $j \in q$, we have to represent the posterior probabilities in equation (4) and (5) of Proposition 1. Therefore, each node in the BN will have two states: r_1 is the first state and encodes the case in which the corresponding variable is less than m and r_2, the second state, encodes the case in which the variable is equal to m. As a consequence:

– the CPT for a node encoding a sensitive variable y_j has size equal to 2 and, assuming that the user has not knowledge about the domain of the sensitive field, the prior distribution is $(\frac{1}{2}, \frac{1}{2})$;
– because a node encoding a binary max query has two parents, then its CPT has size equal to $2^3 = 8$.

Because there are l nodes encoding the sensitive variables and $l - 1$ binary max nodes, then the total CPT for the BN grows linearly with the size of the query:

$$totalCPTsize = (l-1)2^3 + 2l = 8(l-1) + 2l = 10l - 8. \tag{7}$$

Example 5. Let q be the max query in Example 4 and $m = 8$ its answer, then in order to compute the posterior probabilities in (4) and in (5) of Proposition 1, we insert evidence on node encoding q as shown in Fig. 2 a).

If the user knows that $max\{y_1, y_2\} = 8$, inserting evidence on the corresponding binary max node as in Fig. 2 b), we compute, for y_1 and y_2, the probabilities given by (4) and (5), and, for y_3, the probabilities given by Proposition 2.

If the user knows that $max\{y_1, y_2\} < 8$, inserting evidence on the corresponding binary max node as in Fig. 2 c), we compute, for y_3, the probabilities given by Proposition 3.

Finally, if the user knows that $y_3 < 8$, inserting evidence on the corresponding node as in Fig. 2 d), we compute, for y_1 and y_2, the probabilities given by Proposition 3.

4.3 On-Line Max and Min Auditing

We build the BN for the on-line max and min auditing problem at run-time, that is we execute a temporal transformation after each max or min user query and decide whether or not to answer the query. From now on, we assume that:

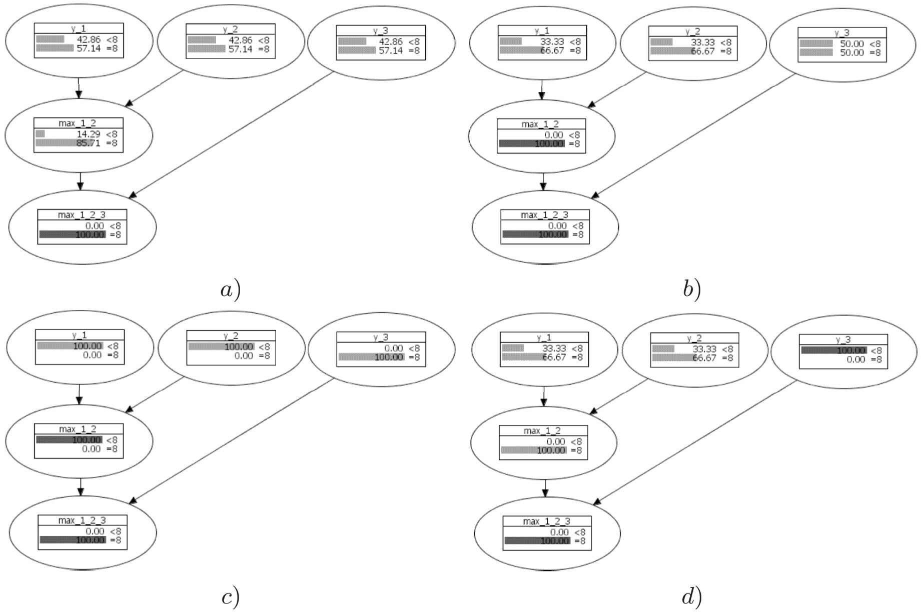


Fig. 2. Temporal transformation for a max 3-query. a) $\forall j \in \{1,2,3\}$, $P(y_j = 8|m = 8) = \frac{4}{7}$. b) $P(y_3 = 8|m = 8, max\{y_1, y_2\} = 8) = \frac{1}{2}$. c) $P(y_3 = 8|m = 8, max\{y_1, y_2\} < 8) = 1$. d) $\forall j \in \{1,2\}$ $P(y_j = 8|m = 8, y_3 < 8) = \frac{2}{3}$.

- $Z = \{z_1, ..., z_n\}$ is a permutation of $Y = \{y_1, ..., y_n\}$ such that $z_j \geq z_{j+1}$, $\forall j \in \{1, ..., n-1\}$;
- if $q = \{i_1, ..., i_l\}$, with $i_j < i_{j+1} \quad \forall j \in \{1, ..., l-1\}$, is a max l-query then $max\{z_{i_1}, ..., z_{i_l}\} = z_{i_1}$;
- if $q = \{i_1, ..., i_l\}$, with $i_j < i_{j+1} \quad \forall j \in \{1, ..., l-1\}$, is a min l-query then $min\{z_{i_1}, ..., z_{i_l}\} = z_{i_l}$;
- for each z_j, $\mu_j^{(t)}$ and $\lambda_j^{(t)}$ are respectively its best upper bound and its best lower bound after $\{q_1, ..., q_t\}$ (see Definition 1 and Definition 2).

Given $\{q_1, q_2, ..., q_t\}$ a set of max and min queries already submitted and $\{m_1, m_2, ..., m_t\}$ the set of the corresponding answers, then the BN, representing user knowledge without prior knowledge about the domain of the sensitive field, is such that:

- a node encoding the sensitive variable z_j, with $j \in \bigcap_{k=1}^{t} q_k$, has the following states:
  - r_1 and r_2, with $r_1 < r_2 = \mu_j^{(t)}$ and prior probability distribution equal to $(\frac{1}{2}, \frac{1}{2})$, if j is in only max queries;
  - r_1 and r_2, with $\lambda_j^{(t)} = r_1 < r_2$ and prior probability distribution equal to $(\frac{1}{2}, \frac{1}{2})$, if j is in only min queries;
  - r_1, r_2 and r_3, with $\lambda_j^{(t)} = r_1 < r_2 < r_3 = \mu_j^{(t)}$ and prior probability distribution equal to $(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$, if j is in max and min queries;

Table 1. $n = 5$, $r = 4$. Table with a duplicate sensitive value.

X	x_1	x_2	x_3	x_4	x_5
Z	9	8	8	5	4

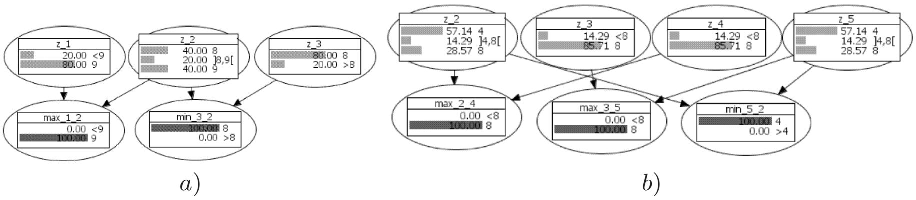


Fig. 3. Examples. a) *tol* = 0.8. Privacy is breached. b) To manage duplicates values.

- if $\exists j \in K | j \notin \bigcap_{k=1}^{t} q_k$, then there is not a node encoding the sensitive variable z_j;
- a binary max node with value equal to M has two states: r_1 and r_2, with $r_1 < r_2 = M$;
- a binary min node with value equal to m has two states: r_1 and r_2, with $m = r_1 < r_2$.

Our model is able to:

1. deny if the privacy is breached (Definition 3). See Example 6;
2. deal with “duplicated values” of the sensitive field in an efficient way. See Example 7.

In the following examples, we consider Table 1.

Example 6. Let *tol* = 0.8 be the tolerance value. If the user submits the max query $q_1 = \{1, 2\}$ and the min query $q_2 = \{2, 3\}$, then the corresponding BN is shown in Fig. 3 a). Thus, because the posterior probabilities $P(z_1 = 9|m_1 = 9, m_2 = 8) = P(z_3 = 8|m_1 = 9, m_2 = 8) = 0.8 = tol$, then the auditor has to deny the second query.

Moreover, we can see that nodes encoding z_1 and z_3 have two states because they are respectively in a max query and in a min query, the node encoding z_2 has three states because it is in a max and in a min query.

Example 7. Let *tol* = 0.9 be the tolerance value. If the user submits the max query $q_1 = \{2, 4\}$, the max query $q_2 = \{3, 5\}$ and the min query $q_3 = \{2, 5\}$, then the corresponding BN is shown in Fig. 3 b). In addition to the information given by the answers to the queries, the user also knows that $max\{z_3, z_4\}$ must be 8 since one of z_2 or z_5 has to be 4. In the previous work [10], the auditor needs to maintain, in addition to submitted queries, also inferred queries, as for instance $max\{z_3, z_4\} = 8$, with a possible blow up in the number of queries that need to be maintained. This could not happen in the absence of duplicates since the first two queries could never have the same answer. In our model, even if there is not a node encoding the query $max\{z_3, z_4\}$, this additional knowledge is captured and, moreover, if the user submits the max query $q_4 = \{3, 4\}$, there is no need to add the corresponding node.

4.4 Dealing with Implicit Delivery of Information Deriving from Denial

In this section, the implicit delivery of information, that derives from denying the answer to a query, is considered. Intuitively, denials leak information because users can ask *why* a query is denied, and the reason is in the data.

As a simple example, assume that a query is denied only if some value is compromised. Assume that the user submits the first max query $q_1 = \{1, 2, 3\}$ and the auditor answers 8. Assume also that the user then submits the second max query $q_2 = \{2, 3\}$ and the auditor denies the answer and finally the user submits the third max query $q_3 = \{1, 2\}$ and the auditor provides the answer 8. The denial tells the user that if the true answer to the second query were given then some value could be determined.

If the decision to allow or deny a query depends on the actual data, it reduces the set of possible consistent solutions for the underlying data. Thus, the decision to deny or answer the current query q_t must be independent from the actual answer m_t.

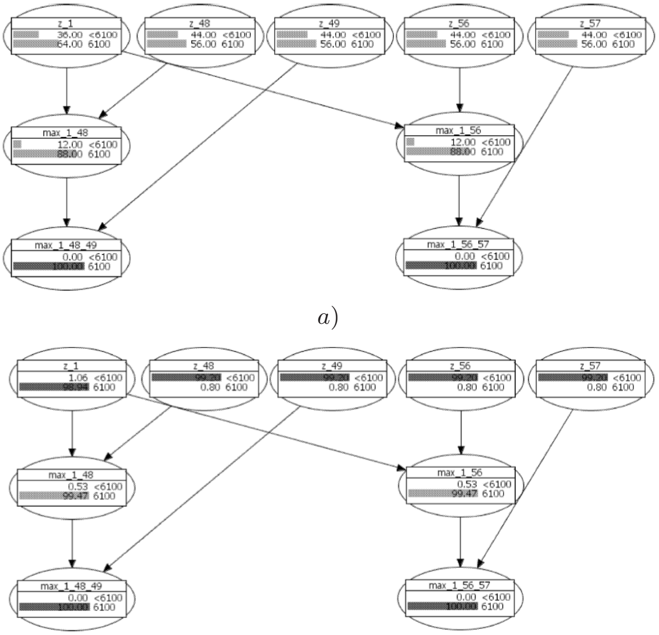
Therefore, the auditor denies the answer to a query, not only if the privacy is breached, but also in the following cases:

1. the probability that a sensitive variable is equal to a value is greater or equal to a given tolerance threshold (even if this value is not the actual value of the sensitive data item). See Example 8;
2. for a possible answer to q_t, the probability that a sensitive variable is equal to a value is greater or equal to a given tolerance threshold (even if this value is not the actual value of the sensitive data item). See Example 9.

Example 8. Let $tol = 0.8$ be the tolerance value, then the third query in Example 7 is denied, because $P(z_3 = 8|m_1 = 8, m_2 = 8, m_3 = 4) = P(z_4 = 8|m_1 = 8, m_2 = 8, m_3 = 4) = 0.8571 > tol$. We can see that the actual value of z_3 is 8, but the actual value of z_4 is 5.

In general, if there is j such that its posterior probability $P(z_j = x|m_1, ..., m_t) > tol$ then the query is denied even if $z_j \neq x$.

Example 9. Let $tol = 0.8$ be the tolerance value. The user submits the max query $q_1 = \{1, 2, 3\}$, the auditor provides the answer $m_1 = 8$; the corresponding BN is shown in Fig. 2 a). We suppose that the user submits the max query $q_2 = \{1, 2\}$: if the answer is $m_2 = 8$ (see Fig. 2 b)) then the privacy is not breached; if $m_2 < 8$ (see Fig. 2 c)) then the private association $(x_3, 8)$ is disclosed with probability equal to 1. Thus, independently from the actual value of m_2, the answer is denied.



a)

b)

Fig. 4. Comparison. a) BN does not model user knowledge on probability distribution of the sensitive field. b) BN models user knowledge on probability distribution of the sensitive field.

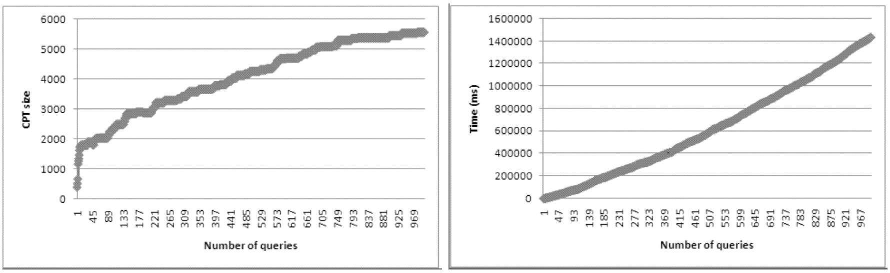
4.5 Dealing with Known Probability Distribution of the Sensitive Field

In this section, we assume that the probability distribution of the sensitive field is known. Thus, we do not consider anymore that each state is equally probable, but the prior probabilities depend on probability distribution of the sensitive field. We show how the auditor denies the answer to a query if the BN models the probability distribution of the sensitive field, and gives the answer otherwise.

Example 10. Consider the baseball dataset in [15], it consists of 377 records. We have added a field ID, in such way that $(ID, Salary)$ is the private association, with ID the field identifying the baseball player and $Salary$ the sensitive field. The dataset comprises 210 distinct values of $Salary$. We assume that the probability distribution of the sensitive field is public, and in particular it is known to the user.

Let $tol = 0.8$ be the tolerance value. Given the queries $q_1 = \{1, 48, 49\}$ and $q_2 = \{1, 56, 57\}$ then $m = m_1 = m_2 = 6100$.

If the BN does not model user knowledge about the probability distribution of the sensitive field then each node encoding a sensitive variable has prior



a) *b)*

Fig. 5. Experimentation. a) The CPT size rises to around 5500 after 1000 queries. b) The time to answer 1000 queries is around 1400000.

probability distribution equal to $(\frac{1}{2}, \frac{1}{2})$, and after q_2 the BN is in the state of Fig. 4 a). Therefore, the auditor knows that the private association $(x_1, 6100)$ is disclosed with probability equal to 0.64, then he provides the answer to q_2.

Else, if the BN models user knowledge about the probability distribution of the sensitive field then each sensitive variable has prior probability distribution equal to $(0.9973, 0.0027)$, because $P(z_j < 6100) = 0.9973 \quad and \quad P(z_j = 6100) = 0.0027$. After q_2 the BN is in the state of Fig. 4 b). Therefore, the auditor knows that the private association $(x_1, 6100)$ is disclosed with probability equal to 0.9894, then he denies the answer to q_2.

Because, it is very unlike that an user knows exactly the probability distribution of the sensitive variable, it is natural to approximate this knowledge. If $P(z_j = 6100) \approx 0.05$ then $P(z_1 = 6100 | m_1 = 6100, m_2 = 6100) \approx 0.8470$. Also in this case, the auditor has to deny the answer.

Remark 1. In Example 10, we can see that there is a large difference between the prior and the posterior probability; we think that the probabilistic definition of privacy can be improved if this difference is considered.

5 Experimentation

The experimentation is conducted on a computer with the following properties: HP Compaq dc7100; Pentium(R) 4 CPU 2.80 GHz; 2 GB of RAM. In the experimentation, we run sequences of 1000 queries and set tolerance (see Definition 3) equal to $tol = 0.8$. We consider the baseball dataset in Example 10. Each max or min query is generated in random way with length in the range $[2, ..., n]$. From Fig. 5 a), we can see that the CPT size grows quickly with the first queries, then it grows much more slowly. The time to answer a single query (see Fig. 5 b)) grows linearly with the number of queries. Table 2 reports the statistical variables for the time to answer a single query. Finally, in order to analyze the utility of our model, in Fig. 6 we show how the probability to denial grows; it rises to around 0.3 after 1000 queries.

Table 2. Experimentation. Statistical analysis for time(ms) to answer a single query.

MEAN	1435.82
STANDARD DEVIATION	1140.863
MAX	6671
MIN	125
MEDIAN	1125

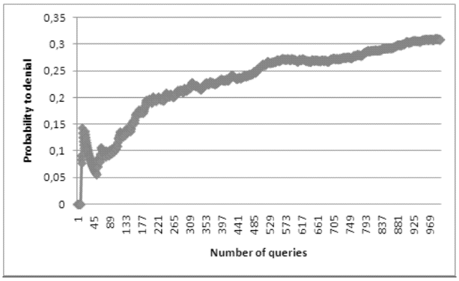


Fig. 6. The probability to denial rises to around 0.3 after 1000 queries

6 Conclusion and Future Work

We propose a novel approach to reasoning under uncertainty in on-line auditing in Statistical Databases. We have demonstrated how our model is able to:

- deal with on-line max and min auditing without maintaining query logs;
- deal with a probabilistic definition of privacy, independently from the probability distribution of the sensitive field;
- manage efficiently duplicated values of the sensitive field;
- provide a graphical representation of user knowledge;
- capture user prior knowledge;
- consider the case in which denial leaks information.

The goal of our future work is fourfold:

1. to quantify the utility of the auditing scheme, as exact analysis of utility for max and min queries is an open problem;
2. to improve the definition of probabilistic privacy, considering the difference between prior and posterior probability;
3. to use a BN as a unifying framework including the interactions among the various domains of uncertainty;
4. to include combinations of different statistical queries (sum, mean, count, etc.).

References

1. Adam, N.R., Wormann, J.C.: Security-control methods for statistical databases: A comparative study. ACM Comput. Surv. 21, 515–556 (1989)
2. Canfora, G., Cavallo, B.: A bayesian approach for on-line max auditing. In: Proceedings of the Third International Conference on Availability, Reliability and Security (ARES), pp. 1020–1027. IEEE Computer Society Press, Los Alamitos (2008)
3. Canfora, G., Cavallo, B.: A bayesian approach for on-line max and min auditing. In: Post-Proceedings of the 2008 International workshop on Privacy and Anonymity in Information Society (PAIS). ACM International Conference Proceeding, vol. 261, pp. 12–20 (2008)
4. Chin, F.Y.: Security problems on inference control for sum, max, and min queries. Journal of the ACM 33, 451–464 (1986)
5. Dobkin, D., Jones, A., Lipton, R.: Secure Databases: Protection against User Influence. ACM Trans. Database Syst. 4, 97–106 (1979)
6. Heckerman, D.: Causal independence for knowledge acquisition and inference. In: Proceedings of Ninth Conference on Uncertainty in Artificial Intelligence, pp. 122–127 (1993)
7. Kenthapadi, K., Mishra, N., Nissim, K.: Simulatable auditing. In: PODS, pp. 118–127 (2005)
8. Kleinberg, J., Papadimitriou, C., Raghavan, P.: Auditing boolean attributes. Journal of Computer and System Sciences 66, 244–253 (2003)
9. Malvestuto, F.M., Mezzini, M., Moscarini, M.: Auditing sum-queries to make a statistical database secure. ACM Transactions on Information and System Security 9, 31–60 (2006)
10. Nabar, S.U., Marthi, B., Kenthapadi, K., Mishra, N., Motwani, R.: Towards robustness in query auditing. In: 32^{nd} International Conference on Very Large Data Bases, pp. 151–162 (2006)
11. Olesen, K.G., Kjaerulff, U., Jensen, F., Jensen, F.V., Falck, B., Andreassen, S., Andersen, S.K.: A munin network for the median nerve-a case study in loops. Applied Artificial Intelligence 3, 385–403 (1989)
12. Pearl, J.: Probabilistic reasoning in intelligent systems: networks of plausible inference. Morgan Kaufmann, San Francisco (1988)
13. Reiss, S.P.: Security in databases: A combinatorial study. Journal of the ACM 26, 45–57 (1979)
14. Srinivas, S.: A generalization of the noise-or-model. In: Ninth Annual Conference of Uncertainty on AI, pp. 208–218 (1993)
15. Watnik, M.R.: Pay for play: Are baseball salaries based on performance? Journal of Statistics Education 6 (1998)

A Remote Analysis Server - What Does Regression Output Look Like?

Christine M. O'Keefe[1] and Norm M. Good[2]

CSIRO Mathematical and Information Sciences,
Preventative Health National Research Flagship
[1] GPO Box 664, Canberra, ACT 2600 Australia
[2] PO Box 10842 Adelaide Street BRISBANE QLD 4000 Australia
{Christine.O'Keefe,Norm.Good}@csiro.au

Abstract. This paper is concerned with the problem of balancing the competing objectives of allowing statistical analysis of confidential data while maintaining standards of privacy and confidentiality. *Remote analysis servers* have been proposed as a way to address this problem by delivering results of statistical analyses without giving the analyst any direct access to data. Several national statistical agencies operate successful remote analysis servers, see for example [1,12].

Remote analysis servers are not free from disclosure risk, and current implementations address this risk by "confidentialising" the underlying data and/or by denying some queries. In this paper we explore the alternative solution of "confidentialising" the output of a server so that no confidential information is revealed or can be inferred.

In this paper we first review relevant results on remote analysis servers, and provide an explicit list of measures for confidentialising the output from a single regression query to a remote server, as suggested by Sparks et al. [22,23]. We give details of a fully worked example, and compare the confidentialised output from the query to a remote server with the output from a traditional statistical package.

Keywords: Remote access facilities, remote server, analysis server, model server, confidentiality.

1 Introduction

National statistical agencies, health administration agencies and others currently face a dilemma. On the one hand, analysis of the growing electronic data archives they hold can be vital for informed decision making, effective policy development and evaluation of the impact of decisions, policies and interventions. On the other hand, the use and analysis of these data archives must be conducted in such a way as not to compromise standards of privacy and confidentiality.

A high level discussion of the problem of enabling the use of health data while protecting privacy and confidentiality typically discusses two broad categories, which are often used in combination. The first is *restricted access,* where access is only provided to approved individuals for approved analyses, possibly at a

J. Domingo-Ferrer and Y. Saygın (Eds.): PSD 2008, LNCS 5262, pp. 270–283, 2008.

restricted data centre, and possibly with further measures such as restrictions on the types of analyses which can be conducted and restrictions on the types of outputs which can be taken out of the room. The second is *restricted or altered data,* where less than the full data set is published or the data are altered in some way before publication. Restricted data might involve removing attributes or observations, aggregating geographic classifications or aggregating small groups of data. Common examples of techniques for altered data include the addition of noise, data swapping or the release of synthetic data, see [3,4,24].

Current technological methods for resolving the dilemma in practice fall into three main approaches, as follow. In the *de-identification* approach, identifying attributes are removed from the data set and the remaining fields are released to analysts with no further modification and under strict controls. In *statistical disclosure control* the released data set is de-identified and then "confidentialised" by making various modifications such as rounding, deleting values, adding random noise to data or releasing synthetic data designed to be similar to the original data [3,4,24]. *Remote analysis servers* are designed to deliver useful results of user-specified statistical analyses with acceptably low risk of a breach of privacy and confidentiality. Remote analysis servers do not release any data, or sometimes only a limited sample. Instead, they accept queries which are run on the original data and only confidentialised results are returned to the analyst.

The approaches of de-identification and statistical disclosure control fall into the restricted or altered data category. Remote analysis servers use elements of restricted access in combination with elements from the additional broad categories *restricted queries* and *restricted output.* As the names suggest, restricted queries does not allow the full range of analyses to be performed while restricted output involves some confidentialisation of the output of the server. The main challenge in the area of remote servers is to design systems for restricted output. In some applications remote servers may restrict data (for example to a random 95% sample) or alter data through making only confidentialised data available.

Each of these approaches must be implemented within an appropriate legislative and policy environment and governance structure, and with user community management and IT security including user authentication, access control, system audit and follow-up. It is useful to develop a range of solutions because different scenarios may have different requirements.

In the remainder of this Introduction we provide a review of relevant results on remote analysis servers, considering the cases of single queries and multiple interacting queries separately. In Section 2 we provide an explicit list of measures

Table 1. Categories and approaches to balancing data use with confidentiality

	De-identification	Stat. disclosure control	Remote Server
Restricted access			X
Restricted data	X	X	
Restricted queries			X
Restricted output			X

suggested by Sparks et al. [22,23] for confidentialising the output of a single regression model fitting query to a remote server. Section 3 provides the details of an example of regression fitting on a real data set, and compares the confidentialised response to the query from a remote server as well as the response from a traditional statistical package. The final section is a discussion of the results.

1.1 Remote Analysis Servers - Single Queries

This paper is about *remote analysis servers*, which do not provide data to users, but rather allow statistical analysis to be carried out remotely. A user submits a statistical query, an analysis is carried out on the original data in a secure environment, then the user receives the results of the analysis. The query could be submitted either as a user-written piece of code or through making selections on a menu-driven interface. The results of the analysis may be filtered or modified to protect confidentiality, or no results provided. For reviews of systems in use in national statistical agencies, see [13,20]. Despite the technical challenges in addressing, for example, missing data, outliers, selection bias testing and assumption checking [22], it seems to be generally agreed that remote analysis servers will play an important role in the future of data dissemination [17].

Early proposals combined a remote access server for query restriction with statistical disclosure control techniques on the underlying data set [5,6,11,21].

An important advance has been the development of *table servers* which disseminate marginal sub-tables of a large contingency table. More specifically, the target database of a table server is a large, high-dimensional contingency table whose cells contain counts or sums. Users submit requests for marginal subtables of the full tables to the server, where each sub-table is specified by the variables that it contains. Potential responses to the user include the requested table, a projection of it, an otherwise modified version of the requested table or a notice of refusal, where the choice of response is made to minimise the disclosure risk. In *static* mode, the set of allowable responses of a for a given table is pre-computed. In *dynamic* mode, the response to each query depends on the queries that the server has previously answered, in that the disclosure risk of a given query is assessed in the light of responses given to previous queries. For more information on table servers, see for example [2,9,10].

There is increasing interest among researchers and practitioners in *model servers*, which respond to requests for relevant output such as estimated parameters and standard errors from statistical models involving a response and one or more predictor variables. Most effort has been directed at linear regression. Reiter [16] noted that to be most useful the remote server should also provide some way for users to check the fit of their models, without disclosing actual data values.

Reiter [16] suggested releasing synthetic regression diagnostics - that is, simulated values of residuals and response and explanatory variables for a regression, constructed to mimic the relationships among the actual data residuals and explanatory variables . Users then treat these synthetic values like ordinary diagnostic quantities, for example by examining scatter lots of the synthetic residuals versus the synthetic explanatory variables or versus the synthetic fitted values. Several examples on simulated and real data sets are provided. For regressions involving categorical explanatory variables, in particular logistic and multinomial regressions, the release of grouped diagnostics has been proposed as a way to release diagnostics which do not reveal individual data values [18]. Several examples on simulated and real data sets are provided. Note that synthetic data is an example of the restricted data category of approaches to enabling analysis while protecting confidentiality.

All remote analysis servers are not free from the risk of disclosure, especially in the face of multiple, interacting queries [7,16,18,19,22,23].

Sparks et al. [22,23] have explored disclosure risks associated with the response generated by a single request to a remote analysis server, while still giving useful output. Disclosure risks are described for exploratory data analysis, the fitting of survival models, linear regression models, tree based models and generalized linear models, and time series analyses. Mitigating strategies are proposed which reduce the risk of a user reading or inferring any individual record attribute value. Examples from biostatistics are provided and a software demonstrator (Privacy-Preserving Analytics® (PPA®)) is described. Some of the methods proposed involve the modification or restriction of standard statistical analyses submitted through a menu-driven interface, whereas others involve modifications to the output of fitted models. In particular, they do not involve applying any traditional statistical disclosure techniques to the underlying microdata (except in the case of using a random 95% sample of the microdata in some analyses). The approaches are examples of the restricted queries and restricted output categories.

Amongst the variety of restricted queries and restricted output methods proposed for various different types of analysis, Sparks et al. [23] suggest a new approach to releasing diagnostic checks of regression fit and checks of assumptions while reducing disclosure risk. The suggestion is to replace each scatterplot (such as residuals versus each explanatory variable and residuals versus fitted value) with a display of parallel boxplots constructed on the actual data, and to replace each qq-plot of actual data values versus specified distribution quantiles by a fitted robust non-parametric regression line, after removing outliers.

A comparison of the fundamentally different approaches of Reiter [16] and Sparks et al. [23] is provided in [14].

1.2 Remote Analysis Servers - Multiple Queries

Gomatam, Karr, Reiter and Sanil [7] describe disclosure risks associated with multiple, interacting queries to model servers, primarily in the context of regression servers, and propose quantifiable measures of risk and data utility. A detailed illustration is given in the case of a static regression server responding to requests for regression analyses on data with one sensitive variable, which the agency wishes to prevent intruders being able to predict too accurately from multiple released regressions with other variables as explanatory variables. If no transformations of the variables are permitted, the authors show how to determine the optimal set of answerable queries within the set of all regression queries, by balancing disclosure risk with data utility.

In the case of multiple queries to remote analysis servers, it is still vital to ensure that each single query does not reveal confidential information, so techniques such as those described in Section 1.1 are needed as well as additional measures to address the additional risks posed by multiple, interacting queries.

2 Techniques for Confidentialising Output from a Single Query to a Remote Regression Server

Sparks et al. [23] have proposed methods by which the outputs from a range of individual statistical queries can be modified so that the results contain no directly identifying information, the exact value of any unit record is not revealed and only very little information about any unit record is revealed. These modifications produce results which are unlikely to enable the identification of individuals either directly or through inference. In this paper we provide examples of the operation of these techniques for the special case of fitting regression models.

2.1 Explicit Statement of Techniques

In this section we give an explicit statement of the techniques proposed by Sparks et al. [23] in the case of regression model fitting. The methods combine elements from each of the broad categories: restricted access, restricted data and restricted analyses of Section 1.

Restricted Access

1. If the model has one of the following properties, do not return any results.
   (a) One or more explanatory variables is a factor for which there is a level with few values.
   (b) Interactions between factors have too few values in the interaction levels.
2. As the analyst fits more and more models to the data, some subsets of the fitted models will not be permitted.
   (a) A different subset of the observations is used for each subset model.

Restricted Data

3. All queries will be run on a sample of the target data of predetermined size, depending on the level of authorisation of the analyst.
   (a) The same sample is used for the same analyst and data for all similar queries, however different types of queries will be run on different samples of the same size.

Restricted Analyses

4. Use robust regression instead of traditional regression
   (a) If using the rlm function to fit a robust linear model in R, then PPA selects automatically and randomly from the three available choices for the so-called ψ-*function* (namely, Huber, Hampel or Tukey bisquare) and use the same function whenever the same query is submitted. Note that rlm is *robust* in that it downweights influential observations/outliers.

5. The scale of the response and explanatory variables cannot be changed.
6. No new variables can be constructed from the data and no values can be added to the explanatory variables.
7. Transformations of explanatory variables is not permitted, except for a limited range of simple transformations such as log and square root.
8. No tranformation of factors is allowed.
9. The choice of Box-Cox power tranformation of the response values is automated.
10. At most two-way interactions between variables can be included in models.
    (a) Factor interactions with a small number of values in any cell are not permitted.

Restricted Output

11. The estimated covariance matrix, the fitted values and the residual values are not provided.
12. If the model error is too small, provide only the level of significance of the parameters, not the values.
13. Estimated regression coefficients are rounded, where the rounding amount is determined by the effect on the predicted response values.
    (a) The same rounding is provided for each subset model fitted.
14. Diagnostic plots are confidentialised, as follows:
    (a) A display of parallel boxplots is provided for model residuals.
    (b) Smoothed qq-plots are provided.
    (c) In plots of residuals against observation number for each subset model, the order of the observations will be randomised.

The procedure for constructing a display of parallel boxplots as in 14(a) is explained in full in [23] and [14]. In particular, each bin contains a predetermined minimum number of observations and continuous variables are discretised by using quantile methods. Also, the plots are *winsorised*, that is, any observation which is more than 2.64 standard deviations away from the mean is set to the mean plus or minus 2.64 standard deviations. This does not affect the calculation of the mean or standard deviation, it is just used in constructing the boxplot.

3 Example of Regression Model Fitting

In this section we provide a comprehensive example, describing first the data and the analysis to be performed. We display the output generated by the query to CSIRO's PPA software demonstrator. For comparison, we also display traditional, unaltered output for the same analysis of the same data from the R package [15].

For this example we will use publicly available data from a study to test the safety and efficacy of estrogen plus progestin therapy to prevent recurrent coronary heart disease in postmenopausal women. The study is called the *Heart and Estrogen/Progestin Replacement Study (HERS)* and is described in [8].

The HERS data *HERSdata* contains information on the characteristics of 2763 participants in the HERS study, who were all women younger than 80 years with coronary disease and who were postmenopausal with an intact uterus. The mean age of the participants was 66.7 years. We will use the (continuous) variables: age in years (age), education in years (educyrs), body mass index (bmi) and systolic blood pressure (sbp), and the (discrete) variables/factors: ethnicity (raceth), diabetes comorbidity (diabetes), insulin used (insulin), previous coronary artery bypass graft surgery (pcabg), at least one drink per day (drinkany) and attendance at exercise program or regular walking (exercise).

We wish to fit a regression model with sbp as the response variable and all other characteristics as explanatory variables.

In the next two sections we will discuss the fitting of this model using the PPA software and with the traditional *glm* function of R, showing details of the inputs and outputs. In the case of PPA we show how the analysis inputs and outputs implement the measures discussed in Section 2.

3.1 PPA Analysis

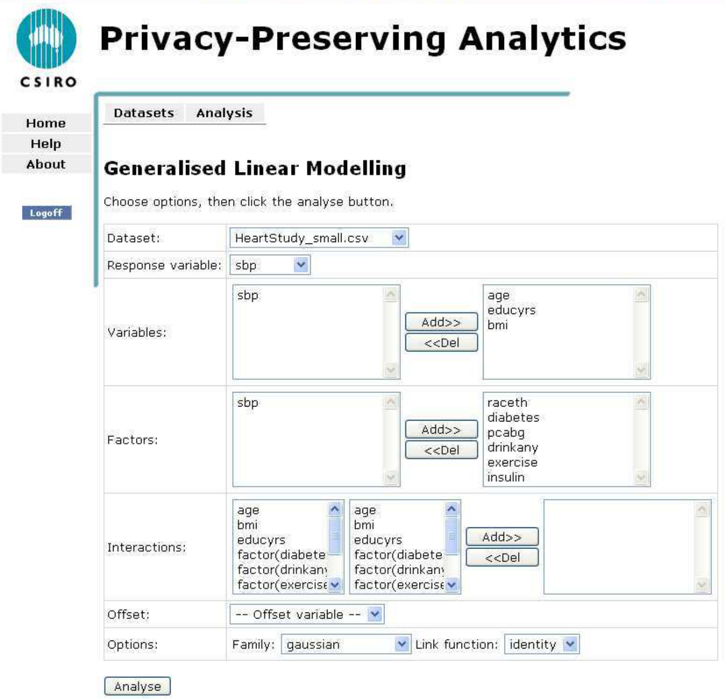
Generalised linear modelling in PPA is offered through a menu-driven interface. The user selects, from drop-down menus, the data and response variable. In the existing early prototype of the PPA software, the user selects from lists the continuous variables and factors as predictors, but these will be deduced automatically from the response variable and metadata in future versions. The user selects from lists the interactions to be included in the model. Further drop-down menus allow the user to select the offset variable, the family and the link function. Finally, pressing the *Analyse* button submits the specified query.

Figure 1 shows the PPA input screen for the analysis described. The use of the menu-driven screen implements the Restricted Analyses techniques 5, 6, 7, 8, 9 and 10 described in Section 2 for protecting confidentiality. The Restricted Access techniques 1 and 2, Restricted Data technique 3 and Restricted Analyses technique 4 are implemented in the software behind the interface.

The PPA software provides the following output to the analysis, implementing the Restricted Output techniques 11, 12, 13 and 14 of Section 2.

Details of the call to R

- datafile - HERSdata.csv
- response - sbp
- mainlist - c(age, educyrs, bmi)
- factorlist - c(raceth, diabetes, pcabg, drinkany, exercise, insulin)
- interlist - c() (interactions)
- infamily - gaussian
- inlink - identity
- variance - constant
- offset - NULL
- scriptFile - linearmodel/glm.r
- zeroInflation - FALSE
- overDispersion - FALSE
- subsetSize - 0.95



Privacy-Preserving Analytics

Generalised Linear Modelling

Choose options, then click the analyse button.

Dataset: HeartStudy_small.csv

Response variable: sbp

Variables: sbp | age, educyrs, bmi

Factors: sbp | raceth, diabetes, pcabg, drinkany, exercise, insulin

Interactions: age, bmi, educyrs, factor(diabete, factor(drinkan), factor(exercise | age, bmi, educyrs, factor(diabete, factor(drinkan), factor(exercise

Offset: -- Offset variable --

Options: Family: gaussian Link function: identity

Fig. 1. Screen shot of PPA query input interface

- rows in initial data: 2763; rows in sampled data: 2620
- ∼ age + educyrs + bmi + factor(raceth) + factor(diabetes) + factor(pcabg) + factor(drinkany) + factor(exercise) + factor(insulin)
- Call: glm(formula = as.formula(paste(response, formula)), family = eval(familyexp), data = indata) where *indata* is a random 95% sample of the HERS data.

Analysis of Deviance table. Terms added sequentially (first to last)

	Df	Deviance	Resid. Df	Resid. Dev
NULL			2619	949964
age	1	27997	2618	921968
educyrs	1	10207	2617	911761
bmi	1	6065	2616	905696
factor(raceth)	2	1894	2614	903802
factor(diabetes)	1	14145	2613	889656
factor(pcabg)	1	6814	2612	882842
factor(drinkany)	1	95	2611	882747
factor(exercise)	1	3216	2610	879531
factor(insulin)	1	202	2609	879329

Summary results

Coefficients:	Estimate	Pr(> | t |)
(Intercept)	103.748	p<0.005 ***
age	0.530	p<0.005 ***
educyrs	-0.560	p<0.005 ***
bmi	0.124	0.05<p<0.1 .
factor(raceth)Latina, Asian, other	-3.164	0.1<p<0.2
factor(raceth)white	-2.319	0.05<p<0.1 .
factor(diabetes)yes	5.449	p<0.005 ***
factor(pcabg)yes	3.462	p<0.005 ***
factor(drinkany)yes	-0.449	p>0.5
factor(exercise)yes	-2.336	p<0.005 **
factor(insulin)yes	-1.124	0.2<p<0.5

Significance codes: "0" 0; "***" 0.001; "**" 0.01; "*" 0.05; "." 0.1; " " 1.
Dispersion parameter for gaussian family taken to be 337.0369.
Null deviance: 949964 on 2619 degrees of freedom.
Residual deviance: 879329 on 2609 degrees of freedom.
AIC: 22697.
Number of Fisher Scoring iterations: 2.

Data and Diagnostic Plots

Figure 2 shows displays of parallel boxplots provided as confidentialised output for each explanatory variable against the response variable sbp. Use of the label

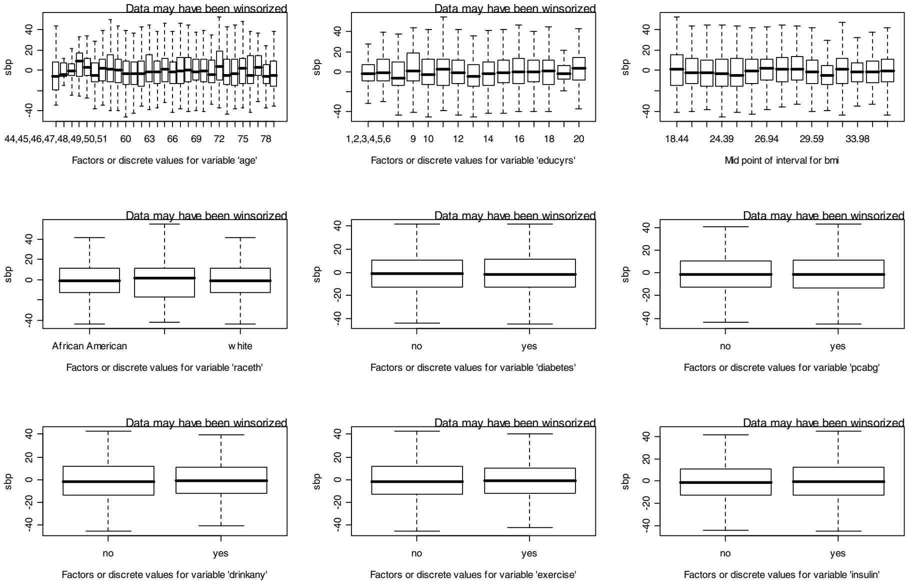


Fig. 2. PPA confidentialised plots of explanatory variables against response variables

Factors or discrete values for variable ... on the x-axis indicates that the variable being plotted is either discrete or a factor. The x-axis label *Mid point of interval for* ... indicates that the variable being plotted is continuous, but has been discretised. The note *Data may have been winsorized* as described in Section 2.

Other plots provided by PPA as confidentialised output include displays of parallel boxplots of partial residuals for each variable in the model as in Figure 3. Normality assumptions can be assessed with smoothed qq-plots. If there are outliers present the following text appears on the plot: *"Some extreme values may have been removed."* In such cases it may be more appropriate to fit robust regression models.

For space reasons the confidentialised plot of residual values against observation numbers is not shown, however this figure is also drawn as a display of parallel boxplots.

3.2 Traditional R Analysis

A generalised linear model is fitted in R by submitting the following query to the *glm* function:

$$\text{Model} = \text{glm}(\text{sbp} \sim \text{age} + \text{educyrs} + \text{bmi} + \text{factor}(\text{raceth}) + \text{factor}(\text{diabetes}) + \text{factor}(\text{pcabg}) + \text{factor}(\text{drinkany}) + \text{factor}(\text{exercise}) + \text{factor}(\text{insulin})).$$

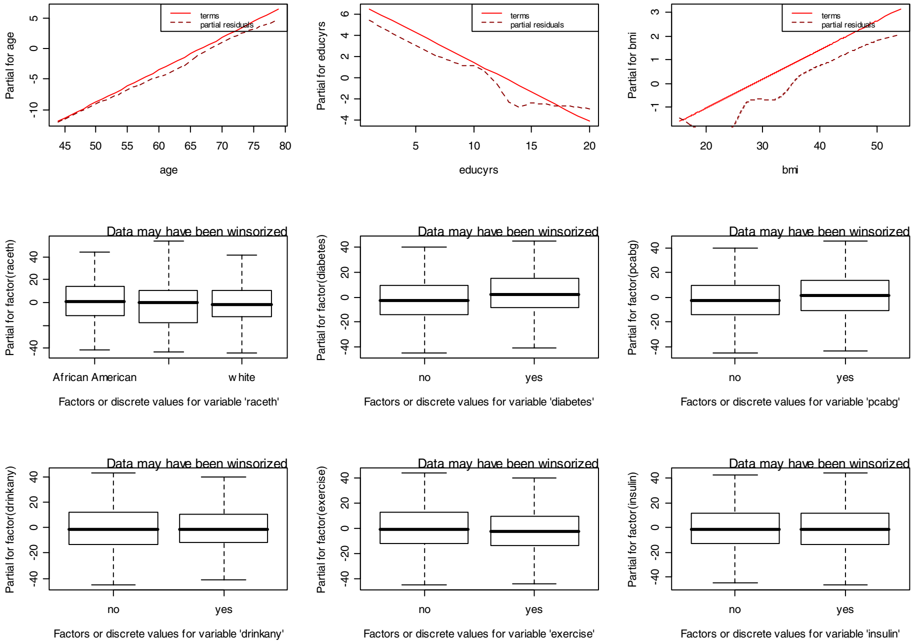


Fig. 3. PPA confidentialised term plots for each explanatory variable

The resulting *model object* contains information that can be accessed in a number of ways.

The command *Anova(Model)* will output the analysis of deviance table showing the effect of addition of each variable to the model. The Analysis of Deviance table is in the same format as the PPA confidentialised Analysis of Deviance table given in Section 3.1 above. The only difference is that the R analysis is run on the full HERSdata data, while PPA was run only on the random 95% sample indata.

The command *Summary(Model)* will output the Summary Results as shown below. Further options allow the analyst to print covariance and/or correlation matrices and to explore aliased coefficients. Other output contained within the model object includes the model residuals, fitted values and the ability to fit predicted values.

Summary results

Coefficients:	Estimate	Std. Error	t value	Pr(> | t |)
(Intercept)	105.89708	4.75504	22.270	$< 2e^{-16}$ ***
age	0.50444	0.05376	9.384	$< 2e^{-16}$ ***
educyrs	-0.60095	0.13800	-4.355	$1.38e^{-05}$ ***
bmi	0.11480	0.06763	1.698	0.08970 .
factor(raceth) Latina, Asian, other	-2.60109	2.27871	-1.141	0.25377
factor(raceth)white	-2.07604	1.33316	-1.557	0.11953
factor(diabetes)yes	5.77120	0.98320	5.870	$4.89e^{-09}$ ***
factor(pcabg)yes	3.58667	0.71624	5.008	$5.86e^{-07}$ ***
factor(drinkany)yes	-0.33294	0.75055	-0.444	0.65737
factor(exercise)yes	-2.46810	0.73606	-3.353	0.00081 ***
factor(insulin)yes	-1.15651	1.41227	-0.819	0.41291

Significance codes: “0” 0; “***” 0.001; “**” 0.01; “*” 0.05; “.” 0.1; “ ” 1.
Dispersion parameter for gaussian family taken to be 336.0279.
Null deviance: 996807 on 2753 degrees of freedom.
Residual deviance: 921725 on 2743 degrees of freedom.
9 observations deleted due to missingness.
AIC: 23849.
Number of Fisher Scoring iterations: 2

Deviance Residuals

Min	1Q	Median	3Q	Max
-52.32	-12.68	-1.31	11.13	77.29

Data and Diagnostic Plots

The user would typically investigate a number of diagnostics, such as the term plots in Figure 4.

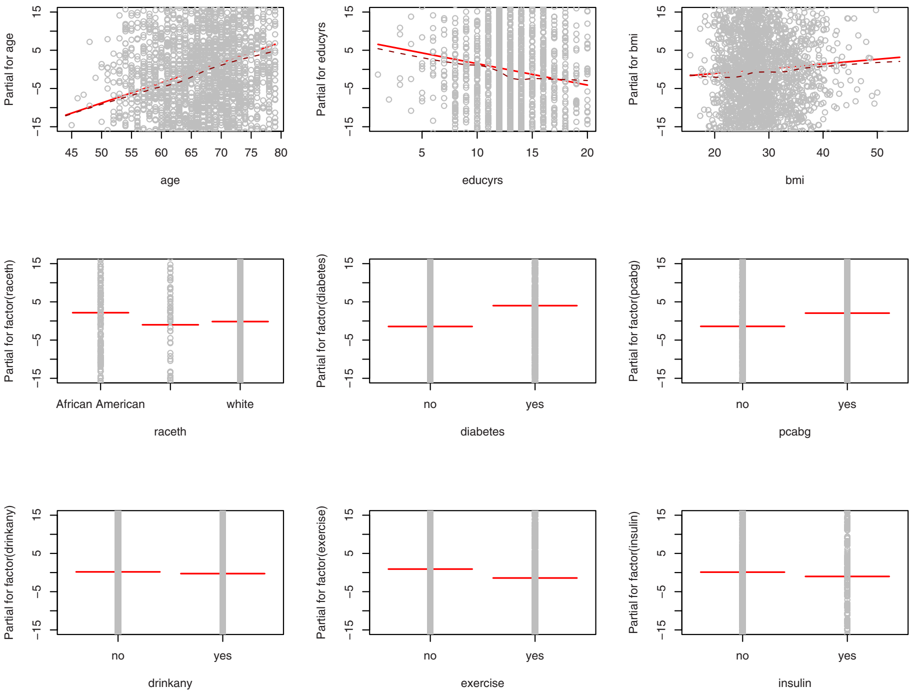


Fig. 4. Traditional term plots of residuals against variables

4 Discussion

This paper has provided detailed examples of confidentialised output from a remote analysis server and of unaltered data from a traditional statistical package, using the same analysis query on the same data.

The first observation to make is that the remote analysis server restricts the queries that can be submitted by providing only a menu-driven interface. Some other confidentialisation measures are implemented in software behind the interface, though the only one that restricts or alters the data in any way is the selection of a random 95% sample of the data for analysis.

In terms of model output, the remote analysis server does not provide the Deviance Residuals displayed by R. However it does provide an Analysis of Deviance table, parameter estimates and diagnostic plots which are confidentialised versions of those provided by the statistical package R. The use of a random 95% sample of the data in PPA leads to sampling error and therefore the parameter estimates from PPA and the traditional analysis may differ.

In situations in which a data custodian does not wish to give analysts access to a confidential data set, we believe that this example provides evidence that the confidentialised regression model fitting output of a remote analysis server may be an acceptable substitute for traditional regression model fitting output in some applications. Although analysts prefer to have unrestricted access to data, a remote analysis server seems to be an acceptable alternative to no access at all.

Finally, in some models there may be sensitive factors that may identify an entity such as a hospital or surgeon, for example. Future versions of PPA will allow an analyst to include such factors in a model, however their parameter estimates will not be published.

References

1. Australian Bureau of Statistics Remote Access Data Laboratory (RADL), http://www.abs.gov.au/websitedbs/D3310114.nsf/home/CURF:+Remote+Access+Data+Laboratory+(RADL)?OpenDocument
2. Dandekar, R.A., Domingo-Ferrer, J., Torra, V.: Maximum utility-minimum information loss table server design for statistical disclosure control of tabular data. In: Domingo Ferrer, J., Torra, V. (eds.) Privacy in Statistical Databases. LNCS, vol. 3050. Springer, Berlin (2004)
3. Domingo-Ferrer, J., Torra, V. (eds.): Privacy in Statistical Databases. LNCS, vol. 3050. Springer, Berlin (2004)
4. Doyle, P., Lane, J.I., Theeuwes, J.J.M., Zayatz, L.: Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies. Elsevier, Amsterdam (2001)
5. Duncan, G.T., Mukherjee, S.: Microdata Disclosure Limitation in Statistical Databases: Query Size and Random Sample Query Control. In: Proceedings of the 1991 IEEE Symposium on Security and Privacy, pp. 278–287 (1991)
6. Duncan, G.T., Pearson, R.W.: Enhancing access to microdata while protecting confidentiality: prospects for the future. Statistical Science 6, 219–239 (1991)
7. Gomatam, S., Karr, A.F., Reiter, J.P., Sanil, A.: Data dissemination and disclosure limitation in a world without microdata: A risk-utility framework for remote access servers. Statistical Science 20, 163–177 (2005)
8. Grady, D., Applegate, W., Bush, T., Furberg, C., Riggs, B., Hulley, S.B.: Heart and Estrogen/progestin Replacement Study (HERS): Design, Methods, and Baseline Characteristics. Controlled Clinical Trials 19, 314–335 (1998)
9. Karr, A.F., Lee, J., Sanil, A.P., Hernandez, J., Karimi, S., Litwin, K.: Web-based systems that disseminate information but protect confidentiality. In: McIver, W.M., Elmagarmid, A.K. (eds.) Advances in Digital Government: Technology, Human Factors and Public Policy, pp. 181–196. Kluwer, Amsterdam (2002)
10. Karr, A.F., Dobra, A., Sanil, A.P.: Table servers protect confidentiality in tabular data releases. Communications of the ACM 46 (2003)
11. Keller-McNulty, S., Unger, E.A.: A database system prototype for remote access to information based on confidential data. Journal of Official Statistics 14, 347–360 (1998)
12. Luxembourg Income Study, www.lisproject.org
13. O'Keefe, C.M.: Privacy and the Use of Health Data - Reducing Disclosure Risk, electronic. Journal of Health Informatics 3(1), e5 (2008)
14. O'Keefe, C.M., Good, N.: Risk and Utility of Alternative Regression Diagnostics in Remote Analysis Servers. In: Proceedings of the 55th Session of the ISI International Statistical Institute, Lisbon, Portugal, 22-29 August (2007)

15. The R Project for Statistical Computing, www.r-project.org
16. Reiter, J.P.: Model diagnostics for remote-access regression servers. Statistics and Computing 13, 371–380 (2003)
17. Reiter, J.P.: New Approaches to Data Dissemination: A Glimpse into the Future (?). Chance 17, 12–16 (2004)
18. Reiter, J.P., Kohnen, C.N.: Categorical data regression diagnostics for remote servers. Journal of Statistical Computation and Simulation 75, 889–903 (2005)
19. Reznek, A.P.: Recent Confidentiality Research Related to Access to Enterprise Microdata. In: Prepared for the Comparative Analysis of Enterprise Microdata (CAED) Conference, Chicago IL, USA (2006)
20. Rowland, S.: An examination of monitored, remote access microdata access systems. In: National Academy of Sciences Workshop on Data Access, October 16-17 (2003)
21. Schouten, B., Cigrang, M.: Remote access systems for statistical analysis of microdata. Statistics and Computing 13, 371–380 (2003)
22. Sparks, R., Carter, C., Donnelly, J., Duncan, J., O'Keefe, C.M., Ryan, L.: A framework for performing statistical analyses of unit record health data without violating either privacy or confidentiality of individuals. In: Proceedings of the 55th Session of the International Statistical Institute, Sydney (2005)
23. Sparks, R., Carter, C., Donnelly, J., O'Keefe, C.M., Duncan, J., Keighley, T., McAullay, D.: Remote Access Methods for Exploratory Data Analysis and Statistical Modelling: Privacy-Preserving Analytics$^{\mathrm{TM}}$. Comput. Methods Programs Biomed (to appear, 2008)
24. Willenborg, L., de Waal, T.: Elements of Statistical Disclosure Control. Lecture Notes in Statistics, vol. 155. Springer, Heidelberg (2001)

Accuracy in Privacy-Preserving Data Mining Using the Paradigm of Cryptographic Elections

Emmanouil Magkos[1], Manolis Maragoudakis[2],
Vassilis Chrissikopoulos[1], and Stefanos Gridzalis[2]

[1] Department of Informatics, Ionian University
Plateia Tsirigoti 7, Kerkyra, Greece, 49100
{emagos,vchris}@ionio.gr

[2] Department of Information and Communications Systems Engineering
University of the Aegean, Karlovassi, Samos
{mmarag,sgritz}@aegean.gr

Abstract. Data mining technology raises concerns about the handling and use of sensitive information, especially in highly distributed environments where the participants in the system may by mutually mistrustful. In this paper we argue in favor of using some well-known cryptographic primitives, borrowed from the literature on large-scale Internet elections, in order to preserve accuracy in privacy-preserving data mining (PPDM) systems. Our approach is based on the classical homomorphic model for online elections, and more particularly on some extensions of the model for supporting multi-candidate elections. We also describe some weaknesses and present an attack on a recent scheme [1] which was the first to use a variation of the homomorphic model in the PPDM setting. In addition, we show how PPDM can be used as a building block to obtain a Random Forests classification algorithm over a set of homogeneous databases with horizontally partitioned data.

Keywords: Privacy and accuracy; Homomorphic encryption; Distributed systems; Frequency mining.

1 Introduction

Data mining aims at extracting valuable, non obvious information from large quantities of data [2]. This technology has broad applications in areas related to market research, as well as to financial and scientific research. Despite the potentials for offering valuable services, there have been concerns about the handling and use of sensitive information by data mining systems. The problem is even more intense nowadays with the proliferation of the Web and ICT technologies, and the progress in network, storage and processor capacity, where an enormous pool of private digital data can be easily gathered, or inferred from massive collections of public data such as Facebook.com, by using well-known data mining techniques. Even when access to sensitive data is controlled, public data can sometimes be used as a path towards private data [3].

J. Domingo-Ferrer and Y. Saygın (Eds.): PSD 2008, LNCS 5262, pp. 284–297, 2008.

Privacy concerns may also prevent building accurate data mining models. Traditionally, data mining algorithms have been applied in centralized collections of data. With distributed databases, data may be *horizontally* or *vertically* partitioned among a set of mutually mistrustful sites, where each site may hold similar data about different people or different data about the same set of people, respectively. This is also known as the *Server to Server* (S2S) model [4]. In a fully distributed setting, also known as the *Client to Server* (C2S) model [4], customers may be on hold of their own collections of sensitive data. Such data may need to be correlated with other clients' data, for example in order to provide some useful service. The traditional data warehousing approach, where dispersed data are gathered into a central site for building the data mining model, raises privacy concerns as organizations and people are reluctant to reveal their private data for legal, commercial or personal reasons. The simple approach of performing data mining at each site independently and then combine the results (*e.g.*, [5]) cannot always be possible (*e.g.*, in the $C2S$ setting) or accurate enough [6].

The need for privacy in statistical databases is driven by law, compliance, ethics, as well as for practical reasons: it would enable collaboration between data holders (*e.g.*, customers, organizations), if they were assured that their sensitive information would be protected. To this end, *Privacy Preserving Data Mining* (PPDM) has been evolved as a new branch of research in the data mining community [7]. Especially in distributed statistical databases, where there is a need to extract statistical information (*e.g.*, sum, average, entropy, Information Gain, etc) without compromising the privacy of the individuals [8].

A very common approach in the PPDM literature has been *data perturbation*, where original data are perturbed and the data mining model is built on the randomized data. For example, the data perturbation approach has been used for *classification* [9] and building *association rules* [10,11]. Typically, such approach involves a trade-off between two conflicting requirements: the privacy of the individual data and the accuracy of the extracted results [8,12]. In addition, there are cases where the disclosure of some randomized data about a client may reveal a certain property of the client's private information, an attack known as *privacy breach* [10,12]. Alternatively, and orthogonally to our research, the privacy preserving issue can also be regarded as an *access control* problem concerning aggregate data in more or less controlled environments. In this regard, multilevel and multilateral security in database information systems (*e.g.*, [13]), trusted platforms, query restriction policies and inference control [8,14,15] as well as anonymization techniques [16] have also been proposed in the literature.

Traditionally, the use of cryptographic primitives has also been well studied by the database security community [17]. In the academic literature for PPDM, following the line of work that begun with Yao [18], most theoretical results are based on the *Secure Multiparty Computation* (SMC) approach (*e.g.* [19,6,20]). SMC protocols are interactive protocols, run in a distributed network by a set of entities with private inputs, who wish to compute a function of their inputs in a privacy preserving manner. The goal is that no more information is revealed to an entity in the computation than can be inferred from that participant's input and output [21]. SMC has been used for mining association rules on both horizontally [20] and vertically partitioned databases [6]. Classification models that use the SMC approach involve decision trees [19,22] and a naive Bayes Classifier for horizontally partitioned data [23], as well as decision trees for vertically partitioned data [24]. A disadvantage of this approach is that SMC protocols require multiple communication rounds among the participants, and privacy usually comes at a high performance and communication cost [22]. Most protocols in the SMC family are efficient as long as the number of participants is kept small (*e.g.*, two or three parties).

Our Contribution. In this paper we explore whether it is possible to use efficient cryptography in order to perform privacy preserving data mining, *e.g.*, in statistical databases, while maintaining the accuracy of the results. To this end we argue in favor of borrowing knowledge from a broad literature dealing with cryptographic elections via the Internet. We discuss some weaknesses and describe an attack on a recent PPDM scheme of Yang, Zhong and Wright [1] which, to our best knowledge, was the first work that used a variation of the classical homomorphic model [25] for online elections. Our PPDM approach will be based on the classical homomorphic model of Cramer, Gennaro and Schoenmakers [25] for online elections, and more particularly on some recent extensions proposed in [26,27] for multi-candidate elections. We show how this approach could be used to mine frequencies on a large set of customer databases. As an example, we also propose the use of PPDM as a building block to obtain a Random Forests classifier learning algorithm over a set of homogeneous databases with horizontally partitioned data.

2 PPDM Based on the Homomorphic (Election) Model

We argue that research for privacy preserving data mining could borrow knowledge from the vast body of literature on Internet voting systems [28]. These systems are not strictly related to data mining but they exemplify some of the difficulties of the multiparty case. Such systems also tend to balance well the efficiency and security criteria, in order to be implementable in medium to large scale environments. Furthermore, these systems fall within our distributed computing scenario and have similar architecture and security requirements. In an Internet election for example, an election authority receives several encrypted *yes/no* votes (*e.g.*, yes $= 1$ and no $= 0$) and declares the winning candidate. In this setting the goal is to protect the *privacy* of the voters (*i.e.*, unlinkability between the identity and the vote that has been cast), while also establishing eligibility of the voters, *accuracy* and verifiability for the election result.

The most efficient schemes in the literature for cryptographically secure online elections follow the *homomorphic model* [25]. This model is a general framework that allows usage of any randomized encryption scheme with several "nice" algebraic properties, in order to protect the privacy of the encrypted votes and establish accuracy of the decrypted results in a universally verifiable way. With homomorphic encryption there is an operation $\oplus$ defined on the message space and an operation $\otimes$ defined on the cipher space, such that the "product" of the encryptions of any two private inputs is the encryption of the "sum" of the inputs: $E(M_1) \otimes E(M_2) = E(M_1 \oplus M_2)$. This property allows, for example, either to tally votes as aggregates or to combine shares of votes (*e.g.*, [29,30]), without decrypting single votes.

In [25] each client signs and then submits an encryption of her vote to a *bulletin board* [31], together with a *zero-knowledge* proof [32] that the vote is valid. The homomorphic model does not require interactions between clients, and only one flow of data is sent to the server. Privacy is established in a strong cryptographic sense: original inputs are encrypted using the randomized encryption scheme to preclude *chosen-plaintext* [33] attacks on the published encryptions; in addition, no encrypted input is ever decrypted, but instead it is combined with the other inputs to get the encrypted aggregate. The homomorphic property of the encryption scheme allows every participant to verify that the final results are accurate, by performing a multiplication of the encrypted inputs and comparing the encrypted aggregate to the value published on the bulletin board. Robustness in such protocols is established by using *threshold cryptography* [34], where the power of the election authority is divided among a set of n independent servers, in a way that a set of $t \leq n$ honest servers are able to cooperate and compute the decrypted outcome. As a result, the privacy of clients is assured against any coalition of less than t servers.

Compared with the election setting, the threat model in PPDM seems to be relaxed. Adversaries in distributed systems for data mining are considered as *semi-honest* (also referred to as *honest but curious*) [22]. This means that they are legal participants that follow the protocol specification but try to learn additional information given all the messages that were exchanged during the protocol. This fact favours the adoption of the homomorphic model in PPDM systems. First of all, there is no need for clients to construct complex zero-knowledge proofs on the correctness of their inputs. Furthermore, a strong notion of privacy for cryptographic elections, known as *receipt-freeness* or *uncoercibility* [35] is not an issue here, as the scenario of coercing the clients to reveal (or, sell) their private inputs does not seem realistic in the PPDM setting. In addition, the *universal verifiability* requirement in online elections, where any outsider is able to verify correctness of the final tally, can also be relaxed and replaced with a requirement for *atomic verifiability* (*i.e.*, where every participant in the protocol is able to verify the accuracy of the results). For all these reasons, we may be able to construct and choose among lightweight versions of some well-known cryptographic schemes for online elections that follow the homomorphic model, and adopt them to our PPDM setting.

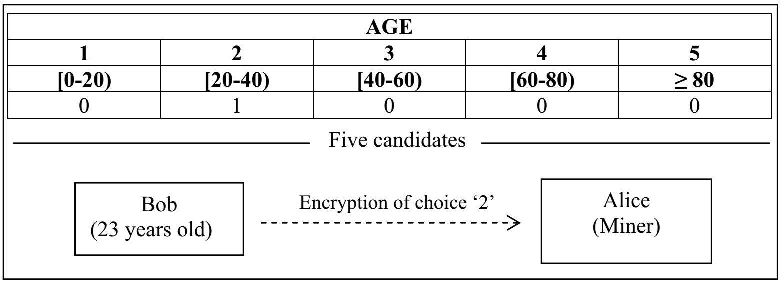
2.1 Extending the Classical Homomorphic Model

In this section we look at some very efficient extensions of the homomorphic model, where *1-out-of-L* or *k-out-of-L* selections are allowed (*e.g.*, [26,27]). In this way, the overall bits of information that a database sends to the miner could be increased, leading to new possibilities.

Multi-candidate protocols have been first investigated in [29] and further studied in [25], where the computation of the final tally grows exponentially with the number L of candidates: $\Omega(\sqrt(C)^{L-1})$, with C being the number of clients. Baudron et al [26] proposed the use of the *Paillier cryptosystem* [36] for conducting homomorphic elections with multiple candidates. The Paillier scheme provides a trapdoor to efficiently compute the discrete logarithm, thus making computation of the tally very efficient, even for large values of C and L. They also presented a threshold version of the Paillier cryptosystem, to be used in the election setting. We briefly recall the Paillier cryptosystem, leaving out some complex cryptographic details on the key generation and decryption functions [36]. Let $N = pq$ be an RSA modulus where p and q are two large primes, and g be an integer of suitable order modulo N^2. The public key is (N, g) while the secret key is the pair (p, q). To encrypt a message $M \in Z_n$, choose a random $x \in Z_n$ and compute $c = g^M x^N (modN^2)$. The knowledge of the trapdoor (p, q) allows to efficiently decrypt c and determine M. The reader may refer to [36,26] for further details.

The protocol in [26] is a *1-out-of-L* protocol, where all the choices are in the set $(1, M, M^2, ...M^{L-1})$, with M being an integer larger than the number C of clients. A client, who wishes to select the m^{th} candidate, encrypts her input with the Paillier scheme, and then signs and publishes the result $c = g^{M^m} x^N (modN^2)$ on a bulletin board. During the tallying stage, the authorities compute the "product" of the encrypted inputs and then cooperate to decrypt the tally using threshold Paillier decryption [26]. The decrypted tally can then be written in M-ary notation: $T = k_0 M^0 + k_1 M^1 + ... + k_{L-1} M^{L-1} (modN)$, which will directly reveal all k_i's, where $0 \leq k_i \leq C$ is the number of selections for candidate i. The decryption process is publicly verifiable, due to the homomorphic property of the Pallier scheme [26].

In a more recent work, Damgard et al [27] also proposed a generalization of the Paillier cryptosystem and discussed its applicability to homomorphic elections. The size of each ciphertext in [27] is logarithmic in L, while the work for computing the final tally is also reduced, compared with [26]. They also proposed a threshold variant of the generalized system.

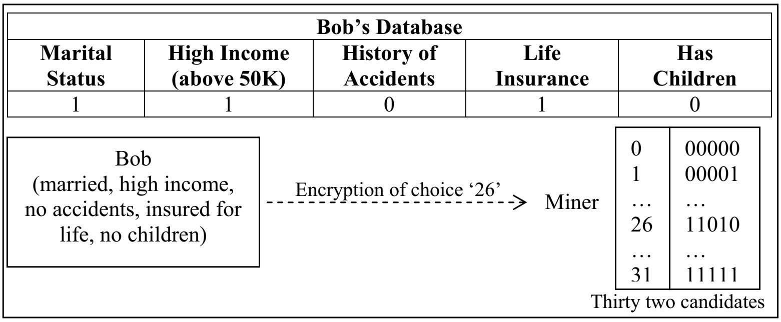


AGE				
1	**2**	**3**	**4**	**5**
[0-20)	**[20-40)**	**[40-60)**	**[60-80)**	**≥ 80**
0	1	0	0	0

Fig. 1. A multi-candidate setting with *1-out-of-5* choices

PPDM-1 Approach: Mining with *1-out-of-L* Protocols. In the usual (*yes/no*) setting (*e.g.*, yes = 1 and no = 0), a client who does not want to participate may give false information. Or, in a fully distributed setting for example, where each client retains control of her transactions, the client may decide not to participate, although we consider this as a privacy violation. As a result, the null input should also be considered in homomorphic protocols. Furthermore, knowledge cannot always be represented with *yes/no* decisions. For example, a client may have to answer which group (*e.g.*, among L age groups) her age belongs to. These are some reasons why we are interested in multi-candidate schemes, where in the simplest *1-out-of-L* case each client makes one selection out of L candidates. For simplicity, we assume that all questions to a database can be reduced to a set of *yes/no* answers, as shown in Figure 1.

PPDM-2 Approach: Mining with *k-out-of-L* Protocols. *1-out-of-L* protocols can easily be adapted to support up to *k-out-of-L* selections. An easy generalization, with some loss of efficiency, would be to send up to k encrypted messages [27]. Our proposal is to encode all possible L-bit numbers as separate candidates, thus producing a set of 2^L candidates. Figure 2 depicts our trivial approach in the fully distributed setting, where the problem of allowing *k-out-of-L* selections from a database record with L features is reduced to a *1-out-of-*2^L multi-candidate protocol. Protocols with up to *k-out-of-L* selections could also be used in a partially distributed scenario, where the full database is horizontally partitioned into a small set of client partitions, with each client possessing R full records of customers' transactions. In this case, Bob would send R encrypted messages to the miner, where R is equal to the rows of the table in his database.



Bob's Database				
Marital Status	**High Income (above 50K)**	**History of Accidents**	**Life Insurance**	**Has Children**
1	1	0	1	0

Bob
(married, high income, no accidents, insured for life, no children)
Encryption of choice '26'
Miner

0	00000
1	00001
…	…
26	11010
…	…
31	11111

Fig. 2. A trivial way to turn a *1-out-of-L* scheme into a *k-out-of-L* scheme

3 Reviewing the (Yang et al) Scheme

In this section we briefly describe the work in [1], which is, to our best of knowledge, the first scheme that used a variant of the homomorphic election model in order to build a privacy preserving frequency mining algorithm. This algorithm is then used in [1] as a building block to design a protocol for naive Bayes learning. The authors in [1] also discuss the application of this algorithm to other data mining techniques such as decision trees and association rule mining. A fully distributed setting is considered, where the clients' database is horizontally partitioned, and every client possesses her own data.

We briefly describe the PPDM protocol of [1], where a miner mines a large number of client data sets to compute frequencies of values. Let G be a group where the Discrete Logarithm problem is hard. All operations are done $modp$, where p is a publicly known and sufficiently large prime number. In a system with n clients, each client possesses two pairs of keys: $(x_i, X_i = g^{x_i})$, $(y_i, Y_i = g^{y_i})$, with g being a (publicly known) generator of the group G. Each client U_i knows her private keys (x_i, y_i), with values (X_i, Y_i) being the corresponding public keys. Furthermore, the protocol requires all clients to know the values X and Y, where $X = \prod_{i=1}^{N} X_i$, and $Y = \prod_{i=1}^{N} Y_i$. Each client is able to give a *yes/no* answer d_i to any question posed by the miner and the miner's goal is to learn $\sum_{i=1}^{N} d_i$. In the protocol of [1], depicted in Figure 3, all clients in the system use a variant of the ElGamal encryption scheme [37]. For correctness and privacy analysis, please refer to [1].

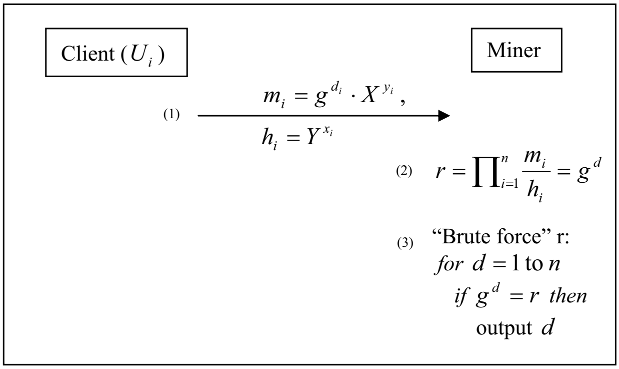


Fig. 3. A schematic representation of the protocol in [1]

Observe that the computation of the tally (*i.e.*, the result d that equals the sum of the plaintext inputs) in the scheme of [1], as well as in the classical homomorphic model of [25], involves a brute-force attack on the value g^d in order to find the discrete logarithm. This stands because there are no trapdoors to determine d from g^d in ElGamal variants [25]. In settings with only two candidates (*e.g.*, *yes/no*) this is a relatively easy computation, at least for a moderately sized number of clients. However the same is not true for multi-candidate selections in large-scale systems. To address this issue, in Section 2.1 we discussed some very efficient protocols for computing the tally in multi-candidate protocols with very large numbers of clients.

3.1 Weaknesses and Attacks

We briefly describe two weaknesses of the protocol in [1]. The first weakness is a minor one and refers to the need that each client must choose new x_i and y_i values after each execution of the protocol. This is actually a requirement in every randomized encryption scheme, where new randomness is used to increase the cryptographic security of the encryption process against chosen plaintext attacks [33]. For example, in Figure 3, if the client U_i uses the same x_i and y_i values during two successive runs, it will be trivial for an attacker (in the semi-honest model) to find out U_i's answers by trial and error.

The above weakness cannot be considered as an attack, since the authors in [1] write a remark about the need for updating the (x_i, y_i) values. However we rather consider this as a scalability issue: Prior to the execution of each run of the protocol (*e.g.*, possibly during a key setup phase) each client must obtain or compute the numbers X and Y which are functions of the public keys (X_i, Y_i) of all system clients. In a fully distributed and large-scale scenario, where a very large number of system clients hold their own records, it may be difficult to pre-compute and/or publicize these values, turning the key setup phase into a complex procedure, especially in cases where the number of participants is not constant through different runs of the system.

A DOS Attack. We also discuss a second weakness of the scheme in [1], which, under preconditions, could lead to a *Denial Of Service* (DOS) attack. We are unaware of any mention of this attack in the literature, and therefore briefly describe it here. We argue that a single client may be able to disrupt the system. Indeed, in a system with say three clients U_1, U_2, U_3, let us assume that U_2 does not send her input, because of a system crash. Then the protocol executes as in Figure 4 and a result cannot be found.

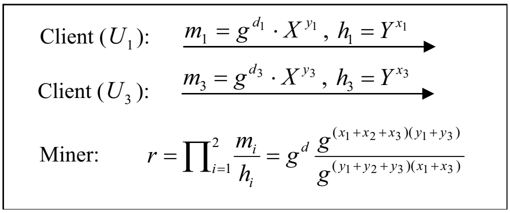

Client (U_1): $m_1 = g^{d_1} \cdot X^{y_1}$, $h_1 = Y^{x_1}$
Client (U_3): $m_3 = g^{d_3} \cdot X^{y_3}$, $h_3 = Y^{x_3}$
Miner: $r = \prod_{i=1}^{2} \frac{m_i}{h_i} = g^d \frac{g^{(x_1+x_2+x_3)(y_1+y_3)}}{g^{(y_1+y_2+y_3)(x_1+x_3)}}$

Fig. 4. A run with two active clients in a system with three registered clients

One could argue that in the semi-honest threat model, all clients will adhere to the protocol specification and will not abstain from the protocol, however this is an unrealistic assumption, especially in large-scale protocols (*e.g.*, 10000 was the number of clients used in the experimental results in [1]). Furthermore, the semi-honest model does not preclude non-malicious system crashes or network failures. Observe that a client does not know a priori who will participate in the protocol, so the obvious fix of constructing the values X and Y as a function of the number of active participants will not work.

4 A Generic Random Forests (RF) Classifier

4.1 Introducing Standalone RF

Nowadays, numerous attempts in presenting ensemble of classifiers towards increasing the performance of the task at hand have been introduced. A plethora of them has portrayed state-of-the-art results in terms of precision and recall measures. Examples of such techniques are Adaboost, Bagging and Random Forests [38].

Random forests [39] are a combination of tree classifiers such that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest. The generalization error of a forest of tree classifiers depends on the strength of the individual trees in the forest and the correlation between them. Using a random selection of features to split each node yields error rates that compare favorably to Adaboost, and are more robust with respect to noise. While traditional tree algorithms spend a lot of time choosing how to split at a node, Random Forests put very little effort into this. Compared with Adaboost, Random Forests portray the following characteristics:

1. The accuracy is as good as Adaboost and sometimes better.
2. They are relatively robust to outliers and noise.
3. They are faster than bagging or boosting.
4. They provide useful internal estimates of error, strength, correlation and variable importance.
5. They are simple and easily parallelized.

A random forest multi-way classifier $\Theta(\mathrm{x})$ consists of a number of trees, with each tree grown using some form of randomization. The leaf nodes of each tree are labeled by estimates of the posterior distribution over the data classes. Each internal node contains a test that best splits the space of data to be classified. A new, unseen instance is classified by sending it down every tree and aggregating the reached leaf distributions. The process is depicted in Figure 5:

Randomness can be injected at two points during training: in sub-sampling the training data so that each tree is grown using a different subset; and in selecting the node tests. In our approach, we shall discuss the former situation, and argue that using privacy preserving protocols in randomly selected instance vectors supports the creation of robust RF, thus allowing for effective Data Mining in horizontally-partitioned data sets. For vertically partitioned type of partitioned data, the latter approach needs to be taken into consideration. However, for the time-being this is out of the paper's scope.

4.2 Privacy-Preserving RF for Horizontally Partitioned (HP) Data

By the term horizontally partitioned data, we mean that parties (≥ 3) collect values from the same set of features but for different objects. Their goal is to

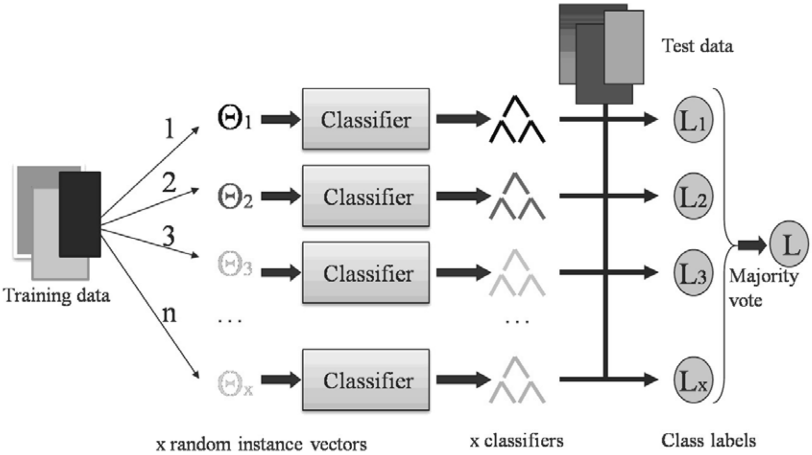


Fig. 5. Hierarchical decomposition of an RF classifier on a non-distributed data set

find an improved function for predicting class values of future instances, yet without sharing their data among each other. Thus, we enroll a collaborative approach where data need to be shared in a secure manner, and the final model will predict class labels without knowing the origin of the test instance. Similar to previous approaches such as [20], classification is performed individually, on each party's site, but the main contribution on the field is that during training, data from other parties are used in order to enhance randomness, thus increase the obtained classification accuracy. However, an assumption needs to be taken into account: data are sharing a common distribution. For example, suppose we have three different bank institutions, sharing financial information on their customers in a HP manner (*e.g.*, they all use features such as *age, occupation, income, marital status* and *sex*). In order to have a robust RF classifier, data has to follow a similar distribution among banks, meaning that if one bank owns data on a specific group of customers (*e.g.*, professors) and the others own data about a totally different group (*e.g.*, farmers), the obtained accuracy would be severely deteriorated. We exploit the two strengths of RF *i.e.*, randomness and voting. The former deals with the issue of premature termination of the tree learning process while the latter confronts data sparseness problems in an effective manner. In this work, we shall provide a protocol that allows for injecting randomness into trees during learning and allow voting over the majority class among all parties at classification time. More specifically, we shall discuss Random Input Forests (RI) learning from HP data sets abd using the forest structure to classify previous unseen test instances originating from one of the distributed database parties. Prior to this analysis, an introduction to *Out-Of-Bag* (OOB) selection of samples is included.

OOB Estimates to Monitor Error, Strength, and Correlation. Out-of-bag samples for tree T_i in a forest are those training examples that are not used to construct tree T_i. As [39] portrayed, they give unbiased estimates of error on future data, since we do not need to use cross validation. Furthermore, oob samples enhance internal strength and correlation. By strength, we denote the notion of a tree being able to be a fairly good model on its own. Correlation among trees is related with the fact that errors are canceled out between different trees. Therefore, our framework uses the following procedure: Each new training set is drawn, with replacement, from the training set of the other parties. Then a tree is grown on the new training set using random feature selection. The trees grown are not pruned. Exact measures of strength and correlation are described in [39,40] and will not be explained so forth.

4.3 Random Input Forests

Our privacy-preserving protocol for training random forests at each party by inserting randomness from different ones is consisted of two distinct phases. At the former one, each party is collaborating using the procedure proposed by [20], in order to collect the whole set of available values per each attribute. This knowledge is particularly important for the next phase, where each party will require a certain number of instances from the others (again, we note that more than three parties are needed). The complete algorithm is as follows:

Each party selects K trees to grow:

- Build each tree by:
  - Selecting, at random, at each node a small set of features (F) to split on (given M features). From related research, common values of F are:
    1. $F = 1$
    2. $F = \log_2(M) + 1$

    F is held constant while growing the forest. Create a random instance based on the values of the complete feature set and ask the other parties to vote if they own it. (based on the afore-mentioned PPDM approach). Since F is significantly smaller that M, the number of candidate instances that each party will create is computationally efficient to be handled by the PPDM approach.
  - For each node split on the best of this subset (using oob instances)
  - Grow tree to full length. There is no pruning.

To classify an unseen, new instance X, collect votes (again using the PPDM approach) from every tree in the forest of each party and use general majority voting to decide on the final class label.

5 Conclusions

In this paper we discussed privacy issues in distributed data mining and argued in favor of borrowing knowledge from a broad literature dealing with cryptographic elections via the Internet. The goal is to use efficient cryptography in order to perform privacy preserving data mining in statistical databases, while maintaining the accuracy of the results. We proposed a PPDM approach based on recent homomorphic schemes for multi-candidate elections [26,27] in order to cryptographically protect privacy in large-scale distributed data mining applications, without sacrificing scalability and efficiency. We reviewed a recent scheme [1] that used a variation of the classical homomorphic model [25] for online elections, discussed some weaknesses and described a security attack. Furthermore we proposed the use of the PPDM approach as a building block to obtain a Random Forests classification algorithm over a set of homogeneous databases with horizontally partitioned data. The introduction of a privacy preserving classifier from the domain of ensemble classifiers is a novelty of this work since such approaches have presented the most promising results as regards to precision and recall measures in real-world Data Mining applications.

We believe that research in cryptographic PPDM must be continued and practical solutions that balance the tradeoff between efficiency and security must be sought. More particularly, further research on cryptographic PPDM should take into account the various kinds of databases to work with, as well as the various data mining technologies that need to be supported.

References

1. Yang, Z., Zhong, S., Wright, R.N.: Privacy-preserving classification of customer data without loss of accuracy. In: SDM 2005 SIAM International Conference on Data Mining (2005)
2. Chen, M.S., Han, J., Yu, P.S.: Data mining: An overview from a database perspective. IEEE Transactions on Knowledge and Data Engineering 08, 866–883 (1996)
3. Clifton, C., Marks, D.: Security and privacy implications of data mining. In: 1996 ACM SIGMOD Workshop on Data Mining and Knowledge Discovery, Montreal, Canada, pp. 15–19 (1996)
4. Zhang, N., Wang, S., Zhao, W.: A new scheme on privacy-preserving data classification. In: KDD 2005: Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining, pp. 374–383. ACM, New York (2005)
5. Prodromidis, A., Chan, P., Stolfo, S.J.: Meta-learning in distributed data mining systems: Issues and approaches. Advances in Distributed and Parallel Knowledge Discovery, 81–114 (2000)
6. Vaidya, J., Clifton, C.: Privacy preserving association rule mining in vertically partitioned data. In: KDD 2002: Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 639–644. ACM, New York (2002)
7. Verykios, V.S., Bertino, E., Fovino, I.N., Provenza, L.P., Saygin, Y., Theodoridis, Y.: State-of-the-art in privacy preserving data mining. SIGMOD Rec. 33, 50–57 (2004)
8. Adam, N.R., Wortmann, J.C.: Security-control methods for statistical databases: A comparative study. ACM Comput. Surv. 21, 515–556 (1989)
9. Agrawal, R., Srikant, R.: Privacy-preserving data mining. In: Proc. of the ACM SIGMOD Conference on Management of Data, pp. 439–450. ACM Press, New York (2000)

10. Evfimievski, A., Srikant, R., Agrawal, R., Gehrke, J.: Privacy preserving mining of association rules. In: KDD 2002: Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 217–228. ACM, New York (2002)
11. Rizvi, S.J., Haritsa, J.R.: Maintaining data privacy in association rule mining. In: VLDB 2002: Proceedings of the 28th international conference on Very Large Data Bases, VLDB Endowment, 682–693 (2002)
12. Liu, K., Kargupta, C.G.,, H.: A survey of attack techniques on privacy-preserving data perturbation methods. In: Aggarwal, C., Yu, P. (eds.) Privacy-Preserving Data Mining: Models and Algorithms. Springer, Heidelberg (2008)
13. Morgenstern, M.: Security and inference in multilevel database and knowledge-base systems. SIGMOD Rec. 16, 357–373 (1987)
14. Domingo-Ferrer, J. (ed.): Inference Control in Statistical Databases. LNCS, vol. 2316. Springer, Heidelberg (2002)
15. Woodruff, D., Staddon, J.: Private inference control. In: CCS 2004: Proceedings of the 11th ACM conference on Computer and communications security, pp. 188–197. ACM, New York (2004)
16. Samarati, P., Sweeney, L.: Generalizing data to provide anonymity when disclosing information (abstract). In: PODS 1998: Proceedings of the seventeenth ACM SIGACT-SIGMOD-SIGART symposium on Principles of database systems, p. 188. ACM, New York (1998)
17. Maurer, U.: The role of cryptography in database security. In: SIGMOD 2004: Proceedings of the 2004 ACM SIGMOD international conference on Management of data, pp. 5–10. ACM, New York (2004)
18. Yao, A.C.C.: How to generate and exchange secrets (extended abstract). In: FOCS, pp. 162–167 (1986)
19. Lindell, Y., Pinkas, B.: Privacy preserving data mining. In: Bellare, M. (ed.) CRYPTO 2000. LNCS, vol. 1880, pp. 36–54. Springer, Heidelberg (2000)
20. Kantarcioglu, M., Clifton, C.: Privacy-preserving distributed mining of association rules on horizontally partitioned data. IEEE Trans. on Knowl. and Data Eng. 16, 1026–1037 (2004)
21. Goldwasser, S.: Multi party computations: past and present. In: PODC 1997: Proceedings of the sixteenth annual ACM symposium on Principles of distributed computing, pp. 1–6. ACM, New York (1997)
22. Pinkas, B.: Cryptographic techniques for privacy-preserving data mining. SIGKDD Explor. Newsl. 4, 12–19 (2002)
23. Kantarcoglu, M., Vaidya, J.: Privacy preserving naive bayes classifier for horizontally partitioned data. In: IEEE ICDM Workshop on Privacy Preserving Data Mining, Melbourne, FL, pp. 3–9 (2003)
24. Du, W., Zhan, Z.: Building decision tree classifier on private data. In: CRPIT 1914: Proceedings of the IEEE international conference on Privacy, security and data mining, pp. 1–8. Australian Computer Society, Inc., Darlinghurst (2002)
25. Cramer, R., Gennaro, R., Schoenmakers, B.: A secure and optimally efficient multi-authority election scheme. European Transactions on Telecommunications 8, 481–490 (1997)
26. Baudron, O., Fouque, P.A., Pointcheval, D., Stern, J., Poupard, G.: Practical multi-candidate election system. In: PODC 2001: Proceedings of the twentieth annual ACM symposium on Principles of distributed computing, pp. 274–283. ACM, New York (2001)
27. Damgard, I., Jurik, M., Nielsen, J.: A generalization of paillier's public-key system with applications to electronic voting (2003)

28. Gritzalis, D. (ed.): Secure electronic voting: trends and perspectives, capabilities and limitations. Kluwer Academic Publishers, Dordrecht (2003)
29. Cramer, R.J., Franklin, M., Schoenmakers, L.A., Yung, M.: Multi-authority secret-ballot elections with linear work. Technical report, Amsterdam, The Netherlands (1995)
30. Schoenmakers, B.: A simple publicly verifiable secret sharing scheme and its application to electronic voting. In: Wiener, M. (ed.) CRYPTO 1999. LNCS, vol. 1666, pp. 148–164. Springer, Heidelberg (1999)
31. Benaloh, J.D.C.: Verifiable secret-ballot elections. PhD thesis, New Haven, CT, USA (1987)
32. Goldreich, O., Micali, S., Wigderson, A.: Proofs that yield nothing but their validity or all languages in np have zero-knowledge proof systems. J. ACM 38, 690–728 (1991)
33. Diffie, W., Hellman, M.E.: New directions in cryptography. IEEE Transactions on Information Theory IT-22, 644–654 (1976)
34. Desmedt, Y.G., Frankel, Y.: Threshold cryptosystems. In: Brassard, G. (ed.) CRYPTO 1989. LNCS, vol. 435, pp. 307–315. Springer, Heidelberg (1990)
35. Hirt, M., Sako, K.: Efficient receipt-free voting based on homomorphic encryption. In: Preneel, B. (ed.) EUROCRYPT 2000. LNCS, vol. 1807, pp. 539–556. Springer, Heidelberg (2000)
36. Paillier, P.: Public-key cryptosystems based on discrete logarithms residues. In: Eurocrypt 1999. LNCS, vol. 1592, pp. 221–236. Springer, Heidelberg (1999)
37. Gamal, T.E.: A public key cryptosystem and a signature scheme based on discrete logarithms. In: Blakely, G.R., Chaum, D. (eds.) CRYPTO 1984. LNCS, vol. 196, pp. 10–18. Springer, Heidelberg (1985)
38. Breiman, L.: Bagging predictors. Machine Learning Journal 26, 123–140 (1996)
39. Breiman, L.: Random forests. Machine Learning Journal 45, 32–73 (2001)
40. Breiman, L.: Looking inside the black box. In: Wald Lecture II, Department of Statistics, California University (2002)

A Privacy-Preserving Framework for Integrating Person-Specific Databases

Murat Kantarcioglu[1], Wei Jiang[2], and Bradley Malin[3]

[1] University of Texas at Dallas, Department of Computer Science
muratk@utdallas.edu
[2] Purdue University, Department of Computer Science
wjiang@cs.purdue.edu
[3] Vanderbilt University, Department of Biomedical Informatics
b.malin@vanderbilt.edu

Abstract. Many organizations capture personal information, but the quantity of records needed to detect statistically significant patterns is often beyond the grasp of a single data collector. In the biomedical realm, this problem has pressed regulatory agencies to require funded investigators to share research-derived data to public repositories. The challenge; however, is that shared records must not reveal the identity of the subjects. In this paper, we extend a secure framework in which data holders contribute and query encrypted person-specific data stored on a third party's server. Specifically, we develop protocols that enable data holders to merge personal records, thus creating larger profiles and diminishing duplication. The repository administrator can merge records via encrypted identifiers without decrypting or inferring the contents of the joined records. Our model is more practical than prior secure join methods because each data holder needs only a single interaction with the central repository. We further present an extension to the protocol that permits the revelation of k-anonymous demographics, such that the administrator can perform joins more efficiently with the guarantee that each record can be linked to no less than k individuals in the population. We prove the privacy preserving features of our protocols and experimentally evaluate their efficiency in a real world Census dataset.

1 Introduction

The tensions between data sharing and data privacy are felt in many environments. In this paper, we focus on recent issues in the biomedical community, which illustrates real policy and technology challenges, but also opportunities for solutions. Consider, in the United States, the National Institutes of Health (NIH) expects the timely sharing of final research data from investigators receiving $500,000 or more in any year of an NIH-funded grant [1]. The goal is to facilitate the dissemination of data generated with NIH funds for use by other researchers. More recently, the NIH recognized that the integration of information technology into healthcare, in combination with the decreasing cost of DNA sequencing and storage technologies, has enabled the collection of detailed

J. Domingo-Ferrer and Y. Saygın (Eds.): PSD 2008, LNCS 5262, pp. 298–314, 2008.

patient-specific health and genetic data [2]. Thus, in its policy for genome wide association studies (GWAS), the NIH specifies that all data derived through NIH-funded GWAS studies, whether it be DNA sequences or information derived from an individual's medical record, must be shared to an NIH-managed centralized repository [3]. These are noble endeavors, designed to facilitate information reuse, but they are challenging because, at the same time, investigators must ensure that the anonymity of their subjects is not compromised. Failure to do so will have adverse legal and social consequences that significantly harm public support of biomedical research. It is important to recognize this problem is not unique to the United States. Across the globe, organizations are working to integrate person-specific DNA and clinical information from disparate facilities in the hopes of generating statistically significant research results [4,5,6].

In earlier work, we began to address this challenge with the introduction of a privacy enhancing framework that permits data holders to submit and query biomedical data housed in a centralized repository managed by a third party [7]. The framework enables data holders to store person-specific data, such as DNA sequences, on a third party's server in an encrypted format. The cryptographic basis of the framework is homomorphic, which enables the third party to execute queries for researchers, such as count queries, without decrypting the records. For example, a researcher can ask "How many DNA sequence contain pattern X?" and, while the third party will learn the result (i.e., the fraction of records that satisfy the criteria), it cannot learn which records satisfied the criteria.

The knowledge gained through count queries; however, does not support certain biomedical applications. For instance, in the healthcare realm, patients are mobile and their data can be collected by multiple locations, such as when a patient visits one hospital for primary care and a second hospital to participate in a clinical trial [8]. To facilitate robust biomedical investigations and prevent duplication of entries, it is beneficial to merge data that corresponds to the same patient. In traditional databases, such merges are achieved through joins on common attributes. The framework presented in [7]; however, is based on semantic security, so equivalent identifiers will appear different after encrypted. Therefore, we need to develop a new protocol to achieve join queries.

We exploit the fact that many organizations, such as healthcare providers, collect identifying information on their consumers. For instance, it is common for hospitals to use a patient's Social Security Number (SSN) and/or demographics for administrative purposes [9,10]. Such identifiers have been validated as excellent keys for data merging [11], but their disclosure is strictly prohibited by most organizations' policies and federal regulations. Despite restrictions, we can share SSNs, and other identifiers, in an encrypted manner for data merging purposes when the encryptions are semantically secure [12]. In this paper, we demonstrate how to join patient-specific identifiers within an existing encrypted framework. Moreover, by utilizing the concept of k-anonymity

[13,14], we propose an approach to speed up encrypted joins by revealing person-specific specific features to limit the number of potential joins that must be evaluated.

The rest of the paper is organized as follows: Section 2 highlights the existing work that is most related to the protocols proposed in this paper, including an overview of our secure framework. In Section 3, we develop a secure equi-join procedure within the context of the framework. Then, in Section 4 presents a more efficient protocol using data k-anonymization techniques. Section 5 provides an empirical analysis of the protocol with a real world Census dataset. Section 6 discusses some lessons learned, and Section 7 concludes the paper.

2 Related Work

To date, several generic privacy preserving data integration frameworks [15,16] have been proposed. These frameworks generally involve three basic privacy-preserving components: 1) schema matching, 2) joins, and 3) query processing. For the most part, research has been performed to address specific challenges in each component. For example, in [17], a cryptographic protocol for schema matching has been proposed. In addition, several equality joins have been proposed in the past for different settings [18,19,20,21]. Usually, these protocols involve expensive cryptographic operations and they are not directly scalable for large data sets. To enable more efficient solutions, hash-based noise addition techniques [22] and anonymization based approaches have been [23] proposed. Compared to previous work; however, the protocols we introduce in this paper require participants to have only single interaction. Also, in our work, we enable each data holder's records to be incrementally added to the central data repository.

In prior work, we introduced a framework to support integration and querying of a database of encrypted genomic sequences and affiliated patient-specific data [7]. The goal of the framework is to enable 1) the secure transfer and centralized storage of person-specific DNA sequences in a database and 2) the support queries and data mining tasks as they would be performed on the original sequences. To achieve these goals, the framework incorporates four types of participants: data holders, data users, a data site, and a key holder site. As a running example, imagine that the set of data holders are hospitals and that the set of data users are biomedical researchers. We assume each hospital maintains some demographic information (e.g., sex, age), clinical information (e.g., medical diagnosis), and DNA records. We further assume that the participants do not collude and are semi-honest [24], such that all participants can use information they observe to infer knowledge, but they do not deviate from the framework's specification. The data site (DS) and key holder site (KHS) are crucial to the security components of the architecture. Specifically, KHS manages the keys that encrypt patient information and queries and the keys to decrypt the query results. In contrast, the encrypted DNA and patient data is stored and processed at DS. We summarize the participants' roles and the framework in Appendix A.

3 Secure Queries and the Equi-Join

We focus on how to execute equi-join queries on the encrypted data stored at DS. Such queries are necessary for adding and integrating new datasets to the database already stored at DS. Here, we present a novel protocol to perform secure equality joins, termed as *Secure-Equijoin*, and applicable to non-interactive environments with independently encrypted attributes. The proposed protocol uses the Paillier cryptosystem (key properties are presented in Appendix B).

We adopt the following notation for this paper: $\theta^h = \{\theta^h_1, \ldots, \theta^h_\alpha\}$ is dataset of α records in relational form, where each row (e.g., θ^h_i) indicates an individual's data. θ^h_{ij} represents the value of the j^{th} attribute of the i^{th} individual in θ^h. E_{pk} and D_{pr} respectively are the Paillier's encryption and decryption functions with public key pk and private key pr. $\theta^{h_1} \bowtie \theta^{h_2}$ indicates the join of datasets θ^{h_1} and θ^{h_2} on common attributes (e.g., encrypted primary key).

Protocols 1 and 2 depict the pseudo-code of *Secure-Equijoin* as executed by DS and KHS, respectively. We assume a patient's record in a database is associated with identifying attributes, such as Social Security Number (SSN) or various demographics. The *Secure-Equijoin* is initiated by a hospital to encrypt the tuples in its database, θ^{h_1}, which is then sent to DS. After receiving encrypted tuples from two hospitals, $E_{pk}(\theta^{h_1})$ and $E_{pk}(\theta^{h_2})$, DS calculates $\theta^{h_1} \bowtie \theta^{h_2}$.

To evaluate the equi-join, DS securely calculates if two encrypted records are equivalent. Without loss of generality, we assume the join is performed using attributes $j_1 \ldots j_m$. Let θ^h_{ij} be the value of the j^{th} attribute of i^{th} tuple of θ^h. DS must inspect whether two encrypted tuples $E_{pk}(\theta^{h_1}_{i_1})$ and $E_{pk}(\theta^{h_2}_{i_2})$ match. Using the homomorphic properties of Paillier encryption, DS checks if $\left(\theta^{h_1}_{i_1 j_1} = \theta^{h_2}_{i_2 j_2}\right) \wedge \cdots \wedge \left(\theta^{h_1}_{i_1 j_m} = \theta^{h_2}_{i_2 j_m}\right)$ is true by checking if $M_{i_1,i_2} = \sum_{v=1}^{m} \left(\theta^{h_1}_{i_1 j_v} - \theta^{h_2}_{i_2 j_v}\right) \cdot r_v = 0 \bmod n$, where $r_1, \ldots r_m$ are non-zero random values. DS calculates $E_{pk}(M_{i_1,i_2})$ on encrypted data via evaluating:

$$E_{pk}(M_{i_1,i_2}) = (+_h)_{v=1}^{m} \left[\left(E_{pk}(\theta^{h_1}_{i_1 j_v}) +_h (E_{pk}(\theta^{h_2}_{i_2 j_v}) \times_h (-1)) \right) \times_h r_v \right]$$

When the decrypted value of $E_{pk}(M_{i_1,i_2})$ is 0, then two records correspond to the same patient with high probability. The main reason behind this observation is the fact that if all the attributes match then for each v, $(\theta^{h_1}_{i_1 j_v} - \theta^{h_2}_{i_2 j_v}) = 0$, and then M_{i_1,i_2} is 0. As proven below, if any of the attributes fail to match then it is highly unlikely that M_{i_1,i_2} is 0.

3.1 Correctness of Equi-Join Protocol

Here, we prove that $M_{i_1,i_2} = 0$ gives the correct join result with high probability. We first derive a lemma that states the probability of computing a 0 through homomorphic addition, when there is at least one non-zero value, is very low.

Lemma 1. *Given fixed $a_1, \ldots, a_m \in \{0, \ldots, n-1\}$ with at least one non-zero a_j value and uniformly randomly chosen $r_1, \ldots, r_m \in \{1, \ldots, n-1\}$. Let $S_m = \sum_{i=1}^{m} a_i \cdot r_i \bmod n$, then $Pr[S_m = 0] \leq \frac{1}{n-1}$.*

Algorithm 1. DS-Equi-Join

```
Require: Encrypted datasets $E_{pk}(\theta^{h_1})$ and $E_{pk}(\theta^{h_2})$; $j_1, \ldots, j_m$ are join attributes
for all $E_{pk}(\theta^{h_1}_{i_1}) \in E_{pk}(\theta^{h_1})$ do
    for all $E_{pk}(\theta^{h_2}_{i_2}) \in E_{pk}(\theta^{h_2})$ do
        for $v = 1$ to $m$ do
            $E_v \leftarrow \left(E_{pk}(\theta^{h_1}_{i_1 j_v}) +_h (E_{pk}(\theta^{h_2}_{i_2 j_v}) \times_h (-1))\right) \times_h r_v$
        end for
        $E_{pk}(M_{i_1,i_2}) \leftarrow E_1 +_h E_2 +_h \ldots +_h E_m$
    end for
end for
Send all permuted $E_{pk}(M_{i_1,i_2})$ values to KHS
```

Algorithm 2. KHS-Equi-Join

```
Require: $E_{pk}(M_{i_1,i_2})$'s from DS
for all $E_{pk}(M_{i_1,i_2}$ do
    if $D_{pr}(M_{i_1,i_2}) = 0$ then
        $(i_1, i_2)$ matches
    end if
end for
Send all matching $(i_1, i_2)$ pairs to DS
```

Proof. Let us assume a_j is not equal to zero (it exists due to the initial assumption) and all operations are modulo a large prime n.[1] Given any $x \in \{0, \ldots, n-1\}$, we can easily state the following inequality:

$$Pr[a_j \cdot r_j = -x] = Pr[r_j = -x \cdot (a_j)^{-1}] = \begin{cases} 0, & x = 0 \\ \frac{1}{n-1}, & \text{else} \end{cases} \leq \frac{1}{n-1}$$

Using the above inequality, we have: $Pr[S_m = 0] = \sum_{x=0}^{n-1}(Pr[a_j \cdot r_j = -x | S_m - a_j \cdot r_j = x] \cdot Pr[S_m - a_j \cdot r_j = x])$. Thus, for any x, $Pr[a_j \cdot r_j = -x] \leq \frac{1}{n-1}$,

$$Pr[S_m = 0] \leq \frac{1}{n-1} \left(\sum_{x=0}^{n-1} Pr[S_m - a_j \cdot r_j = x] \right) \tag{1}$$

Since all operations are modulo n, $S_m - a_j \cdot r_j$ will only takes values between $\{0, \ldots, n-1\}$, this implies that:

$$\left(\sum_{x=0}^{n-1} Pr[S_m - a_j \cdot r_j = x] \right) = 1 \tag{2}$$

[1] To uphold protocol security, we recommend choosing values of n, the modular base, on the order of 1024 bits. If, for instance, we join two tables with 10 million tuples each, the expected number of mismatches is significantly smaller than one (i.e., $\frac{10^7 . 10^7}{2^{1024}-1}$). Thus, for any database with the less than 2^{512} tuples, the error introduced by our scheme can be made arbitrarily small by increasing size of n.

Equations (1) and (2) concludes our proof. □

Lemma 1 provides intuition regarding the general properties of homomorphic addition. In the context of our protocol, this lemma can be used to prove Theorem 1. Basically, assume that we use homomorphic encryption to subtract the two patients' values in the same attribute (e.g., date of birth). Then, if the values match, the homomorphic subtraction will be a random value, or a false non-match with very low probability.

Theorem 1. *Given two encrypted tuples $E_{pk}(\theta_{i_1}^{h_1})$ and $E_{pk}(\theta_{i_2}^{h_2})$, if $\theta_{i_1}^{h_1}$ and $\theta_{i_2}^{h_2}$ matches, then $M_{i_1,i_2} = 0$ (M_{i_1,i_2} is defined as above); on the other hand, if $M_{i_1,i_2} = 0$ then $\theta_{i_1}^{h_1}$ and $\theta_{i_2}^{h_2}$ matches with probability at least $1 - \frac{1}{n-1}$.*

Proof. Due to the definition of M_{i_1,i_2}, if $\theta_{i_1}^{h_1}$ and $\theta_{i_2}^{h_2}$ matches then for all v, $\left(\theta_{i_1 j_v}^{h_1} - \theta_{i_2 j_v}^{h_2}\right) = 0$. This implies that $M_{i_1,i_2} = 0$. Let us consider the case where $M_{i_1,i_2} = 0$ but $\theta_{i_1}^{h_1}$ and $\theta_{i_2}^{h_2}$ does not match. This implies for some non-zero $a_v = \left(\theta_{i_1 j_v}^{h_1} - \theta_{i_2 j_v}^{h_2}\right)$ values, $\sum_{v=1}^{m} (a_v \cdot r_v) = 0 \bmod n$ for non-zero uniformly randomly chosen $r_v \in \{1, \ldots, n-1\}$. According to Lemma 1, the probability of such an event is less than $\frac{1}{n-1}$. This implies that if $M_{i_1,i_2} = 0$ then two tuples match with probability bigger than $1 - \frac{1}{n-1}$. □

3.2 Security of the Equi-Join Protocol

The protocol is secure within the framework with respect to DS because it does not have access to the private keys. In addition, due to semi-honest model, we assume that DS follows the protocol and only asks KHS for the decryption for the properly constructed $E_{pk}(M_{i_1,i_2})$ values. Thus, we consider security with respect to KHS. Specifically, KHS observes only encrypted values of either 0, which corresponds to a match, or a random value, which corresponds to a non-match. Since the encryption scheme is semantically secure, KHS cannot learn anything regarding the corresponding values of the encrypted data.

3.3 Communication and Computational Cost

Assume *Secure-Equijoin* is performed using m attributes, and let $|\theta^{h_a}|$ indicate the number of tuples in θ^{h_a}. According to Protocol 1, for each tuple pair, we perform $2m - 1$ homomorphic additions, m modulo inverses and m homomorphic multiplications. Since each homomorphic multiplication is equivalent to an exponentiation, which is much more expensive than the other operations, we define the computational complexity in terms of the number of exponentiations. Each tuple pair requires m exponentiations and there are $|\theta^{h_1}| \cdot |\theta^{h_2}|$ such pairs; as a result, the number of exponentiations for the *Secure-Equijoin* protocol is bounded by $O(|\theta^{h_1}| \cdot |\theta^{h_2}| \cdot m)$.

For each tuple pair, DS sends the M_{i_1,i_2} value to KHS. Assuming an s-bit long n value, the communication complexity is bounded by $O(|\theta^{h_1}| \cdot |\theta^{h_2}| \cdot s)$.

4 *k*-Anonymity for Secure Equi-Join

The *Secure-Equijoin* protocol is impractical with large datasets because it requires testing each new and existing record as a potential join, such that the running time increases quadratically with the number of tuples. To overcome this limitation of the protocol, we propose a method that relaxes the semantically secure protections afforded by the homomorphic cryptosystem to anonymity sets of size k. We append non-encrypted patient-specific values (e.g., demographics) to encrypted data (e.g., DNA) in a manner that satisfies a formal privacy model. Specifically, hospitals disclose non-encrypted patient-specific data in a manner that satisfies k-anonymity (basic properties are presented in Appendix C).

4.1 *k*-Anonymity as Hash Key

In essence, k-anonymized values serve as hashed keys by which DS can partition encrypted tuples into buckets that are much smaller than the number of tuples in the database. In doing so, DS can perform the secure equi-join procedure on the homomorphically encrypted identifiers, such as SSNs, without testing every combination of tuples in the cross-product of the submitted databases. Moreover, by k-anonymizing the data, we ensure that every tuple in a bucket is linkable to no less than k patients. Thus, after joining encrypted values, DS is unable re-identify a tuple to less than k patients.

To implement this model, we assume the hospitals' databases contain a common set of quasi-identifying attributes, such as a patient's residential zip code and age. Each hospital encrypts all remaining attributes via the public key of DS. The hospitals then k-anonymize the quasi-identifying values of their datasets.

4.2 Joins with Equivalent Populations

First, we consider the case when all hospitals have data on the same population. In this scenario, each hospital k-anonymizes its dataset (with the same anonymization algorithm, k values and generalization schema) and submits the result to DS. When DS performs a join, it constructs buckets corresponding to each combination of k-anonymous values. For each bucket, DS executes the *Secure-Equijoin* protocol. At the completion of the protocol, every tuple from each location will be joined with data stored at DS.

Claim. The joined database resulting from Secure-Equijoin at DS is k-anonymous with respect to the attributes in the quasi-identifier.

Proof Sketch. The union of the joined quasi-identifying values is equivalent to the quasi-identifying values of any tuples involved in the join. Since there are k or more tuples in each bucket, after all tuples are joined with their corresponding partners, their quasi-identifying values do not change. Thus, the resulting database is k-anonymous. □

4.3 Joins with Overlapping Populations

Next, we address how to join data when hospitals collect records on overlapping populations of patients. In this case, hospitals cannot k-anonymize their databases independently, as was performed in the prior section. If this occurs, the same patient's data can be k-anonymized in different ways at different hospitals. As a consequence, data that is joined at DS could violate the k-anonymity model. For instance, consider the record $\{25, 47906\}$ defined over the attributes *AGE* and *ZIP CODE* and the generalization hierarchies in Figure 3 (See Appendix C). This tuple could be k-anonymized to $\{[23, 45], 479{*}{*}\}$ at one hospital and $\{25, 47{*}{*}{*}\}$ at another hospital. In each submitted database, the tuples are k-anonymous, however, after joining the encrypted values, DS can infer that the corresponding demographics must be $\{25, 479 * *\}$, which is more specific than both of the submitted tuples. If the number of tuples with the combination of these demographic values is less than k, then the join violates k-anonymity.

Algorithm 3. k-Equijoin

Require: k: anonymity threshold; $V_1, \ldots, V_m$: a set of value generalization hierarchies

```
1: DS: Send $T[Q]$ to hospital $h$
2: Hospital $h$:
   (1) Compute $C \leftarrow$ Get-Candidate($T^h, T[Q], V_1, \ldots, V_s$)
   (2) Anonymize $C$ based on $T[Q]$ and send $C$ to DS
3: DS: Compute $C' \leftarrow$ Equi-Join($C, T$) and send $C'$ to hospital $h$
4: Hospital $h$:
   (1) Compute $\mathit{\Gamma} \leftarrow (T^h - C) \cup C'$
   (2) $k$-anonymize $\mathit{\Gamma}$ and send it to DS
```

To prevent this inference leak, we present a protocol that enables hospitals to coordinate and, subsequently, ensure all data stored at DS satisfies the k-anonymity framework. We call this protocol *k-Equijoin* and its key steps are presented in Protocol 3. Before delving into the details of the protocol, we provide an informal overview. Let T represent the database stored at DS. We partition T into $T[Q]$ and $T[\hat{Q}]$. The first component, $T[Q]$, represents the projection of T on the quasi-identifier attributes. The second component, $T[\hat{Q}]$, represents the encrypted portion of T. Similarly, data at hospital h is represented as $T^h[Q]$ and $T^h[\hat{Q}]$. To initiate the protocol, h submits a request to DS to transfer its patient-specific records. At this point, we must consider two scenarios: 1) a base case in which h is the first hospital to submit data and 2) a general case in which h is not the first submitter. In the base case, DS has yet to receive data from any hospital, so h will k-anonymize the quasi-identifying attributes in its database and encrypt the remaining attributes. Then, h will send T^h to DS for storage. In the more general case, DS has already received and stored data from one or more hospitals. So, hospital h partitions his data into records that DS: 1) may have already received from other hospitals and 2) definitely has not received.

Hospital h will k-anonymize the first set of records in the same schema as they were submitted to DS by other hospitals. The second set of records, which we call Γ, can be k-anonymized by h without regard to records at DS because they are the first to be to submitted. Thus, h generates and sends $\Gamma[Q]$ to DS.

Now, we present the protocol more formally. To produce consistent data, the degree of k-anonymization and the generalization algorithms are fixed for the execution of the protocol. In addition, we assume there exists a fixed set of value generalization hierarchies available to the hospitals. Without loss of generality, we assume that some data is already stored at DS. *k-Equijoin* utilizes a function called *Get-Candidate* to produce a set of data tuples in T^h whose projection on quasi-identifier attributes can be anonymized to some tuples in $T[Q]$. For example, let V_1 and V_2 be the value generalization hierarchies (VGHs) presented in Figure 3 (Appendix C), which correspond to the attributes *AGE* and *ZIP CODE*. Using set representation, let $t = \{[46, 90], 475{*}{*}\}$ be one of the records in $T[Q]$. Based on the two VGHs and t, we compute a set γ such that any value in γ can be generalized to some value in t based on the given VGHs. Consequently, given t, $\gamma = \{50, 53, 70, 75, 80, 47500, 47535\}$. Suppose $t^h = \{50, 54339\}$ is one of the records in $T^h[Q]$. Since $t^h \nsubseteq \gamma$, t^h cannot be generalized to t. On the other hand, if $t^h = \{50, 47500\}$, then t^h can be generalized to t, and t^h is called a candidate for the future join process at DS. The set C contains all possible candidates computed according to $T[Q]$ and VGHs.

In Step 1 of *k-Equijoin*, DS sends the k-anonymous portion of the centralized database to hospital h.[2] Next, in Step 2, h computes the set of its tuples that could potentially join to tuples in the centralized database at DS. Then, h k-anonymizes his local database $T^h[Q]$ and sends the k-anonymized portion along with the corresponding encrypted portion to DS. Then, in Step 3, after DS receives C, DS will locate the tuples in C that can potentially join to its data using the *Secure-Equijoin* protocol. After this computation, DS notifies hospital h which tuples can definitely not be joined with existing records, denoted as the set C'. Finally, in Step 4, h k-anonymizes the remaining tuples with those in C', and sends the k-anonymized tuples to DS.[3]

Claim. The database at DS after all hospitals execute k-Equijoin is k-anonymous with respect to the attributes in the quasi-identifier.

Proof Sketch. Tuples that can be joined successfully at DS do not violate k-anonymity. Since hospital h locally k-anonymizes the other data tuples (denoted as Γ in Algorithm 3), these tuples create new buckets at DS. Each of the new buckets contains at least k tuples and each tuple in the new buckets do not violate k-anonymity. Similarly, the extension of this protocol to all hospitals

[2] Note, before sending $T[Q]$, DS can eliminate all duplicates in $T[Q]$ to reduce communication costs at least by k times.

[3] Note that a small number of tuples may not be k-anonymized. When this occurs, h will not send these tuples to DS. However, these data can be combined with future collected data. If the size of combined data is greater than k, h can initiate k-Equi-Join protocol again with DS.

preserves the k-anonymity criteria. When multiple locations submit their data to DS, the *k-Equijoin* protocol is executed sequentially; i.e., hospital 1 executes the protocol, followed by hospital 2, and so on. Each execution of the *k-Equijoin* protocol with a new hospital preserves the k-anonymity of the database at DS. Since each execution of the protocol preserves the k-anonymity criteria, the final database at DS must be k-anonymous as well. This holds for all hospitals and the database at DS that results from the sequential execution of *k-Equijoin* must satisfy k-anonymity. □

We hypothesize that *k-Equijoin* will reduce the number of cryptographic match evaluations that DS must perform in comparison to the *Secure-Equijoin* protocol because tuples corresponding to the same patient must reside in the same bucket. We experimentally investigate this hypothesis below.

5 Experiments

To evaluate the proposed protocols, we selected the U.S. Census income dataset, which is publicly available on the UC Irvine Machine Learning Repository [25]. This dataset contains person-specific records that were extracted from the 1994 and 1995 Current Population Surveys. It contains 286,775 tuples without missing values. There are 40 demographic and employment-related attributes.

5.1 Secure-Equijoin

We prototyped our protocols in Java and executed the secure join query experiments using relations of 100, 200, 300, and 400 tuples extracted from the Census income dataset. Please note that for relations of size 100, we need to do compare 10000 tuple pairs. Similarly for data set size 400, we need to compare 160000 tuple pairs. For simplicity, we executed join queries of the form $\theta^{h_1} \bowtie \theta^{h_1}$ with different numbers of attributes in the equi-join criteria. The experimental results are summarized in Figure 1 and indicate that join queries are computationally expensive and thus very time consuming. As expected, the running time of the join protocol increases linearly with the number of attributes in the join and quadratically with the size of the relation. For instance, it took around an hour to compute an integration of two datasets with 100 patients each (i.e., a join operation that involves 10000 tuple pair comparisons) across four attributes.

From a practical perspective, we investigated the degree to which specialized software implementations could decrease the time necessary to complete secure equi-joins. We simulated the homomorphic encryption-decryption function in the C programming language with the GMP library. Our results indicate that we can achieve an order of 10 speed-up. This implies that we can complete two million tuple pair comparisons in less than a day, which may be an acceptable amount of time for some biomedical research queries. Yet, as the size of the database grows, the savings afforded by specialized code is significantly outpaced by the increased time required to evaluate possible tuple pairs in the homomorphic space. Thus, scaling the basic join protocol to large datasets is not

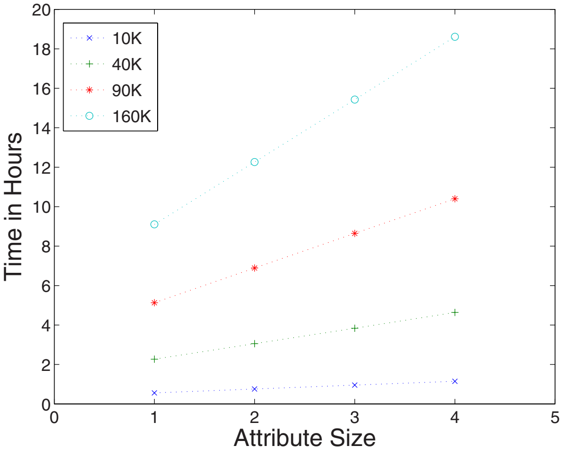


Fig. 1. Execution Time for Join Queries

feasible for the large databases that will be employed in biomedical data mining endeavors.

5.2 *k*-Equijoin

To evaluate the effect of k-anonymous demographics on secure joins, we applied the Datafly [26] algorithm, which generates k-anonymous datasets through generalization hierarchies using a greedy heuristic. We selected 4 of the 40 attributes in the dataset to represent the quasi-identifier (i.e., the generalizable attributes). Table 2 summarizes the statistics regarding the k-anonymous income dataset with k equal to 5, 10, 15, and 20. To orient the reader with respect to the results in this table, let θ^{h_1} refer to the income dataset and θ^{h_2} refer to a particular hospital's dataset. Table 2 can be interpreted as follows: given $k = 20$, $\forall \theta_{i_2}^{h_2} \in \theta^{h_2}$, the expected number of tuples in θ^{h_1} that potentially match θ^{h_2} is 196. In other words, when the quasi-identifying attributes of θ^{h_2} is 20-anonymous, the average number of exponentiations is bounded by $196 \cdot |\theta^{h_2}|$ instead of $|\theta^{h_1}| \cdot |\theta^{h_2}|$, which occurs when we do not apply k-anonymity. This result implies that for $|\theta^{h_1}| = 286,775$ and $|\theta^{h_2}| = 1000$, by applying k-anonymity, we can reduce the number of exponentiations needed from $286,775,000 = 286,775 \cdot 1000$ to $196,000 = 196 \cdot 1000$. We have increased efficiency by almost 1500! Thus, we have confirmed our hypothesis that when k is not large, the number of secure equality checks required by the equi-join protocol can be reduced drastically.

6 Discussion

The limiting factor in the applicability of our join protocols is the computational power needed for exponentiations and the bandwidth necessary for

Table 1. Census Dataset Description

Attribute	Values	VGH Height
Age	91	5
Marital Status	7	3
Race	5	2
Sex	2	2

Table 2. Anonymization Statistics

k	Min	Max	Avg	Med	Std
5	5	1964	122	19	313
10	10	1964	150	30	341
15	15	1964	174	41	363
20	20	1964	196	51	379

communication between the data site (DS) and key holder site (KHS). We believe that our protocols will be more efficient when implemented in secure computer hardware. Here, we suggest several potential hardware-based improvements.

First, significant efficiency gains for our protocols can be achieved through cryptography accelerators that are tailored to execute expensive exponentiation operations. Based on reported results with hardware accelerators , the combination of more efficient software implementations (e.g., in the GMP library of the C language) with hardware accelerators could substantially decrease the time needed to complete an exponentiation in comparison to our Java-based experiments. This implies that secure joins of relations, without the use of k-anonymous keys, on databases of 10,000 could be achieved in less than a day. We leave the implementation of our algorithms using crypto accelerators as a future work.

Second, we can decrease the communication cost by co-locating KHS and DS. Specifically, we envision a system in which the functions of the KHS are performed by a secure co-processor that resides on the same server as DS. A secure co-processor is a single-board computer consisting of a CPU, memory and special-purpose cryptographic hardware contained in a tamper-resistant shell; certified to level 4 under FIPS PUB 140-1 (One example of such a secure co-processor is the IBM 4758 Cryptographic co-processor [27]). When installed on the server, it is capable of performing local computations that are completely hidden from the server. If tamper is detected, the secure co-processor clears the internal memory. The implementation of KS functionality through a secure co-processor on the same machine as DS will decrease the communication cost.

7 Conclusions

In this paper, we presented a framework by which person-specific biomedical data can be stored and queried in a centralized encrypted repository. We demonstrated that the administrator of the repository can perform joins of encrypted databases without decrypting or inferring the contents of the joined records. Furthermore, we presented an efficient extension to the join protocol that reveals patient-specific demographics in a manner that satisfies a formal privacy model, i.e., k-anonymity. In doing so, we allow the administrator to perform efficient joins with the guarantee that each record can be linked to no less than k patients in the population. This research is notable in that it demonstrates how centralized biomedical data repositories can be integrated and updated with data distributed healthcare organizations without violating privacy regulations. In future research, we intend to implement this research in real world settings and extend it to secure computer architectures, such as secure coprocessors.

References

1. National Institutes of Health: Final NIH statement on sharing research data. NOT-OD-03-032 (2003)
2. National Institutes of Health: Genome-wide studies in biorepositories with electronic medical record data. RFA-HG-07-05 (2007)
3. National Institutes of Health: Policy for sharing of data obtained in nih supported or conducted genome-wide association studies. NOT-OD-07-88 (2007)
4. Benkner, S., Berti, G., Engelbrecht, G., Fingberg, J., Kohring, G., Middleton, S., Schmidt, R.: Gemss: grid-infrastructure for medical service provision. Methods of Information in Medicine 44, 177–181 (2005)
5. Anonymous: Medicine's new central bankers. The Economist (2005)
6. Barbour, V.: UK Biobank: a project in search of a protocol? Lancet 361, 1734–1738 (2003)
7. Kantarcioglu, M., Jiang, W., Liu, Y., Malin, B.: A cryptographic approach to securely share and query genomic sequences. IEEE Transactions on Information Technology in Biomedicine (in press, 2008)
8. Malin, B., Sweeney, L.: How (not) to protect genomic data privacy in a distributed network: using trail re-identification to evaluate and design anonymity protection systems. Journal of Biomedical Informatics 37, 179–192
9. Helliker, K.: A new medical worry: identity thieves find ways to target hospital patients. Wall Street Journal (2005)
10. Quantin, C., Allaert, F., Avillach, P., Fassa, M., Riandey, B., Trouessin, G., Cohen, O.: Building application-related patient identifiers: what solution for a european country? Int. J. Telemed Appl., 678302 (2008)
11. Grannis, S., Overhage, J., McDonald, C.: Analysis of identifier performance using a deterministic linkage algorithm. In: Proceedings of the 2002 American Medical Informatics Annual Fall Symposium, pp. 305–309 (2002)
12. Berman, J.: Zero-check: a zero-knowledge protocol for reconciling patient identities across institutions. Archives of Pathology and Laboratory Medicine 128, 344–346 (2004)
13. Sweeney, L.: *k*-Anonymity: a model for protecting privacy. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems 10, 557–570 (2002)
14. Samarati, P.: Protecting respondents' identities in microdata release. IEEE Transactions on Knowledge and Data Engineering 13, 1010–1027 (2001)
15. Clifton, C., Kantarcioglu, M., Foan, A., Schadow, G., Vaidya, J., Elmagarmid, A.: Privacy-preserving data integration and sharing. In: Proc. of the 9th ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery (2004)
16. Bhowmick, S., Gruenwald, L., Iwaihara, M., Chatvichienchai, S.: Private-iye: A framework for privacy preserving data integration. In: Proceedings of the 22nd International Conference on Data Engineering Workshops (ICDEW 2006). IEEE Computer Society, Los Alamitos (2006)

17. Scannapieco, M., Figotin, I., Bertino, E., Elmagarmid, A.: Privacy preserving schema and data matching. In: Proceedings of the 2007 ACM SIGMOD International Conference on Management of Data (2007)
18. Agrawal, R., Asonov, D., Kantarcioglu, M., Li, Y.: Sovereign joins. In: ICDE 2006: Proceedings of the 22nd International Conference on Data Engineering (ICDE 2006). IEEE Computer Society, Washington (2006)
19. Kissner, L., Song, D.: Privacy preserving set operations. In: Shoup, V. (ed.) CRYPTO 2005. LNCS, vol. 3621, pp. 241–257. Springer, Heidelberg (2005)
20. Freedman, M.J., Nissim, K., Pinkas, B.: Efficient private matching and set intersection. In: Eurocrypt 2004, Interlaken, Switzerland, International Association for Cryptologic Research (IACR) (2004)
21. Emekci, F., Agrawal, D., El Abbadi, A., Gulbeden, A.: Privacy preserving query processing using third parties. In: Proceedings of ICDE 2006, Atlanta, GA (2006)
22. Pon, R., Critchlow, T.: Performance-oriented privacy-preserving data integration. In: Data Integration in the Life Sciences, pp. 240–256. Springer, Heidelberg (2005)
23. Inan, A., Kantarcioglu, M., Bertino, E., Scannapieco, M.: A hybrid approach to private record linkage. In: Proceedings of the 24th Int'l Conf. on Data Engineering - ICDE 2008 (2008)
24. Goldreich, O.: General Cryptographic Protocols. In: The Foundations of Cryptography, vol. 2. Cambridge University Press, Cambridge (2004)
25. Blake, C., Merz, C.: UCI repository of machine learning databases (1998)
26. Sweeney, L.: Guaranteeing anonymity when sharing medical data, the datafly system. In: Proceedings of the 1997 American Medical Informatics Association Annual Fall Symposium, pp. 51–55 (1997)
27. IBM: IBM PCI cryptographic coprocessor (2004), http://www.ibm.com/security/cryptocards/html/pcicc.shtml
28. Paillier, P.: Public key cryptosystems based on composite degree residuosity classes. In: Stern, J. (ed.) EUROCRYPT 1999. LNCS, vol. 1592, pp. 223–238. Springer, Heidelberg (1999)
29. Sweeney, L.: Achieving *k*-anonymity privacy protection using generalization and suppression. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems 10, 571–588 (2002)

A Secure Architecture

The join protocols proposed in this paper are based on a secure framework [7]. To orient the reader, we briefly walk through the framework and describe the cryptographic mechanisms. Figure 2 summarizes the system.

Step 1 (Key Generation). KHS generates a <public, private> key pair and provides DS with the public key.

Step 2 (Data Encryption). Hospitals encrypt their records using the public key and send the results to DS.

Step 3 (Query Issuance). After the data is encrypted and stored at DS, a researcher sends a query for the database to DS.

Step 4 (Query Processing). DS executes the requested query and sends the encrypted results to KHS.

Step 5 (Result Decryption). KHS decrypts the result using the private key and sends it to the biomedical researcher.

We demonstrated that the framework supports aggregation queries, which are crucial to biomedical data mining tasks. Specifically, we proved that, through the homomorphic properties of Pallier encryption, we can execute count queries without revealing anything other than the query result. We would like to stress that our proposed system is secure under semi-honest model [24]. In the semi-honest model, the participating parties (e.g., KHS and DS) are assumed to follow the prescribed protocol and those parties only try to infer private information by using what is revealed during the protocol execution. In our context, semi-honest model implies that DS only asks KHS to decrypt *encrypted query results* as prescribed by the protocol. The semi-honest model, widely used in many different data mining tasks, is realistic for our purposes because changing complex protocols buried into large software without being detected could be hard. In addition, due to legal concerns, owners of the DS and KHS may not be willing to accept the potential liability of "not following the prescribed protocols".

Data stored at DS is semantically secure, so DS can learn the actual values only with the corresponding private key. However, KHS only issues DS a public key. KHS keeps the private key secret and does not share it with DS. As a result, DS is unable to discover the original patient information. Therefore, the data stored at DS are inherently secure against DS, as well as any biomedical researcher that issues queries in the framework.

B Homomorphic Cryptography

To achieve a simple and flexible architecture, we utilize a semantically secure public-key encryption scheme. The public key encryption scheme adopted in our architecture is probabilistic and possesses a homomorphic property. The homomorphic property allows us to compute the encrypted sum of two plaintext values through the corresponding ciphertexts. Formally, let $E_{pk}(.)$ and $D_{pr}(.)$ represent the encryption function with public key pk and the decryption function with private key pr, respectively. A secure public key cryptosystem is probabilistic and homomorphic if the encryption function satisfies the following features:

Constant Efficient: Given a constant k and a ciphertext $E_{pk}(m)$ of m, we can efficiently compute a ciphertext of km, denoted as $E_{pk}(km) := k \times_h E_{pk}(m)$.

Probabilistic: Given a message m, $c_1 = E_{pk}(m)$ and $c_2 = E_{pk}(m)$, $D_{pr}(c_1) = D_{pr}(c_2)$ but $c_1 \neq c_2$ with high probability.

Additive Homomorphic: Given the encryptions $E_{pk}(m_1)$ and $E_{pk}(m_2)$ of m_1 and m_2, there exists an efficient algorithm to compute the public key encryption of $m_1 + m_2$, denoted as $E_{pk}(m_1 + m_2) := E_{pk}(m_1) +_h E_{pk}(m_2)$.

Our framework can be applied within any additively homomorphic cryptosystem. In this paper, we situate the framework within the Paillier cryptosystem [28] because it has relatively wide-scale adoption and standardization.

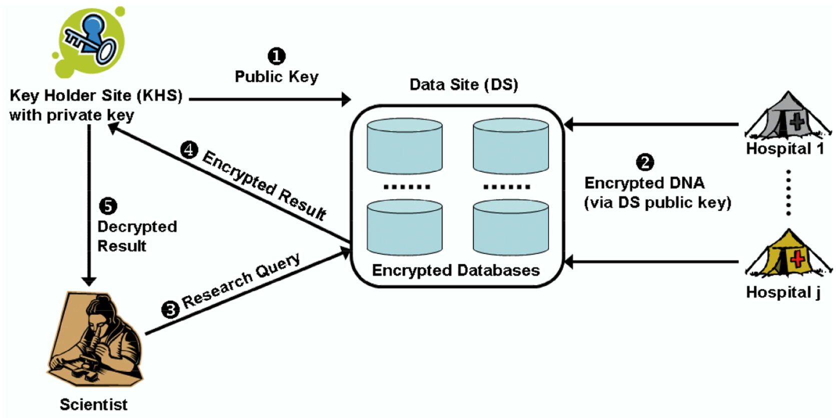


Fig. 2. General Architecture

C *k*-Anonymity

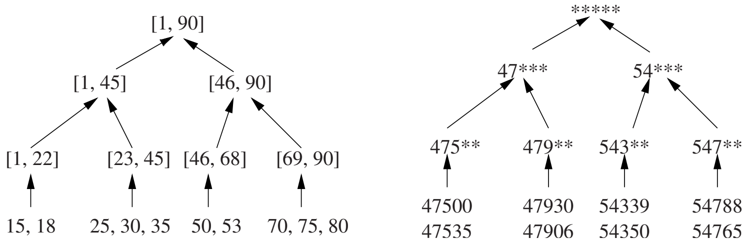
Here, we briefly review k-anonymity [13,29]. Let QI be a set of quasi-identifier attributes that can be used with certain external information to identify a specific individual, T be a dataset represented in a relational form and $T[QI]$ is the projection of T to the set of attributes contained in QI.

AGE	ZIP CODE
25	54339
75	47500
50	47535
30	54788

(a) Original Data

AGE	ZIP CODE
[23, 45]	54***
[46, 90]	475**
[46, 90]	475**
[23, 45]	54***

(b) 2-Anon. Data



(c) VGH of AGE

(d) VGH of ZIP CODE

Fig. 3. Data tables and value generalization hierarchies

Definition 1. *$T[QI]$ satisfies k-anonymity if and only if each record in it appears at least k times.*

The criteria for k-anonymity can be achieved via a number of mechanisms. In this paper, we concentrate on *generalization* [29]. In generalization, values are replaced by more general ones, according to a value generalization hierarchy (VGH). Figure 3 contains VGHs for the attributes *AGE* and *ZIP CODE*. According to the VGH of *AGE*, we say that 25 can be generalized to $[23, 45]$.

As an example, consider Figure 3(a). Here, we show a dataset T with quasi-identifier $QI = \{AGE, \textit{ZIP CODE}\}$. By generalization according to the VGHs, we can derive dataset in Figure 3(b) ($T[Q]$), which satisfies 2-anonymity.

Peer-to-Peer Private Information Retrieval

Josep Domingo-Ferrer and Maria Bras-Amorós

Universitat Rovira i Virgili
UNESCO Chair in Data Privacy
Department of Computer Engineering and Mathematics
Av. Països Catalans 26, E-43007 Tarragona, Catalonia
{josep.domingo,maria.bras}@urv.cat

Abstract. Private information retrieval (PIR) is normally modeled as a game between two players: a user and a database. The user wants to retrieve some item from the database without the latter learning which item. Most current PIR protocols are ill-suited to provide PIR from a search engine or large database: i) their computational complexity is linear in the size of the database; ii) they assume active cooperation by the database server in the PIR protocol. If the database cannot be assumed to cooperate, a peer-to-peer user community is a natural alternative to achieve some query anonymity: a user submits a query on behalf of another user in the community. A peer-to-peer PIR system is described in this paper which relies on an underlying combinatorial structure to reduce the required key material and increase availability.

Keywords: Privacy in statistical databases, private information retrieval, combinatorial designs.

1 Introduction

In private information retrieval (PIR), a user wants to retrieve an item from a database or search engine without the latter learning which item the user is interested in. PIR was invented in 1995 by Chor, Goldreich, Kushilevitz and Sudan [3,4] with the assumption that there are at least two copies of the same database, which do not communicate with each other. In the same paper, Chor *et al.* showed that single-database PIR (that is, with a single copy) is infeasible in the information-theoretic sense. However, two years later, Kushilevitz and Ostrovsky [12] presented a method for constructing single-database PIR based on the algebraic properties of the Goldwasser-Micali public-key encryption scheme [6]. Subsequent developments in PIR are surveyed in [15].

In the PIR literature the database is usually modeled as a vector. The user wishes to retrieve the value of the i-th component of the vector while keeping the index i hidden from the database. Thus, it is assumed that the user knows the physical address of the sought item, which might be too strong an assumption in many practical situations. Keyword PIR [2,12] is a more flexible form of PIR: the user can submit a query consisting of a keyword and no modification in the structure of the database is needed.

J. Domingo-Ferrer and Y. Saygın (Eds.): PSD 2008, LNCS 5262, pp. 315–323, 2008.

We claim that PIR protocols proposed so far have two fundamental shortcomings which hinder their practical deployment:

1. The database is assumed to contain n items and PIR protocols attempt to guarantee maximum privacy, that is, maximum server uncertainty on the index i of the record retrieved by the user. Thus, the computational complexity of such PIR protocols is $O(n)$, as proven in [3,4]. Intuitively, all records in the database must be "touched"; otherwise, the server could rule out some of the records when trying to discover i. For large databases, an $O(n)$ computational cost is unaffordable [1].
2. It is assumed that the database server cooperates in the PIR protocol. However, it is the user who is interested in her own privacy, whereas the motivation for the database server is dubious. Actually, PIR is likely to be unattractive to most companies running queryable databases, as it limits their profiling ability. This probably explains why no real instances of PIR-enabled databases can be mentioned.

If one wishes to run PIR against a search engine, there is another shortcoming beyond the lack of server cooperation: the database cannot be modeled as a vector in which the user can be assumed to know the physical location of the keyword sought. Even keyword PIR does not really fit, as it still assumes a mapping between individual keywords and physical addresses (in fact, each keyword is used as an alias of a physical address). A search engine allowing only searches of individual keywords stored in this way would be much more limited than real engines like Google or Yahoo.

In view of the above, in [11] a system is proposed in which a user masks her target query by ORing it with $k-1$ fake queries and then submits the masked query to a search engine or large database which does not need to cooperate (in fact, it does not even need to know that the user is trying to protect her privacy). Rather than total privacy, this system provides a sort of k-privacy, in that it cloaks the target query within k queries. This system works fine but assumes that the frequencies of keywords and phrases that can appear in a query are known and available: for maximum privacy, the frequencies of the target and the fake queries should be similar.

1.1 Contribution and Plan of This Paper

We present a peer-to-peer private information retrieval system. It has the same practical philosophy of [11]; however, rather than cloaking a query in a set of queries, the system described here cloaks a user in a peer-to-peer user community, because a user submits queries on behalf of her peers and conversely. This approach certainly requires the availability of peers (not needed in [11]) but it has the advantage of not requiring knowledge of the frequencies of all possible keywords and phrases that can be queried.

The new scheme uses a type of combinatorial design called configuration to increase service availability and reduce the number of required keys (see [16,13] for background on designs and configurations). The use of configurations in cryptographic key management is not new (*e.g.* see [13]), but their use in private information retrieval is.

Section 2 presents configurations. Section 3 describes the proposed peer-to-peer PIR protocol. Section 4 assesses the performance and the privacy offered by the protocol. Finally, Section 5 sketches conclusions and future work.

2 (v, b, r, k)-Configurations

We first define a combinatorial design and then a configuration as a special type of design.

Definition 1 (design). *A* design *is a pair $(X, \mathcal{A})$, where X is a set of points and $\mathcal{A}$ is a finite set of subsets of X, called blocks. The degree of a point $x \in X$ is the number of blocks containing x. The rank of $(X, \mathcal{A})$ is the size of the largest block.*

A design is said to be *regular* if all points have the same degree, say r. A design is said to be *uniform* if all blocks have the same size, say k (in which case the design is uniform of rank k). In the next definition we used the notations in [16,13].

Definition 2 ($(\mathbf{v}, \mathbf{b}, \mathbf{r}, \mathbf{k})$-1-design). *A (v, b, r, k)-1-design is a regular and uniform design with $|X| = v$, $|\mathcal{A}| = b$, degree r and rank k.*

A (v, b, r, k)-1-design corresponds to a bipartite semiregular graph with $v+b$ vertices and degrees r and k. A necessary and sufficient condition for the existence of a (v, b, r, k)-1-design is that

$$bk = vr. \tag{1}$$

Definition 3 ($(\mathbf{v}, \mathbf{b}, \mathbf{r}, \mathbf{k})$-configuration). *A (v, b, r, k)-configuration is a (v, b, r, k)-1-design where any two distinct blocks intersect in zero or one point.*

A (v, b, r, k)-configuration corresponds to a bipartite semiregular graph with $v+b$ vertices, degrees r and k, and girth strictly larger than 4. Configurations and their history have been largely studied by Harald Gropp in [7,8,9,10].

The following lemma quantifies the "connectivity" between blocks in a configuration.

Lemma 1. *In a (v, b, r, k)-configuration the number of blocks intersecting any specific block is $k(r-1)$.*

Proof. Consider a (v, b, r, k)-configuration $(X, \mathcal{A})$ and fix a block $A_i \in \mathcal{A}$. For any $x \in A_i$ define

$$\mathcal{B}_x = \{A_j \in \mathcal{A} : x \in A_j\} \setminus \{A_i\}$$

Clearly, $|\mathcal{B}_x| = r - 1$ for all $x \in A_i$. On the other hand, the sets $\mathcal{B}_x$ $(x \in A_i)$ are disjoint. Thus, the number of blocks intersecting A_i can be computed as

$$\left| \bigcup_{x \in A_i} \mathcal{B}_x \right| = \sum_{x \in A_i} |\mathcal{B}_x| = k(r-1). \qquad \square$$

A necessary condition for the existence of a (v, b, r, k)-configuration is $v \geq r(k-1)+1$ [5]. Yet, this condition may not be sufficient.

Finding configurations and even determining if a configuration with a given set of parameters exists is not trivial. One can use the next greedy algorithm to find a (v, b, r, k)-configuration if it exists.

We think of $\mathcal{A}$ as a list of b lists. The ith list in $\mathcal{A}$ correspond to the points in X contained in the ith block. We initialize all lists in $\mathcal{A}$ to empty lists. We use a pair of indices (i, j), where i and j respectively indicate which list in $\mathcal{A}$ and what position in the list we are dealing with. We start with $(i, j) = (1, 1)$ that is, we start by enumerating the first point of the first block.

Then we proceed by appending to the ith list in $\mathcal{A}$ a $(j+1)$th point, while the ith list has less than k elements or by appending a first point to the $(i+1)$th list. When this is not possible because there is no possible point to append satisfying the requirements of a configuration then backtracking is used (the last assignments of points to blocks are deleted in reverse chronological order until one is found that can be changed in a way compatible with the configuration structure).

It is helpful to use a second list of lists corresponding to the v points in X, with the lth list indicating what blocks contain the lth point. It can be used for verifying if a given point at a given position of a block in $\mathcal{A}$ is possible.

The algorithm is the following one:

```
while($0 < i \leq b$)
    $test = \begin{cases} 0 & \text{if } A_{i,j} = NULL \text{ and } j = 1 \\ A_{i,j-1}+1 & \text{if } A_{i,j} = NULL \text{ and } j \neq 1 \\ A_{i,j}+1 & \text{if } A_{i,j} \neq NULL \end{cases}$
    while($test \leq v - k + j$ and appending $test$ to $A_i$ does not lead to a configuration)
        $test = test + 1$
    if($test = v - k + j + 1$)
        $A_{i,j} = NULL$
        $(i, j) = \begin{cases} (i-1, k) & \text{if } j = 1 \\ (i, j-1) & \text{if } j \neq 1 \end{cases}$
        BACKTRACK
    else
        $A_{i,j} = test$
        $(i, j) = \begin{cases} (i, j+1) & \text{if } j \neq k \\ (i+1, 1) & \text{if } j = k \end{cases}$
if($i = 0$)
    output $\emptyset$
if($i = b + 1$)
    output $A$.
```

We refer the reader to [10] for results on existence and particular examples of configurations.

3 A Peer-to-Peer PIR Protocol

Consider a peer-to-peer (P2P) community consisting of b users. Assume a dealer who creates a key pool in the following way:

1. The dealer creates v keys and distributes them into b blocks of size k each according to a (v, b, r, k)-configuration.
2. The dealer confidentially sends one block of k keys to each user (no two users get the same block). E.g., if each user has got a public-private key pair, confidentiality can be achieved by sending the block of keys encrypted under the user's public key. Let A_i be the block assigned to user u_i, for $i = 1$ to b.
3. The dealer erases the v keys from its storage. If a trusted device such as a smart card is used as a dealer, it can be assumed that keys are forgotten by the dealer after distribution.

A variant of the above initialization process is to allow the users to send to the dealer their preferences about which other users they would like to share keys with. The dealer could take this input into account to the extent possible when assigning blocks of keys to users.

At the end of the process, by Lemma 1 the block of keys of each user intersects $k(r-1)$ other users' blocks. Consider now a storage pool consisting of v memory sectors, each corresponding to one key in the key pool.

Note 1. In the above set-up process, users do not need to know each other's identity. When deciding whether identities are to remain pseudonymous or not, one should carefully ponder whether the increased mutual trust derived from mutual knowledge compensates the loss of privacy of users in front of the rest of users in the P2P community. We will henceforth assume user pseudonymity. See Section 4.2 below for further discussion on user privacy.

A protocol for peer-to-peer PIR among the b users is specified next. The keys distributed to the users are used to key a symmetric cipher (*e.g.* see [14]). Also, in what follows we assume that plaintext queries and answers to queries can be distinguished from garbage by a user decrypting them; some kind of redundancy (*e.g* a cyclic redundancy check) can be appended to the query or the query answer to facilitate this distinction.

Protocol 1 (P2P PIR query submission)

1. *In order to submit a query q_i to a database or search engine, user u_i randomly selects one of the k keys in her block. Let x_{ij} be the selected key and $U^i_j = \{u^i_{j1}, \cdots, u^i_{j(r-1)}\}$ be the set of $r-1$ users with whom u_i shares x_{ij} according to the configuration used for key distribution. (Note that the sets $U^i_1, \cdots, U^i_k$ are disjoint due to the configuration structure.)*
2. *u_i reads the memory sector m_{ij} corresponding to key x_{ij} and decrypts it under x_{ij}. Five cases can arise depending on the outcome of decryption:*

(a) *The outcome is garbage. In this case, u_i encrypts q_i using a symmetric cipher keyed by x_{ij} and records the encrypted query in sector m_{ij}.*
(b) *The outcome is a query q_j issued by some user in U^i_j, who expects u_i to submit it on her behalf. In this case, u_i submits q_j to the database/search engine and records in sector m_{ij} the answer obtained after encrypting it under key x_{ij}. Thereafter, u_i goes back to Step 1 to select a new key and obtain assistance in submitting q_i to the database/search engine from someone in the group of $r-1$ users sharing the new key with u_i.*
(c) *The outcome is the answer to a previous query q'_j issued by some user in U^i_j and previously submitted by u_i to the database/search engine on behalf of that user. Since this answer has not yet been read by the user in U^i_j (a user is assumed to erase the query answer when she reads it), u_i goes back to Step 1 to select a new key and obtain assistance in the submission of her own query q_i.*
(d) *The outcome is a query q'_i previously issued by u_i, who expects some user in U^i_j to submit it on u_i's behalf. Since there is a previous query pending to be serviced by some user in U^i_j, u_i goes back to Step 1 to select a new key and obtain assistance with the submission of her new query q_i.*
(e) *The outcome is the answer to a previous query q'_i issued by u_i and previously submitted by some user in U^i_j to the database/search engine on behalf of u_i. In this case, u_i reads the answer, then encrypts her new query under key x_{ij} and finally records the encrypted query in sector m_{ij}.*

It can be seen that Protocol 1 will iterate until u_i can have her query submitted to the database/search engine by some other user. If a user does not have queries to submit and never runs Protocol 1, she does not contact the database; if the number of users contacting the database is very small (*e.g.* only two) there are problems: i) the database may infer who is submitting what query, and ii) the delay until a query answer can be collected can be too long. To prevent this, we require that all users run Protocol 1 at regular time intervals, whether or not they wish to submit actual queries (they can submit fake queries if necessary).

After submitting a query q_i, user u_i follows the protocol below to collect the answer to q_i:

Protocol 2 (P2P PIR query answer collection)

1. *If x_{ij} was the key selected to submit q_i, u_i keeps reading and decrypting m_{ij} at regular time intervals until either the answer to q_i is found (and erased from m_{ij}) or a timeout occurs.*
2. *If there was a timeout, u_i calls Protocol 1 to select a new key and find some other user who can assist her with the submission of q_i.*

4 Performance and Privacy

We examine in this section the influence of the configuration parameters k and r on performance and privacy. The other two parameters do not need discussion:

b is the (fixed) number of users in the P2P community and v is the number of keys and depends on k, r and b according to Equation (1).

4.1 Performance

First we deal with performance in terms of required keys and required storage. The proposed set-up process based on a (v, b, r, k)-configuration is compared with the trivial case in which every user shares a different key with every other user (complete connection graph). It turns out that performance improvement is controlled by parameter r.

Lemma 2. *If $r > 2$ it holds that:*

- *the number of keys and memory sectors required using a (v, b, r, k)-configuration is less than the number of keys and memory sectors required in the case of a complete graph;*
- *the overall number of keys stored by the users with a (v, b, r, k)-configuration is less than in the case of a complete graph.*

Proof. With a complete graph among the b users, the number of required keys and memory sectors is $b(b-1)/2$. Each user stores $b-1$ keys, so that the overall number of keys stored by the users is $b(b-1)$.

With configurations, the number of required keys and memory sectors is $v = bk/r$ (Equation (1)). The overall number of keys stored by the users is bk. Now, from Lemma 1 it follows that $k(r-1) \leq b-1$ (the number of blocks intersecting a specific block cannot be greater than $b-1$); thus

$$\frac{bk}{r} \leq \frac{b(b-1)}{r(r-1).}$$

So for $r > 2$ there is a reduction in the number of required keys and memory sectors with respect to the complete graph case. Similarly, since $bk \leq b(b-1)/(r-1)$, for $r > 2$ there is a reduction in the overall number of keys stored by the users. □

In addition to storage, another performance metric is how long does it take for u_i to get her query submitted and answered. Clearly, the greater the number r with whom u_i shares the selected key x_{ij}, the shorter is the expected waiting time.

Therefore, performance improves as r increases.

4.2 Privacy

If a good symmetric cipher is used for encryption, the encrypted contents stored in any memory sector are indistinguishable from garbage (see [14] for a review of the properties of the output of a symmetric cipher). Thus to an intruder not in $\{u_i\} \cup U_j^i$ the content of sector m_{ij} is indistinguishable from garbage; therefore, such an intruder does not gain any information on the queries submitted nor the query answers received by users in $\{u_i\} \cup U_j^i$.

Within U_j^i, the $r-1$ users do not know in principle to which other user in $\{u_i\} \cup U_j^i$ do the queries and query answers correspond. From this remark and those in the performance section above, one might be tempted to take r as large as possible: a single key shared by all b users (that is, $r = b$ and $k = v = 1$). Yet this does not look like a good solution because then any user can see any query or query answer, which does not seem very secure nor private: even if users are pseudonymous, successive queries by the same u_i are likely to be linkable, with the subsequent profiling and re-identification risk for u_i (*e.g.* a way to link u_i's queries is through her IP address when u_i writes her queries or reads her query answers).

It seems better for u_i to limit (pseudonymous) visibility of her query and its answer to those parties strictly needed: the database/search engine and a set of users just large enough so that the expected waiting time to get the query answer is not too long. Indeed, if u_i can select x_{ij} among $k > 1$ different keys at Step 1 of Protocol 1, where each key is shared by a disjoint set of users (see proof of Lemma 1), users in U_j^i only see on average 1 out of k queries issued by u_i (and 1 out of k query answers received by u_i). Therefore, the risk that a user in U_j^i can profile and thereby re-identify u_i decreases as k increases.

Finally, let us examine the privacy of user u_i in front of the database or search engine. The queries issued by u_i are submitted by the $k(r-1)$ users with whom u_i shares keys. In fact, each u_j in that group of $k(r-1)$ users submits on average a fraction $1/(k(r-1))$ of the queries issued by u_i. But u_j may also submit other queries corresponding to other users different from u_i with whom u_j shares a key. Therefore the query profile of u_i is diffused among the $k(r-1)$ users with whom u_i shares a key and confused among the other queries submitted by those users.

In summary, the greater r, the better is performance; the greater k, the greater is privacy in front of the other users; the greater $k(r-1)$, the greater is privacy in front of the database/search engine.

5 Conclusion

So far in the literature, practical PIR protocols that can be run against an uncooperative search engine or database aim at cloaking the user query among a finite number of fake queries. We have introduced a new paradigm, in which the user herself rather than the query is cloaked; indeed, the user seeks assistance by a P2P community who submit queries on her behalf. Unlike the query cloaking approach, any complex query can be submitted with our proposal and no knowledge of the frequencies of keywords and phrases is required. The level of privacy achieved is proportional to connectivity $k(r-1)$ of the P2P community.

From the combinatorial point of view, we have also contributed a construction of configurations (the structure used for key and storage management).

Acknowledgments and Disclaimer

This work was partly supported by the Spanish Government through projects CONSOLIDER INGENIO 2010 CSD2007-00004 ”ARES” and TSI2007-65406-C03-01 ”E-AEGIS”, and by the Government of Catalonia under grant 2005 SGR 00446. The authors are with the UNESCO Chair in Data Privacy, but their views do not necessarily reflect the position of UNESCO nor commit that organization.

References

1. Beimel, A., Ishai, Y., Malkin, T.: Reducing the servers computation in private information retrieval: Pir with preprocessing. Journal of Cryptology 17, 125–151 (2004)
2. Chor, B., Gilboa, N., Naor, M.: Private information retrieval by keywords. Technical Report TR CS0917, Department of Computer Science, Technion (1997)
3. Chor, B., Goldreich, O., Kushilevitz, E., Sudan, M.: Private information retrieval. In: IEEE Symposium on Foundations of Computer Science (FOCS), pp. 41–50 (1995)
4. Chor, B., Goldreich, O., Kushilevitz, E., Sudan, M.: Private information retrieval. Journal of the ACM 45, 965–981 (1998)
5. Colbourn, C.J., Dinitz, J.H. (eds.): Handbook of Combinatorial Designs, 2nd edn. Chapman and Hall/CRC, London (2007)
6. Goldwasser, S., Micali, S.: Probabilistic encryption. Journal of Computer and Systems Science 28(1), 270–299 (1984)
7. Gropp, H.: On the history of configurations. In: International Symposium on Structures in Mathematical Theories, Bilbo, pp. 263–268. Euskal Herriko Unibertsitatea (1990)
8. Gropp, H.: Configurations between geometry and combinatorics. Discrete Appl. Math. 138(1-2), 79–88 (2000); Optimal discrete structures and algorithms (ODSA 2000)
9. Gropp, H.: Existence and enumeration of configurations. Bayreuth. Math. Schr. 74, 123–129 (2005)
10. Gropp, H.: Configurations. In: Colbourn, C.J., Dinitz, J.H. (eds.) Handbook of Combinatorial Designs. Chapman & Hall/CRC, Boca Raton (2007)
11. Solanas, A., Domingo-Ferrer, J., Bujalance, S.: k-private information retrieval from privacy-uncooperative queryable databases (submitted, 2008)
12. Kushilevitz, E., Ostrovsky, R.: Replication is not needed: single database, computationally-private information retrieval. In: Proc. of the 38th Annual IEEE Symposium on Foundations of Computer Science, pp. 364–373 (1997)
13. Lee, J., Stinson, D.R.: A combinatorial approach to key predistribution for distributed sensor networks. In: Wireless Communications and Networking Conference-WCNC 2005, vol. 2, pp. 1200–1205 (2005)
14. Menezes, A.J., van Oorschot, P.C., Vanstone, S.A. (eds.): Handbook of Applied Cryptography. CRC Press, Boca Raton (1997)
15. Ostrovsky, R., Skeith-III, W.E.: A survey of single-database pir: techniques and applications. In: Okamoto, T., Wang, X. (eds.) PKC 2007. LNCS, vol. 4450, pp. 393–411. Springer, Heidelberg (2007)
16. Stinson, D.R.: Combinatorial Designs: Constructions and Analysis. Springer, New York (2003)

Legal, Political and Methodological Issues in Confidentiality in the European Statistical System

Jean-Marc Museux[1], Martine Peeters[1], and Maria João Santos[2]

[1] Methodology and Research Unit, Eurostat, L-2920 Luxembourg
[2] International Statistical Cooperation Unit, Eurostat, L-2920 Luxembourg
Jean-Marc.Museux@ec.europa.eu, Martine.Peeters@ec.europa.eu, Maria-Joao.Santos@ec.europa.eu

Abstract. The paper discusses the challenges linked to the need of the research community to have access to microdata files for scientific purposes. These needs have to be adequately balanced with the legal requirement of preserving the confidentiality of respondents. The paper presents the policies and instruments available at the European Union to progress in the supply of data to the research community, while respecting the legal requirements. More specifically, the paper explains the current process dealing with research projects and the work of the European Statistical System Network (ESSnet) project for statistical disclosure control. Finally the paper describes future trends that are currently investigated in the European Union, and more specifically the development of remote access facilities, the enhancement of disclosure control tools and the convergence towards common policies in Member States.

Keywords: Statistical Disclosure Control (SDC), access researchers microdata, tabular and micro data protection.

1 Introduction

The objective of this paper is to provide an overview of the various issues related to confidentiality in a European-wide perspective. It aims to give technical experts an idea of the difficulties raised by the multinational and administrative perspective which might not be perceived at first sight. The different perceptions, the lack of well defined standards are sources of divergences that make standard confidentiality much more problematic at European level. This paper calls for a closer partnership between administrative and research community and for a strong scientific research input and responsibility in order to design best practices to feed legal reflection at European level. Statistical confidentiality is a critical issue for Eurostat because it is at the core of the delicate trust data providers have towards statistics compilers. It influences greatly the quality of EU statistics and consequently the relationship between Eurostat and the ESS.

Technical development in the information era confronts the ESS with new challenges – not only with regard to Data Access - and the statistical disclosure control (SDC) problems connected to it.

J. Domingo-Ferrer and Y. Saygın (Eds.): PSD 2008, LNCS 5262, pp. 324–334, 2008.

2 Confidentiality Legal Framework

2.1 General Framework

The right to privacy is a fundamental right. It includes the protection of the person in the context of personal data processing. That means for instance the right to receive certain information, the right to access the data, the right to have the data corrected, etc. Statistical confidentiality primarily aims at safeguarding privacy in the field of statistics and is a key to the necessary trust that has to be maintained between statistical bodies and respondents. Mutual confidence ensures accurate and reliable basic information and eventually high quality statistics.

At EU level, statistical confidentiality is addressed in the following legal acts:

a. Council Regulation (EURATOM, EEC) No 1588/90 of 11 June 1990 on the transmission of data subject to statistical confidentiality to the Statistical Office of the European Communities;
b. Council Regulation (EC) No 322/97 of 17 February 1997 on Community Statistics;
c. Commission Decision 97/281/EC of 21 April 1997 on the role of Eurostat as regards the production of Community statistics;
d. Commission Regulation (EC) No 831/2002 of 17 May 2002 and following amendments implementing Council Regulation (EC) No 322/97 on Community Statistics, concerning access to confidential data for scientific purposes.

The Statistical Code of Practice (CoP) was adopted by the Statistical Programme Committee in 2005. It includes some provisions about the use of a statistical data by researchers and the protection of confidentiality and provides a framework to develop a harmonised activity in this domain. The observance of the CoP is now monitored by means of peer reviews which focus inter-alia on confidentiality protection and availability of microdata for research purposes.

Statistical confidentiality is regulated at EU level only to the extent to which statistical activities carried out by Eurostat and the national statistical authorities for the production of Community statistics are concerned. Specific confidentiality regimes still coexist at national level and differences may appear with the EU statistical confidentiality regime. These differences are less on the substance (the general concepts are common to a very large extent) than on the perception of the issue (the national framework remains the frame of reference), which is equally important.

The existent statistical confidentiality regime is thus not unified in one regulation, which leads to difficulties of interpretation between Member States and the Commission, which creates difficulties in different sectors. Improving the existing framework should contribute to avoiding repeated discussions and even in some cases obstacles when dealing with confidentiality issues in the context of the negotiation of sectoral regulations.

The wide acceptation of an objective basis for declaring data confidential and measuring disclosure risk would definitively ease legal progress in the field of statistical confidentiality. Scientific researcher's authority is certainly required to put a cut off to the endless subjective discussion. Lawyers are waiting for a strong technical input in order to design harmonised legislation.

2.2 Access to Researchers

Microdata sets are becoming important because of increasing interest in accessing them by researchers. This interest has two related drivers. The first is an aspect of modern life—accountable government and transparency. This is reflected in an increasing interest and demand for evidence-based policy, policy analysis, as well as monitoring policies and their impact. This kind of activity requires timely, detailed information and frequently requires more detailed analyses than are presently published by statistical organisations. Sometimes these analyses are seen as being outside the remit of national statistical institutes (NSIs) or even as activities that could compromise the perceived independence of NSIs. Indeed, these analyses are performed often by academic institutions or independent research institutions.

The second driver here is the changing nature of research itself. Modern research cannot be satisfied with aggregate data. Microdata are needed for fine analysis and model building.

In summary the advantage of having microdata sets at Eurostat level is twofold:

- It brings enormous flexibility in the use of the data and in the production of tabulated results according to users' requests, namely additional non-standardized tabulated data can be done at Eurostat (beyond standard tabulations transmitted by Member States). It increases the coherence of the statistical output.
- It allows better use and more thorough analysis of data collected at ESS level, increasing the benefits of the data collection, making better use of public money and eventually lowering the burden on statistical respondents.

Access to microdata for scientific purposes in the European Union: Commission Regulation (EC) No 831/2002

In order to meet the needs of researchers in the EU, two instruments have been developed in the frame of the basic confidentiality legal acts (Council Regulations 1588/90 and 322/97). These two instruments are (1) the Committee on Statistical Confidentiality (CSC) that has the implementation powers in all confidentiality matters and (2) the Commission Regulation 831/2002 concerning access to confidential data for scientific purposes.

At the Committee on Statistical Confidentiality Meeting of 2005 an action plan on confidentiality was approved including, amongst others, specific measures to improve and streamline the implementation of Commission Regulation 831/2002. The implementation of the access to microdata was improved by three means. Firstly, by reducing the administrative delays for establishing the contract with the researchers by a fast track consultation with Member States on research projects of a certain type (typology criteria agreed with member States); secondly, by enlarging the number of datasets available for researchers and, thirdly, by opening the safe centre at Eurostat.

At present microdata for researchers can be provided as anonymised microdata sets for the European Community Household Panel (ECHP), the Labour Force Survey (LFS), European Union Statistics on Income and Living Conditions (EU-SILC), Structure of Earnings Survey (SES) and Community Innovation Survey (CIS). In addition, the Education and Training Statistics Working Group is now discussing anonymisation criteria to distribute Adult Education Survey (AES) microdata files which should start in 2008. Commission Regulation 831/2002 is currently being amended in order to include the Farm Structure Survey (FSS).

Besides access to anonymised microdata sets, researchers can have access to SES and CIS confidential microdata sets in the safe centre located at Eurostat.

The efforts done by the Confidentiality Committee to streamline the administrative procedures associated to providing microdata for researchers have lead in the last four years to a rapid increase of the number of contracts established with researchers (from 18 contracts in 2003 to 130 contracts in 2007).

There are two levels of access to microdata:

- Level one: Confidential data as obtained from the national authorities. They allow only indirect identification of the statistical units concerned. This access is done through the use of a safe centre at Eurostat.
- Level two: Sets of anonymised microdata extracted from the above data. They are individual statistical records which have been modified in order to minimise, in accordance with current best practice, the risk of identification of the statistical units to which they relate. This access is done via distribution of encrypted CD-ROM according to contracts established between Eurostat and the corresponding institutions.

The advantage of possibilities offered by this regulation is that researchers now have the possibility to have access to harmonized datasets spanning all Member States, therefore avoiding the lengthy process of making requests to each MS and benefiting from a output harmonised product. This gives researchers opportunities for pan-European Union research and analyses.

Table 1. Research contracts – Main topics - Year 2006

Microdata from ECHP, EU-SILC, LFS and CIS	
Studies of specific sub-populations	*Studies of specific phenomena*
• Low skilled/unskilled labour force • Early school leavers • Poor • Regions/Europe • Long-term unemployed • Married women • Female labour force • Divorced • Temporary Workers • Persons at end of working life • Youth • Elderly • Disabled • Immigrants • Older workforce • SME • Researchers	• Mobility • Income inequality and distribution • Transition employment <-> unemployment • Fiscal, subsidies and insurance policies • Intra-family transfers • Inequality in income and education • Job structure changes (self employment, formal/informal) • Wage changes • Educational choices/training/ life long learning strategies • School-work transition • Labour market participation and fertility • Childcare • Discrimination • Real estate investments • Public pension schemes • Growth of cities • Innovation

The table above presents a synthesis of the projects related to submitted requests for microdata access to Eurostat in 2006.

Article 3 of Commission Regulation 831/2002 foresees a fairly straightforward and simple request process for researchers from the following categories of organisations:

- universities and other higher education organisations established under Community law or by the law of a Member State;
- organisations or institutions for scientific research established under Community law or under the law of a Member State;
- national statistical institutes of the Member States;
- the European Central Bank and the national central banks of the Member States.

For the other bodies, article 3 of the regulation lays down the condition that they must first be approved by the CSC if they wish to make requests to access confidential data for scientific purposes. The prerequisite to achieve admissibility is that the institution has demonstrated that it fulfils a set of criteria. The CSC has approved these criteria at its meeting of 10 December 2004. Specific services of EU Institutions, which carry out statistical activities, may be considered eligible for access for scientific purpose to specific confidential microfiles provided that the equivalent guarantees are provided. Commission Decision 2004/452/CE lists the bodies that have been considered admissible. Universities based outside Europe can also be considered as admissible; the University of Cornell (USA) was the first to be included in this list. The efforts will continue to extend the list of other bodies than can be regarded as admissible.

2.3 Initiatives at ESS Level on Microdata Access

Cenex on statistical disclosure control

The Task Force on Centres of Excellence set up by the Statistical Programme Committee (SPC) proposed to launch during 2005 a pilot project on the concept of Centres of Excellence (Cenex). Briefly, the concept consisted of setting up a team of national statistical organisations that provided expertise on a specific domain, developing tools or knowledge that benefited not only the participating organisations but the rest of the ESS.

Statistical disclosure control (SDC) was considered one of the two subjects that integrated the pilot phase of Cenex. For the generation of comparable statistical information across countries it was essential to assure that similar methods and tools were used to protect confidentiality in the published information. As long as member states compiled their statistics using different SDC-methods, the availability of useful European microdata was very much hampered.

The Cenex on SDC was launched by end 2005 and lasted for one year. This pilot was a success as it allowed participating NSIs to work together on various SDC themes, share the expertise and created necessary synergies. As a result the detailed inventory was done on the situation in the ESS regarding SDC methods and tools, as well as legal and administrative environments. The project was also beneficial to non-participating NSIs, since it permitted them to share their practices, compare the results with other NSIs and profit from the core work of the Cenex partners. The main outputs of this Cenex were a Handbook, an upgrading of the ARGUS SDC software and several training and scientific actions.

2.4 Considerations at International Level on Microdata Access

The 2003 Conference of European Statisticians (CES) agreed that supporting research with microdata is an important activity of the NSIs. CES set up a Task Force to develop guidelines on managing confidentiality while facilitating microdata access for research. The Task Force, chaired by Dennis Trewin, had prepared guidelines which were endorsed by the CES plenary session in June 2006. They addressed the need to unify the approaches internationally and to agree on core principles for dissemination of microdata. They also suggested moving towards a risk management rather than a risk avoidance approach in the provision of microdata. The principles agreed are general enough to be applicable in countries with different level of development and are accompanied by examples of good practices.

Eurostat and the Office of National Statistics in the United Kingdom organised in October 2006 a workshop on microdata access having as objective to reinforce the essential principles of microdata access as set out in the CES Guidelines; to share experiences of policies and practices in Member States regarded to be avant-garde on research access to microdata; to develop a shared understanding of the appropriate levels of capacity and support to the NSIs' activity in this context; and to give rise to actions that can be implemented in NSIs.

A follow up Workshop is foreseen by Eurostat on 23-24 October 2008 to continue the efforts to foster greater consistency across countries, facilitate better research access to microdata and improve the administrative arrangements

3 Methodological Issues

In general the legislation at national and European levels is fairly harmonised with respect to what is considered as confidential data. However, when implementing this legislation, the criteria used differ considerably from country to country. These criteria have sometimes an important historical weight; sometimes do not have a solid scientific basis; and in many cases lead to conservative solutions because real risks are not well mastered.

This diversity of interpretations is a consequence of the fact that there is no harmonised approach of disclosure risk. To agree on disclosure risk, one should agree first on the sensitivity of the data (how "private" are the variables in the file) and on the possibility to match these data with external sources, that is, to the presence of key variables or identifying variables. Second, there is a need to find a harmonised way to measure the risk. Methodological work is needed to reconcile the different approaches or to express preference for one of them.

The need to have common core criteria which, while providing a satisfactory harmonisation level, allow for a degree of flexibility to adapt to the specific perception of the society in each country is obvious. This will also have the advantage of having a more solid internationally agreed basis that better justifies national choices made in the release of microdata.

Disclosure protection of EU aggregates

Most of the time, Eurostat compiles EU aggregates on the basis of national aggregates. These are accompanied with a confidentiality flag informing Eurostat that the information should be treated as confidential. In the best situation, Eurostat is also informed on the presence of dominance in these aggregates. However, meta information is not standardised and even sometimes there exists confusion between not publishable because of lack of reliability and confidential as meant in the legal framework.

To declare information as (primary) confidential, Member States use measures of risk of disclosure of individual information (dominance rules, threshold rules) which are not harmonised. The level of protection can vary between Member States depending on different perceptions of the level of disclosure risk and also simply of the perception of the damage of disclosure itself. Distinction is rarely made between variables themselves: some variables might be considered as non sensitive whereas other from the same record could be.

The lack of harmonisation of primary confidential rules causes major methodological problems at Eurostat level. Software packages for handling secondary confidentiality are not designed to deal with such a situation. Regarding this aspect progress has been achieved for Prodcom and Structural Business Statistics (SBS), where confidentiality rules were developed and agreed by Member States specifically for each data collection. Based on these rules and using the Eurostat framework contract for methodological support, a table protection solution based on controlled rounding has been developed to compute publishable rounded values of those EU aggregates which are not publishable directly according to the Prodcom confidentiality rules. An algorithm has been developed and needs to be tested by Eurostat. For SBS, first a study was done to study the feasibility of using the restricted tabular adjustment proposal to identify EU aggregates which would not be published directly but would be replaced by an approximation. The development of the respective algorithm is at the moment on going.

Disclosure protection of microdata

To some extend the same holds when Eurostat has to design, in collaboration with Member States, anonymisation of microdata to be released to researchers. Despite they share common objectives:

- the need to follow Regulation principles on the right for privacy,
- the need to maintain the trust the respondent have in the statistical system,
- the need to monitor the release so to avoid confidentiality breach,

the differences in the perception of the risk and the lack of a universal measure of risk render the possibility of a consensus very thin. Part of the problems lies in the absence of knowledge of real risk.

This situation would be improved if once again, European experts would agree on a harmonised framework, reflecting the state of the art in terms of disclosure control technique, to assess and to measure disclosure risks by practitioners.

4 Future Perspectives

With respect to medium term perspectives, some components are already identified:

(1) The new legal framework

In its 57th meeting of 29 and 30 November 2005, the SPC agreed on an approach for the revision of the basic statistical legal framework, based on a paper presented by

Eurostat. After consultation of the Member States this led to a new proposal that was supported by the SPC in its meeting of 20 September 2007 and is now in the phase of submission to the Council to launch the co-decision process.

Some of the new aspects of the new legal framework concern statistical confidentiality: firstly, the need to enhance the role of NSIs and Eurostat for organisational, coordination and representation purposes was noted. In this context the current SPC is proposed to be replaced by a new Committee, the European Statistical System Committee. This new Committee is entrusted also with the functions of the CSC, which thus will cease to exist. The new legal framework moves towards providing some changes in the confidentiality chapter:

- General issues and definitions

The relevant provisions of Council Regulation 1588/90 on the transmission of data subject to statistical confidentiality to Eurostat have been integrated in the new basic legal act. For the sake of transparency and comprehensiveness, the basic legal act covers all guiding principles for confidentiality, including exceptional cases for passive confidentiality and dissemination of data from public sources or with the agreement of the statistical data subject. A reference has been included to measures to ensure the physical and logical protection of confidential data. Such measures shall be adopted by comitology (regulatory procedure).

- Exchange of confidential data

The exchange of directly identifiable data between Member States and Eurostat and between Member States is allowed by deleting the words ‘which do not permit direct identification’ in Article 14 of Council Regulation 322/97. The revised basic legal act includes an enabling provision for the exchange of confidential data with the European Central Bank (ECB). Transmission of confidential data will be justified only in cases where it is explicitly laid down in a legislative act adopted by the Parliament and the Council. Moreover, a legal guarantee must be given that the data are safely protected and used only for statistical purposes.

- Access for scientific purposes

The basic legal act contains a provision of a general nature enabling access to confidential data for scientific purposes. The approval of Member States is required when the data concerned have been transmitted by them to Eurostat, but, in order to allow more flexibility, the word ‘explicit’ has been removed. All implementing measures concerning the modalities, rules and conditions for such access are to be decided by comitology, more specifically by the regulatory procedure with scrutiny.

(2) ESSnet on SDC follow-up

The work done by the Cenex on SDC showed very big differences in terms of confidentiality treatment across the ESS. It demonstrated as well that more emphasis should be made on the knowledge transfer and building up of the expertise in less experienced countries. The analysis of the results also showed that it is essential to improve and adapt the existing SDC tools to NSIs environment. Since these environments and needs vary from country to country a more customized approach would be needed in order to make the software operational in the ESS. Last but not least, Cenex on SDC emphasized the need for further research in various SDC fields. In order to exploit these first results, and enable their appropriation by all NSIs and improve further SDC methods and tools, a call for proposals for a follow up ESSnet on SDC was launched (ESSnet is the new name given by the SPC to Cenex projects). The project has started in January 2008 and will last 24 months.

The objectives of this project are manifold: dissemination and follow-up of the Cenex on SDC work, especially handbook finalisation; further investigation of NSI needs on the basis of an inventory and on-site visits where necessary; software SDC adaptations and improvements in order to better address NSIs' needs; facilitation and test of the implementation in the different NSI environments; research work in specific SDC fields of interest to NSIs; active involvement of all (not only participating partners) NSIs in ESSnet work via contributions to forum discussions; participation in training actions…

(3) Remote access and delocalization of safe centres

There is a broad agreement among countries that this is a very promising approach. In order to facilitate the European reflection on this subject Eurostat has organized an Expert Group on remote access. The first meeting took place on 15 June 2007 and the second on 23 November 2007.

The action has the following objectives:

- To facilitate integrated European wide access and use of microdata sets from official statistics for scientific purposes;
- To develop a European integrated approach to remote access systems to microdata and to determine a remote access standard.

After the first meeting a roadmap for the future was developed. The actions foreseen depend mostly on the available types of data (either confidential or anonymised data) and on the types of access (access in safe centres to the data stored on-site or remote access to the data stored in the secure server outside). They are combined in different projects with shared responsibility and different leadership of the actors involved. A mixture of instruments financing these actions can be used (FP7, ESSnet, NSIs' resources). The approach for remote access will be planned in a step-wise manner, broken down in five phases:

- In the *first stage* two remote access options will be assessed:
- Remote facilities: data can be seen on the computer screen, manual confidentiality check after submission of final output,
- Remote execution: data not seen on the computer screen, query submitted by researchers, first automatic confidentiality check of the query, final output checked automatically, manual checks on the random basis.

Both options will be deeply analysed and prioritised from the point of view of feasibility and the utility to the researchers.

- In the *second stage* two pilot projects (ECHP and data archive) should be launched and tested.
- In the *third stage* a remote access to the microdata from the network of safe centres located in Member States might be tested. In this case no control of the identity of researcher will be needed. For the testing of the approach the microdata listed in Commission Regulation 831/2002 might be used.

- In the *fourth stage* the access to the microdata produced at national level and not transmitted to Eurostat will be made possible.
- In the *fifth stage* of the project remote access from the workstation of the researcher will be developed and tested.

Eurostat plans to launch the first and second phase by a pilot project that will be the subject of an ESSnet to be launched in 2008. ECHP European datasets could be deposited in the accredited safe centres to be accessed by researchers visiting these centres. If this pilot is successful, other datasets could be used namely business microdata sets (SES and CIS).

In addition, the possibility to establish framework contracts with data archives to host anonymised datasets will be explored in 2008.

(4) Public use files

Public use files (PUF) are the most accessible; widely and freely used microdata products made available by statistical institutes, but their value for policy for much policy relevant research is limited. Nevertheless these files are useful for some research purposes as teaching aids and are a good advertisement for a statistical institute. Continued distribution of PUF is threatened by the increased re-identification risk associated with both technological advances in linking software and widespread availability of administrative records. During the last decade researchers have developed increasingly sophisticated methodologies for restricted data products. The development of a methodology for generating synthetic or virtual data is a relatively recent activity. A key objective of the method is to preserve faithful representations of the original data so that inferences from the synthetic data are as consistent as possible with the inferences that would be drawn from the original data. One attractive feature of the synthetic data approach is that it can be used to create multiple PUF from the same underlying data – targeted at different audiences. The methodology of synthetic files as a measure to replace PUF need to be further researched.

The work at sectoral level to establish the criteria for establishing PUF (such as the work of the EU-SILC Task Force on anonymisation and establishment of PUF) should continue to be promoted in the future via the establishment of sectoral task forces that will define PUF for each survey. This should be facilitated by the fact that, in the new statistical law, PUF are no longer considered confidential

5 Conclusions

(1) There is an important gap between the information contained in statistical data and what a statistical office actually releases. A way to fill this gap is to supply microdata files to researchers.

(2) There are also many risks that have to be mastered - these are related to the legal protection of identification of individuals; to the possibility of bad use of the data; and to the perception of individuals of abusive manipulations of their information. These risks should be well managed, moving from a perspective of risk avoidance to risk management.

(3) The objective is to fill the gap as long as the risks are satisfactorily managed. For this purpose several legal and technical measures can be explored. The legal

measures concern eligibility of researchers and research projects. The technical ones refer to the different methods of confidentiality protection.

(4) International reflections show that although there is a broad consensus in favour of this supply of microdata to researchers, there is a diversity of views on many of the more detailed issues. In particular, criteria for considering a file sufficiently safe for dissemination vary widely.

(5) Commission Regulation (EC) No 831/2002 is a useful legal frame for the supply of microdata to researchers. After a difficult initial start, its implementation was considerably improved.

(6) Several ongoing actions facilitate the harmonisation of methods, practices and tools and streamline the process of supply of data to researchers.

(7) Various lines are being further explored to improve the dissemination of microdata to researchers. First, the continuation of the development of harmonised criteria for anonymisation and for eligibility of researchers/research; second, the legal frame to prevent bad use; third, the application of the Code of Practice; fourth, exploring the possibility of remote access to microdata.

References

1. Eurostat, 19th CEIES seminar: Innovative solutions in providing access to microdata, Lisbon, 26 and 27 September 2002, Office for Official Publications of the European Communities, Luxembourg (2003)
2. UNECE/Eurostat: Monographs of official statistics: Work session on statistical data confidentiality, Luxembourg, 7 to 9 April 2003, Office for Official Publications of the European Communities, Luxembourg, (2004)
3. UNECE/Eurostat: Work session on statistical data confidentiality, Geneva, 9 to 11 November 2005, Office for Official Publications of the European Communities, Luxembourg (2006)
4. Domingo-Ferrer, J., Torra, V. (eds.): PSD 2004. LNCS, vol. 3050. Springer, Heidelberg (2004)
5. Domingo-Ferrer, J., Franconi, L. (eds.): PSD 2006. LNCS, vol. 4302. Springer, Heidelberg (2006)
6. UNECE: Managing Statistical Confidentiality and Microdata Access, UN Economic Commission for Europe, Conference of European Statisticians (2007)

Author Index