Josep Domingo-Ferrer
Luisa Franconi (Eds.)

LNCS 4302

Privacy in Statistical Databases

CENEX-SDC Project International Conference, PSD 2006
Rome, Italy, December 2006
Proceedings

Springer

Lecture Notes in Computer Science 4302

Josep Domingo-Ferrer   Luisa Franconi (Eds.)

Privacy in Statistical Databases

CENEX-SDC Project International Conference, PSD 2006
Rome, Italy, December 13-15, 2006
Proceedings

Springer

Volume Editors

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

Luisa Franconi
ISTAT, Servizio Progettazione e Supporto Metodologico
nei Processi di Produzione Statistica
Via Cesare Balbo 16, 00184 Roma, Italy
E-mail: franconi@istat.it

Library of Congress Control Number: 2006936080

CR Subject Classification (1998): H.2.8, H.2, G.3, K.4.1, I.2.4

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

ISSN 0302-9743
ISBN-10 3-540-49330-1 Springer Berlin Heidelberg New York
ISBN-13 978-3-540-49330-3 Springer Berlin Heidelberg New York

Springer is a part of Springer Science+Business Media

springer.com

Printed in Germany

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India
Printed on acid-free paper SPIN: 11930242 06/3142 5 4 3 2 1 0

Preface

Privacy in statistical databases is a discipline whose purpose is to provide solutions to the conflict between the increasing social, political and economical demand of accurate information, and the legal and ethical obligation to protect the privacy of the individuals and enterprises to which statistical data refer. Beyond law and ethics, there are also practical reasons for statistical agencies and data collectors to invest in this topic: if individual and corporate respondents feel their privacy guaranteed, they are likely to provide more accurate responses.

There are at least two traditions in statistical database privacy: 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. Both started in the 1970s, but the 1980s and the early 1990s saw little privacy activity on the computer science side. The Internet era has strengthened the interest of both statisticians and computer scientists in this area. Along with the traditional topics of tabular and microdata protection, some research lines have revived and/or appeared, such as privacy in queryable databases and protocols for private data computation.

Privacy in Statistical Databases 2006 (PSD 2006) was the main conference of the CENEX-SDC project (Center of Excellence in SDC), funded by EUROSTAT (European Commission) and held in Rome, December 13–15, 2006. PSD 2006 is a successor of PSD 2004, the final conference of the CASC project (IST-2000-25069), held in Barcelona in 2004 and with proceedings published by Springer as LNCS vol. 3050. Those two PSD conferences follow a tradition of high-quality technical conferences on SDC which started 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 in Springer LNCS vol. 2316.

The Program Committee accepted 31 papers out of 45 submissions from 17 different countries in four 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 official statistics and computer science. Covered topics include methods for tabular data protection, methods for individual data (microdata) protection, assessments of analytical utility and disclosure risk, protocols for private computation, case studies and SDC software.

We are indebted to many people. First, to EUROSTAT for sponsoring the CENEX project and PSD 2006. Also, to those who made the conference and these proceedings possible: the Organization Committee (Xenia Caruso, Jordi Castellà-Roca, Maurizio Lucarelli, Jesús Manjón, Antoni Martínez-Ballesté and Micaela Paciello). In evaluating the papers we received the help of the Program

Committee and the following external reviewers: Lisa Dragoset, José Antonio González, Krish Muralidhar, Bryan Richetti and Monica Scannapieco.

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

September 2006

Josep Domingo-Ferrer
Luisa Franconi

Privacy in Statistical Databases - PSD 2006

Program Committee

John Abowd (Cornell University and Census Bureau, USA)
Jordi Castro (Polytechnical University of Catalonia)
Lawrence Cox (National Center for Health Statistics, USA)
Ramesh Dandekar (Energy Information Administration, USA)
Josep Domingo-Ferrer (Rovira i Virgili University, Catalonia)
Mark Elliot (Manchester University, UK)
Luisa Franconi (ISTAT, Italy)
Sarah Giessing (Destatis, Germany)
Jobst Heitzig (Destatis, Germany)
Anco Hundepool (Statistics Netherlands)
Ramayya Krishnan (Carnegie Mellon University, USA)
Julia Lane (NORC/University of Chicago, USA)
Jane Longhurst (Office for National Statistics, UK)
Silvia Polettini (University of Naples, Italy)
Gerd Ronning (University of Tübingen, Germany)
Juan José Salazar (University of La Laguna, Spain)
Maria João Santos (EUROSTAT, European Commission)
Eric Schulte-Nordholt (Statistics Netherlands)
Francesc Sebé (Rovira i Virgili University, Catalonia)
Natalie Shlomo (University of Southampton, UK; Hebrew University, Israel)
Chris Skinner (University of Southampton, UK)
Julian Stander (University of Plymouth, UK)
Vicenç Torra (IIIA-CSIC, Catalonia)
William E. Winkler (Census Bureau, USA)

Program Chair

Josep Domingo-Ferrer (Rovira i Virgili University, Catalonia)

General Chair

Luisa Franconi (ISTAT, Italy)

Organization Committee

Xenia Caruso (ISTAT, Italy)
Jordi Castellà-Roca (Rovira i Virgili University, Catalonia)

Maurizio Lucarelli (ISTAT, Italy)
Jesús Manjón (Rovira i Virgili University, Catalonia)
Antoni Martínez-Ballesté (Rovira i Virgili University, Catalonia)
Micaela Paciello (ISTAT, Italy)

Table of Contents

Methods for Tabular Protection

Utility and Risk in Tabular Protection

Methods for Microdata Protection

Utility and Risk in Microdata Protection

Protocols for Private Computation

Case Studies

Software

A Method for Preserving Statistical Distributions Subject to Controlled Tabular Adjustment

Lawrence H. Cox[1], Jean G. Orelien[2], and Babubhai V. Shah[2]

[1] National Center for Health Statistics, 3311 Toledo Road
Hyattsville, MD
LCOX@CDC.GOV
[2] Scimetrika, LLC, 100 Capitola Drive
Research Triangle Park, NC 27713 USA

Abstract. Controlled tabular adjustment preserves confidentiality and tabular structure. Quality-preserving controlled tabular adjustment in addition preserves parameters of the distribution of the original (unadjusted) data. Both methods are based on mathematical programming. We introduce a method for preserving the original distribution itself, a fortiori the distributional parameters. The accuracy of the approximation is measured by minimum discrimination information. MDI is computed using an optimal statistical algorithm—iterative proportional fitting.

Keywords: minimum discrimination information; iterative proportional fitting; entropy; Kolmogorov-Smirnov test.

1 Introduction

Statistical disclosure limitation (SDL) in tabular data aims to prevent the data user (or snooper) from inferring with accuracy 1) small cell values in categorical data (cell values based on counts of units) or 2) the contribution of any respondent to a cell total in magnitude data (cell values based on aggregates of quantities pertaining to units). SDL in tabular data is driven by a disclosure rule (known as a sensitivity measure) that quantifies notions of "accurate estimate", "safe value", etc. ([1]). SDL can be achieved in categorical data by several methods including rounding ([2]) and perturbation ([3]) but until recently only cell suppression ([4], [5], [6]) was suitable for SDL in tabular magnitude data. Cell suppression is undesirable for several reasons but especially because it thwarts data analysis for the casual user by removing cell values from the tabulations (leaving "holes" in the data) and for the sophisticated user because the removal process is not at random.

Controlled tabular adjustment (CTA) is a method for SDL in tabular data. CTA is a perturbative method, viz., replaces unsafe (sensitive) values by safe values and replaces selected nonsensitive values with nearby values to restore additive structure. For magnitude data in particular, this is an important improvement over cell suppression because it provides the user a fully populated table for analysis. CTA methodology heretofore has been based on mathematical (mostly, linear) programming. Introducing suitable linear objective functions ([7]) and linear constraints to the CTA model ([8])

J. Domingo-Ferrer and L. Franconi (Eds.): PSD 2006, LNCS 4302, pp. 1 – 11, 2006.

enables quality-preserving controlled tabular adjustment (QP-CTA)--CTA that in addition approximately preserves distributional parameters such as means and (co)variances and regressions.

In this paper, we introduce a new form of CTA aimed at preserving the distribution of original data, based on a well-known statistical algorithm for achieving minimum discrimination information (MDI) or Kullback-Leibler distance ([9]). MDI is aimed at preserving the overall distribution and, a fortiori, preserves the distributional parameters. In Section 2, we summarize the CTA problem. In Section 3, we present the new method, MDI-CTA, and in Section 4 examine its computational and statistical performance. Section 5 offers concluding comments.

2 The SDL Problem for Tabular Data

A tabular cell is considered sensitive if the publication of the true cell value is likely to disclose a contributor's identity or data to a third party. Confidentiality protection for tabular data is based on assuring that all released tabular cells satisfy an appropriate disclosure rule. Cells failing to satisfy the rule, called sensitive cells, are assigned protection ranges defined by lower and upper bounds on the true cell value. Values lying between the bounds are treated as unsafe; those at or beyond either bound are safe. The bounds are computed from the disclosure rule, assuring a framework for SDL that is equitable across respondents and sensible mathematically. See [1], [3], [6] for details and examples. Controlled tabular adjustment assigns a safe value to each sensitive cell (often, but not necessarily, one of its bounds) and the original or a nearby value to each nonsensitive cell. This is accomplished via linear programming (LP) to assure that tabular structure is preserved (Note: our method is not LP-based). Keeping adjusted nonsensitive data close to original data appeals to intuitive notions of data quality. It is possible to quantify and enable a number of these notions ([7]), as follows. If each nonsensitive adjustment can be restricted to lie within two multiples of the cell value's estimated standard error, then arguably the adjusted nonsensitive values are indistinguishable statistically from original values. These conditions, when feasible, are enforceable via LP capacity constraints. Capacities can in addition be parameterized to avoid infeasible problem statements. The Euclidean distance between adjusted and original data can be further restricted by choice/manipulation of the LP objective function. Euclidean concepts are related to statistical concepts, but often imprecisely, and consequently Euclidean reasoning goes only so far to address statistical data quality. Cox et al. ([8]) investigate this problem and provide additional linear constraints aimed at approximately preserving distributional parameters and regressions for normal distributions. We take this further and provide a method for preserving arbitrary distributions, a fortiori the distributional parameters.

Each sensitive cell may be adjusted to (or beyond) one of two values: upper or lower safe value. This results into 2^n combinations for n sensitive cells. A rigorously mathematical optimal solution to CTA requires solving a binary integer linear program. Integer programming works well when n is small, but requires computing resources growing exponentially with n. One of two approaches, or a combination, is needed for quality-assured CTA: a heuristic for selecting combinations that are most likely to lead to the optimal solution and/or a stopping rule based on distributional distance (MDI) between adjusted and original data which indicates when a sufficiently good solution has been reached. We examine these issues next.

3 MDI-CTA

We propose an algorithm based on Kullbak-Leibler minimum discrimination information (MDI) and the iterative proportional fitting procedure (IPFP). MDI is a measure of distance between two statistical (distribution) functions. Other measures of distance such as conditional Chi-square or Kolmogorov-Smirnov distance were considered but MDI was preferred for computing the adjustments because it achieves minimal distance and has other desirable properties. We define MDI in Section 3.1 and provide an algorithm to apply MDI to CTA in Sections 3.2-3.3. The MDI solution is evaluated via three standard statistical tests in Section 4.

3.1 Definition of MDI

Kullback and Leibler [9] proposed a statistic denoted discrimination information to measure the "distance" or "divergence" between two statistical populations. A special case of this statistic is Mahalanobis distance. Discrimination information is also referred to as expected weight of evidence, Renyi's information gain, entropy, entropy distance or cross-entropy. The key points of [9] are summarized below.

Consider a set of points ω in a space Ω. Suppose, the hypotheses H_1 and H_2 imply two functions $f_1(\omega)$ and $f_2(\omega)$ over Ω. One way to choose H_1 over H_2 given that H_1 is true is defined by the mean discrimination information:

$\mathrm{I}(f_1 : f_2) = \int_\Omega f_1(\omega)\log\left(\frac{f_1(\omega)}{f_2(\omega)}\right)d\omega$ when the space Ω is continuous and

$\mathrm{I}(f_1 : f_2) = \sum_\Omega f_1(\omega)\log\left(\frac{f_1(\omega)}{f_2(\omega)}\right)$ when Ω is discrete.

Given a probability distribution $\pi(\omega)$ over the set of cells or space Ω such that $\sum_\Omega \pi(\omega) = 1$, assume a family of distributions $P\{p(\omega)\}$ which satisfies certain constraints (e.g., $\sum_\Omega p(\omega) = 1$). The density function $p^*(\omega)$ of P that is closest to $\pi(\omega)$ minimizes (over P) the expression:

$$I(p : \pi) = \sum_\Omega p(\omega)\log\left(\frac{p(\omega)}{\pi(\omega)}\right)$$

Some properties of the MDI are:

- $I(p:\pi)$ is a convex function hence the procedure yields a unique minimum
- If $p^*(\omega)$ is the MDI estimate, it can be shown that for any member $p(\omega)$ of P $I(p:\pi) = I(p^*:\pi) + I(p^*:p)$
- $I(p:\pi) \geq 0$ with equality if and only $\pi(\omega) = p(\omega)$

3.2 Applying MDI to CTA

The CTA problem for a 3-dimensional table can be stated as follows. Given a table with values O_{drc} (with the indices d, r, c representing depth, row and column) in which there are sensitive cells, we want to find the set of adjusted values A_{drc} with which to replace values in the sensitive cells so that $K = \sum_d \sum_r \sum_c A_{drc} \log(A_{drc} / O_{drc})$ is minimized subject to the constraints that all marginals are preserved. For a sensitive cell, A_{drc} is either the lower or upper bound and for the nonsensitive cells A_{drc} correspond to adjustment made to preserve the marginals. Because, as the number of sensitive cells increase, it is not possible to find the minimum by computing K for all possible combinations, we propose an algorithm that consists of initial heuristic steps to select binary up/down directions for change for the sensitive cells, followed by IPFP steps to preserve the marginals and achieve (optimal) MDI subject to the binary choice, and subsequent attempts to improve the solution or confirm global optimality, viz., an adjusted table closest in distribution to the original table conditional on safeness of the sensitive adjustments. This obviates the need to have separate constraints such as preservation of mean, variance or having correlation between the values in the two tables being close to 1.

The first heuristic step finds a local solution for each level of depth, row and column. Assume that there are n_r rows and denote by r_i the number of sensitive cells within the i^{th} row i $(i = 1,\ 2,\ ...n_r)$. Within each row, for each of the possible 2^{r_i} combinations, we adjust the value of the nonsensitive cells so that the sum of the adjusted values for the non-sensitive cells in that row equal the sum of the original values for the non-sensitive cells. Let $T_{1r_{ig}}$ denote the adjusted values over the sensitive cells for the g^{th} of the possible 2^{r_i} combinations, T_{2r_i} denote total of the original values over the non-sensitive cells and T_{+r_i+} denote the sum of all original values in that cell. Adjusted values A_{drc} for the non-sensitive cells for that combination are given by:

$A_{drc} = \frac{(T_{+r_i+} - T_{1r_{ig}})}{T_{2r_i}} O_{drc}$ Next, we compute the Kullback-Leibler MDI value for the row, $K_{ri} = \sum_{row=i} A_{drc} \log(A_{drc} / O_{drc})$, select the combination that produced the minimum value for K_{ri}, and save that combination. For this combination, define $u_{drc} = C_{drc}$

where $C_{drc} = 1$ if the value of the sensitive cell was adjusted up and 0 if the value of the cell was adjusted down.

Implementing the steps described above across columns and depths, define binary values v_{drc} and w_{drc} that come respectively from the combination that produced the lowest MDI for each depth and column. Each sensitive cell is replaced by either a lower or upper safe value independently in each of three steps (by searching for a local solution across row, depth or column). For a given sensitive cell, we assign the lower safe value if and only if at least two of the above steps prefer a lower safe value for that cell; otherwise we assign the upper safe value to that cell. Our preliminary solution for the sensitive cells is obtained by majority rule, we assign:

$$C_{drc} = 0, \text{ if } u_{drc} + v_{drc} + w_{drc} \le 1$$
$$C_{drc} = 1, \text{otherwise.}$$

The method for 2-dimensional tables (which can be generalized to tables of even dimension) breaks the tie by comparing the change in entropy and accepts the one that has smaller change in the entropy function η defined as $\eta(P) = P\log(P)$, and $\eta(0) = 0$. Specifically for a 2-dimensional table, we find the binary value v_{rc} and w_{rc} for each cell using best combination based on the MDI for each row and column. When the values of v_{rc} and w_{rc} disagree, we break the tie with values u_{rc} assigned as follows:

$$u_{rc} = 1, \text{ if and only if } \quad |\eta(U_{rc}) - \eta(O_{rc})| \ge |\eta(L_{rc}) - \eta(O_{rc})|$$
$$u_{rc} = 0, \text{ otherwise}$$

where U_{rc} and L_{rc} represent the upper and lower safe value for the cell.

The steps described above yield a preliminary solution for the sensitive cells. Our next step is to adjust values of the nonsensitive cells so that the marginal totals are preserved, using the iterative proportional fitting procedure (IPFP). Let A_{d++}, A_{+r+}, and A_{++c} denote the marginal totals respectively for each of the d depths, each of the r rows and each of the c columns. The target marginal totals for IPFP are $O_{d++} - A_{d++}$, $O_{+r+} - A_{+r+}$ and $O_{++c} - A_{++c}$ corresponding respectively to a) total of the original values in depth n_d minus total of the sensitive values in that depth, b) total of the original values in row n_r minus total of the sensitive values in that row, and c) total of the original values in column n_c minus total of the sensitive values in that column, respectively. In what follows, we describe how the IPFP algorithm for 2-dimensional tables with r rows and c columns. Generalization to higher dimension is straightforward.

3.3 IPFP Algorithm

For a non-sensitive cell with value O_{rc}, let

$$P_{rc}^{m,r} = \frac{O_{r+} - A_{r+}}{P_{r+}^{m-1,r}} P_{rc}^{m-1,r} \quad \text{and} \quad P_{r+}^{m-1,r} = \sum_{c} P_{rc}^{m-1,r}$$

with summation over original non-sensitive values in a row. For m = 1:

$$P_{rc}^{m-1,r} = O_{rc}\,,\ P_{rc}^{m,c} = \frac{O_{+c} - A_{+c}}{P_{+c}^{m-1,c}} P_{rc}^{m-1,c}\ ,\ P_{+c}^{m-1,c} = \sum_{r} P_{rc}^{m-1,c}\ ,\ P_{rc}^{m-1.c} = P_{rc}^{1,n_r}$$

wherein the starting values for columns are those obtained after the first iteration in the last row n_r. The algorithm stops when

$$| O_{r+} - P_{r+}^{m,r} | < 0.25 \qquad | O_{+c} - P_{+c}^{m,c} | < 0.25$$

for all r and c or if m, the number of iterations, equals 25.

After this IPFP, we now have adjusted values for both sensitive and non-sensitive cells and compute $K = \sum_d \sum_r \sum_c A_{drc} \log(A_{drc} / O_{drc})$. We now select one of the sensitive cells and reverse the safe value from upper to lower or lower to upper for that cell and use IPFP to obtain adjusted values for the nonsensitive cells. If the result is a smaller value for MDI, then we accept the revision. We repeat this process for each of the sensitive cells. This is equivalent to a stepwise search one cell at a time.

There is no guarantee that we may find the optimal solution. Let us assume that at least proportion p of the 2^n possible combinations is better than the one finally found above. If we select a random set of m combination of sensitive cells for replacement by upper or lower safe values, then the chance of finding at least one improved solution is given by $\alpha = 1 - (1 - p)^m$. Solving for m, we obtain: $m = \log(1 - \alpha) / \log(1 - p)$. If p = 0.001, to obtain 99% confidence we need to test 4,603 random combinations. If we do find a better solution then we can repeat the process with a new set of random combinations. The number of random combinations to check will depend on available resources and the desired level of confidence. Note that for each random combination of sensitive values, IPFP needs to be applied to preserve the marginals.

4 Evaluation of MDI-CTA

To evaluate whether the solution from MDI-CTA preserves statistical distribution, we computed correlation and simple linear regression coefficient between the original and adjusted values. If statistically insignificant changes are made to the original data, then one would expect a simple linear regression to yield a y-intercept $b_0 = 0$ and a regression coefficient $b_1 = 1$. Also, we use statistical testing to ascertain whether the original and adjusted values come from the same statistical distribution. To perform this evaluation, we used randomly generated data and tables, a 4x9 table (7 sensitive cells) and a 13x13x13 table (approximately 200 sensitive cells) from Cox et al. [8] and a 30x30 table (105 sensitive cells). Cox provided optimal QP-CTA solutions for the last three tables.

4.1 Change in Distributional Parameters and Regression Coefficient

Table 1 shows high correlation values were obtained between adjusted and original values for randomly generated tables. Regression parameters b_0 and b_1 were close to 0 and 1 in each case. 5000 random repetitions were performed, except for 20x20x20, 20 percent sensitive (100 repetitions). For the 4x9, 30x30 and 13x13x13 tables, results are nearly identical to the (perfect) results of optimal QP-CTA solutions and are not reported below.

Table 1. Results of simulation experiments

Table size	**Percent sensitive**	**MDI from random sample (or all combinations) (Q2.5, Q97.5)**	$\frac{b_0}{mean}$ **regression of adjusted on original**	b_1	**Corr-elation**	**Mean pct. change to nonsensitive (min, max)**
10x10	5%	67.82 (67, 85)	-0.02	1.02	0.99	-0.00 (-0.04, 0.03)
10x10	10%	1617.17 (1695, 2131)	0.02	0.98	0.99	-0.00 (-0.03, 0.04)
10x10	10%	103.083 (106, 146)	0.00	1.00	0.99	-0.01 (-0.06, 0.04)
10x10	30%	181.13 (198, 309)	0.00	1.00	0.95	-0.01 (-0.11, 0.13)
5x5x5	5%	104.94 (104, 170)	-0.01	1.01	0.99	-0.00 (-0.08, 0.04)
5x5x5	10%	678.4 (724, 1277)	-0.02	1.02	0.98	-0.01 (-0.17, 0.09)
5x5x5	10%	373.33 (430, 889)	0.01	0.99	0.98	-0.00 (-0.18, 0.17)
5x5x5	30%	3469.44 (4603,13618)	-0.13	1.12	0.79	-0.01 (-0.34, 0.33)
10x10x10	5%	3173.66 (3307, 3455)	-0.02	1.02	0.95	-0.00 (-0.06, 0.04)
10x10x10	10%	5469.14 (5725, 6424)	-0.03	1.03	0.93	-0.00 (-0.08, 0.07)
10x10x10	20%	5832.39 (6064, 6908)	-0.02	1.02	0.92	-0.00 (-0.06, 0.07)
10x10x10	30%	15948.72 (21071, 26255)	0.06	0.94	0.78	-0.00 (-0.11, 0.15)
20x20x20	5%	11858.12 (12033, 12390)	0.00	1.00	0.99	-0.00 (-0.045, 0.035)
20x20x20	10%	26790.78 (27177, 28003)	-0.01	1.00	0.97	-0.00 (-0.06, 0.05)
20x20x20	20%	27750.5 (27824, 28679)	0.00	1.00	0.97	-0.00 (-0.05, 0.05)
20x20x20	30%	85086.48 (86221, 89708)	0.01	0.99	0.92	-0.00 (-0.09, 0.09)

4.2 Goodness-Of-Fit (GOF) Statistics to Compare Adjusted vs. Original

GOF test were used to ascertain whether the adjusted values have the same statistical distribution as the original values. The comparison of the statistical distribution of original and observed values was performed using Kolmogorov-Smirnov (K-S), Kuiper and Chi-square. Because any CTA solution is conditional on net adjustments to sensitive values that are prespecified (in the case of symmetric protection) or at least lower-bounded, the question arises whether it is appropriate to compare original and adjusted distributions on all cells or only on the nonsensitive cells, viz., conditional on the sensitive adjustments. It is known that the Chi-square GOF can lead to rejection of the null hypothesis even when there are small insignificant departures from a specified theoretical distribution (Pederson and Johnson [14]; Laren et al. [15]). So, if sensitive values (representing larger deviations, in general) were included in the analysis, one could only expect large values for the Chi-square test statistic leading to rejection of the null hypothesis that the distributions are equal. Similarly, one might expect a Chi-square GOF test based only on the nonsensitive cells to lead to rejection of the null hypothesis for large samples. The K-S and Kuiper tests are not likely to be affected by these problems. So, we test conditionally on Chi-square and unconditionally on the other two. Below we describe the K-S GOF test.

The Kolmogorov-Smirnov (K-S) test is widely used to assess whether two samples come from the same distribution. K-S is non-parametric and makes no distributional assumptions on the data. The test uses the empirical distribution function (EDF). Consider two samples $x_1, x_2 \ldots x_\mathrm{n}$ and $y_1, y_2 \ldots y_n$ (equal number of observations not required). For any sample $\{x_j\}$, EDF is given by: $F_X(x) = \frac{1}{n}(\text{number of } x_j \;\le\; x) = \frac{1}{n}\sum_{\mathrm{i}=1}^{\mathrm{n}} I(x \ge x_j)$. We compute the test statistic as: $D = \max_j \mid F_X(x_j) - F_Y(y_j) \mid,\ j = 1,\ 2,\ \ldots n$. The p-value is obtained by computing the probability of observing a larger D value. This test is carried out using the SAS System version 9.1 (SAS Institute, Cary, NC). These p-values are based on asymptotic distribution the quality of which was investigated by Hodges [10]. Several authors (for example Shmidt and Trede [11] or Shroer and Trenkler [12) have shown that K-S test tends to be less powerful than similar tests such as Cramer-von Mises when the two samples differ only with respect to a location parameter. However, when the underlying distributions have extreme values, outliers or are asymmetric the K-S test is preferred ([12], [13]). Because we are dealing with nonnegative magnitude data and no restriction on the upper bound, we believe that the use of the K-S test is justified to assess if the original and adjusted values have the same statistical distribution.

Table 2 reports on change to the original distribution, measured by (unconditional) K-S and Kuiper statistics and a (conditional) Chi-square statistic. Results are encouraging.

Table 2. p-values for comparing original and adjusted values

Table Size	**Percent Sensitive**	**Kolmog-Smir p-values: adjust & orig from same distrib (unconditional)**	**Kuiper p-values: adjust & orig from same distribution (unconditional)**	**Chi-square p-values: adjust & orig from same distribution (conditional)**
10x10	5%	1.00	1.00	1.00
10x10	10%	1.00	1.00	0.00
10x10	10%	0.97	0.98	1.00
10x10	30%	0.97	0.91	0.87
5x5x5	5%	1.00	1.00	0.99
5x5x5	10%	1.00	1.00	0.00
5x5x5	10%	0.99	0.97	0.00
5x5x5	30%	0.0587	0.00	0.00
10x10x10	5%	0.79	0.66	0.00
10x10x10	10%	0.72	0.40	0.00
10x10x10	20%	0.65	0.34	0.00
10x10x10	30%	0.019	0.00	0.00
20x20x20	5%	0.98	0.95	1.00
20x20x20	10%	0.60	0.16	1.00
20x20x20	20%	0.51	0.21	1.00
20x20x20	30%	0.056	0.00	0.00

Conditional = distributional distance computing using only nonsensitive cells (*distance conditional on sensitive values*)
Unconditional = distance computing using all cells

Table 2 demonstrates that MDI-CTA achieves its objective of modifying sensitive values while preserving the original distribution. Note that the hypothesis that original and adjusted values came from the same population was rejected under K-S or Kuiper only for simulated tables with more than 30% of sensitive cells. As expected, Chi-square results are extreme for large tables and tables with a higher proportion of sensitive cells, even when restricted to the nonsensitive deviations.

5 Concluding Comments

Key features and benefits of MDI-CTA are:

- the solution leads to a distribution close(st) to the original
- the optimization criterion is consistent with the statistical objective

- the heuristics are based on partial evaluation of the criterion making computing more efficient
- IPFP is a simple but effective and well-proven technique for tables
- a stepwise approach provides additional chance for improvement
- we provide statistical evaluation for confidence in the final solution
- the algorithm is easily expandable to many dimensions

The algorithm is based on Kullback information that has excellent properties: K = 0, if and only if two distributions are identical. In all other cases K is greater than 0. Further more, K is finite, $A_{rc} = 0$ whenever $O_{rc} = 0$. It is convex and hence has unique minimum. Since, MDI is a good measure of closeness between two distributions, by minimizing it, we are keeping overall statistical information as close to the original as feasible. It has been proven that IPFP minimizes Kullbak-Leibler discrimination information subject to the specified constraints. The solution is also a maximum likelihood solution under a log-linear model. Computational results on change to distributional parameters and distributional distance are encouraging.

Next steps for research include incorporating changes to (selected) zero cells and to marginal totals (normally held fixed by MDI) and enhancing, improving or replacing the binary heuristic. Issues surrounding goodness-of-fit testing—conditional vs. unconditional—selection of the most appropriate test—also need to be considered.

Disclaimer. This paper represents the work of the authors and is not intended to represent the policies or practices of the Centers for Disease Control and Prevention or any other organization.

References

1. Cox, L.H.: Linear sensitivity measures in statistical disclosure control. Journal of Statistical Planning and Inference, Volume 5 (1981) 153-164.
2. Cox, L.H. and Ernst, L.: Controlled rounding. INFOR, Volume 20 (1982) 423-432.
3. Cox, L.H., Fagan, J., Greenberg, B. and Hemmig, R.: Research at the Census Bureau into disclosure avoidance techniques for tabular data. In: Proceedings of the Survey Research Methods Section, American Statistical Association. Alexandria, VA (1986) 388-393.
4. Cox, L.H.: Suppression methodology and statistical disclosure control. Journal of the American Statistical Association, Volume 75 (1980) 377-385.
5. Cox, L.H.: Network models for complementary cell suppression. Journal of the American Statistical Association, Volume 90 (1995) 1153-1162.
6. Fischetti, M. and Salazar-Gonzalez, J.J.: Models and algorithms for optimizing cell suppression in tabular data with linear constraints. Journal of the American Statistical Association, Volume 95 (2000) 916-928.
7. Cox, L.H. and Dandekar, R.A.: A new disclosure limitation method for tabular data that preserves data accuracy and ease of use. In: Proceedings of the 2002 FCSM Statistical Policy Seminar. U.S. Office of Management and Budget, Washington (2004) 15-30.
8. Cox L.H., Kelly J.P., and Patil R.: Balancing quality and confidentiality for multivariate tabular data. In: Domingo-Ferrer, J. and Torra, V. (eds.): Privacy in Statistical Data Bases, Lecture Notes in Computer Science, Volume 3050. Springer-Verlag, Heidelberg (2004) 87–98.

9. Kullback S and Leibler R.A.: On information and efficiency. Annals of Mathematical Statistics, Volume 86 (1951) 79-86.
10. Hodges, J.L. Jr.: The significance probability of the Smirnov two-sample test. Arkiv for Matematik, Volume 3 (1957) 469-486.
11. Schmid F. and Trede M.: A distribution free test for the two sample problem for general alternatives. Computational Statistics and Data Analysis, Volume 20 (1994) 409-419.
12. Shroer G and Trenkler D.: Exact and randomization distributions of Kolmogorov-Smirnov tests two or three samples. Computational Statistics and Data Analysis, Volume 20 (1995) 185-202.
13. Buning H. : Robustness and power of modified Lepage, Kolmogorov-Smirnov and Cramer-Von Mises two sample tests. Journal of Applied Statistics, Volume 29 (2002) 907-924.
14. Pederson, S.P. and Johnson, M.E.: Estimating model discrepancy. Technometrics, Volume 32 (1990) 305-314.
15. McLaren C.E., Legler J.M. and Brittenahm G.M.: The generalized Chi-square goodness-of-fit test. The Statistician, Volume 43 (1994) 247-258.

Automatic Structure Detection in Constraints of Tabular Data

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. Methods for the protection of statistical tabular data—as controlled tabular adjustment, cell suppression, or controlled rounding—need to solve several linear programming subproblems. For large multidimensional linked and hierarchical tables, such subproblems turn out to be computationally challenging. One of the techniques used to reduce the solution time of mathematical programming problems is to exploit the constraints structure using some specialized algorithm. Two of the most usual structures are block-angular matrices with either linking rows (primal block-angular structure) or linking columns (dual block-angular structure). Although constraints associated to tabular data have intrinsically a lot of structure, current software for tabular data protection neither detail nor exploit it, and simply provide a single matrix, or at most a set of smallest submatrices. We provide in this work an efficient tool for the automatic detection of primal or dual block-angular structure in constraints matrices. We test it on some of the complex CSPLIB instances, showing that when the number of linking rows or columns is small, the computational savings are significant.

Keywords: statistical disclosure control, cell suppression, controlled tabular adjustment, linear constraints, multilevel matrix ordering algorithms.

1 Introduction

From an algorithmic point of view, one of the main differences between disclosure control techniques for microdata and tabular data is that the latter must deal with many linear constraints, associated to total and subtotal cells. Current methods for tabular data protection, such as, e.g., cell suppression [4,9,14], controlled tabular adjustment [6,12] and controlled rounding [10,20], deal with

⋆ Supported by the Spanish MCyT Project TIC2003-00997.

⋆⋆ Corresponding author.

J. Domingo-Ferrer and L. Franconi (Eds.): PSD 2006, LNCS 4302, pp. 12–24, 2006.

those linear additivity constraints through mathematical programming technology. Unfortunately, the resulting optimization problems turn out to be computationally expensive, even for continuous variables. For instance, the simplex algorithm, which is the preferred option for many linear programming problems [2], has shown to be inefficient compared to polynomial-time interior-point algorithms [21] when dealing with tabular data constraints [6,20].

One of the most used techniques in mathematical programming for reducing the computational cost of a problem is to exploit its structure, either through decomposition or partitioned basis factorization. Two of the most relevant structures are primal block-angular

$$A = \begin{bmatrix} A_1 & & & \\ & A_2 & & \\ & & \ddots & \\ & & & A_k \\ L_1 & L_2 & \dots & L_k \end{bmatrix}, \tag{1}$$

where k is the number of diagonal blocks, $A \in \mathbb{R}^{m \times n}, A_i \in \mathbb{R}^{m_i \times n_i}, L_i \in \mathbb{R}^{l \times n_i}, i = 1, \dots, k$, and dual block-angular

$$A = \begin{bmatrix} A_1 & & & L_1 \\ & A_2 & & L_2 \\ & & \ddots & \vdots \\ & & A_k & L_k \end{bmatrix}, \tag{2}$$

k, A, A_i being as before, and $L_i \in \mathbb{R}^{m_i \times l}$. Structures (1) and (2) appear in problems with l linking constraints and l linking variables, respectively, and have been extensively studied in the literature [3, Chapter 12]. Classical decomposition procedures, based on the simplex method, are Dantzig-Wolfe for primal block-angular structures [13], and Benders for dual block-angular ones [1]. Specialized interior-point approaches for structured problems have also been recently suggested [8,15]; these are promising approaches for tabular data protection, since, as noted above, interior-point methods outperform simplex implementations in this class of problems. It is worth noting that homogeneous sizes for diagonal blocks A_i benefit the performance of any decomposition approach, mainly if some sort of parallelism is going to be applied.

Unfortunately, current methods and software for tabular data protection do not exploit constraints structure in general tables. Structure has only been exploited for two-dimensional tables with at most one hierarchical variable, whose constraints are modeled as a network [7,9], and for three-dimensional tables, that provide multicommodity network models [5]. For general tables, state-of-the-art protection software, as τ-Argus [17], provide a single constraints matrix, or at most a set of linked submatrices, without detailing each matrix structure. Indeed, the constraints structure is particular to each kind of table, and it is not clear that fully exploiting such structure would be worthy for an optimization algorithm. In addition, writing software for the detection of the particular structure of any table would be a cumbersome task. The purpose of this work is to overcome such difficulties, i.e., to develop and test a tool for automatic detection of structures (1) and (2) in constraints matrices derived from tabular data. The tool developed is based on the multilevel matrix ordering algorithm for unsymmetric matrices of [16]. Unlike other highly recognized and efficient algorithms, such as that implemented in METIS [18], the former is tailored for unsymmetric matrices, which is the case for tabular data constraints. The procedure is applied to a set of standard tabular data instances, being its behaviour instance dependent: for most instances and $k = 2$ a small linking block (i.e., l/m (l/n) for primal (dual) block-angular structures is less than 0.2), whereas for others the relative size of the linking block can be up to 0.4 (for $k > 2$ these relative sizes of the linking block grow). When the relative size of the linking block is small it makes sense to apply a decomposition approach, and we provide preliminary computational results comparing the performance of a linear programming solver depending on whether structure is exploited.

The paper is organized as follows. Section 2 outlines the multilevel matrix reordering algorithm and gives details about its implementation. Section 3 reports results using this algorithm on a standard set of tabular data instances. Finally, Section 4 analyzes the computational savings due to using the reordered matrix in a mathematical programming solver.

2 The Matrix Reordering Algorithm

Given any matrix, obtaining the optimal reordering that transforms the matrix into either the primal block-angular structure (1) or the dual one (2) is a difficult combinatorial optimization problem (in this context, “optimal” means “with the smallest linking block”). Several heuristics have been provided in the past for this problem. We have chosen the recent multilevel approach of [16], which is among the most efficient ones for unsymmetric matrices, and it is implemented in commercial libraries (e.g., in routine HSL_MC66L of the HSL archive, formerly the Harwell Subroutine Library). A comprehensive description of such algorithm is out of the scope of this work, and it will just be outlined; details can be found in [16] and references therein.

We first define some concepts to be used later in the overview of the algorithm. Given a sparse matrix $A \in \mathbb{R}^{m\times n}$, the *net* of column j is the number of rows associated to such column, i.e, $\{i \in \{1,\dots,m\} : a_{ij} \neq 0\}$. A *row partition* is a partition of the set of rows $\{1,\dots,n\}$, i.e., R_1, R_2 such that $R_1 \cap R_2 = \emptyset$ and $R_1 \cup R_2 = \{1,\dots,m\}$. A net is cut by a row partition if there are rows of the net in both R_1 and R_2. The *net-cut* of a row partition is the number of nets cut by this row partition. Note that in a block-diagonal matrix, i.e, without neither linking constraints nor linking columns, the net-cut is zero. Therefore this is the value to be reduced by any heuristic in order to obtain close-to-block-angular structures. The *gain* is the decrement of the net-cut obtained after moving a row from R_1 to R_2 or vice-versa; the gain is negative if the net-cut is increased.

The *edge-weight* of rows (i_1, i_2) is the number of columns shared by these rows, i.e, the cardinality of $\{j \in \{1, \ldots, n\} : a_{i_1,j} \neq 0 \text{ and } a_{i_2,j} \neq 0\}$.

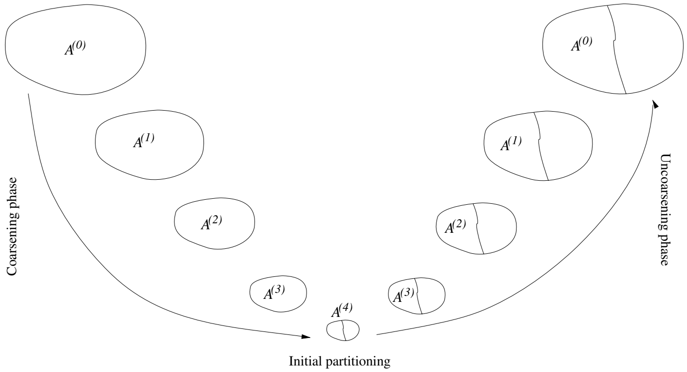


Fig. 1. The three stages of the multilevel ordering algorithm: coarsening, partitioning, uncoarsening; example with 4 levels

The multilevel ordering algorithm of [16] is made of the three following stages, which are shown in Figure 1:

Coarsening phase. Matrix $A = A^{(0)}$ is successively transformed in a sequence of smaller matrices $A^{(1)}, A^{(2)}, \ldots, A^{(r)}$, r being the deepest level, such that the number of rows is reduced at each transformation, i.e., $m = m^{(0)} > m^{(1)} > m^{(2)} > \ldots > m^{(r)}$. The procedure successively collapses the "closest rows" in a single one using the notion of edge-weight defined above. In our particular implementation we used the heavy-edge matching criterion of [18] (see that reference for details).

Partitioning phase. It is based on the classical Kernighan-Lin (KL) heuristic [19] for partitioning graphs. Graphs are known to be associated to symmetric sparse matrices, i.e., $a_{ij} \neq 0$ and $a_{ji} \neq 0$ if there is an edge joining nodes i and j in the graph. The KL heuristic can be extended for unsymmetric matrices if we use the notion of *net-cut* and *gain*, instead of the original one of *edge-cut* (i.e., number of edges in a graph cut by a node partition) of the KL algorithm. In short, the KL algorithm is an iterative procedure that starting from an initial row partition $R_1^{(0)}, R_2^{(0)}$, performs two nested loops. The inner iterations successively look for the row with the largest gain. This row is obtained from the set $R_1^{(0)}$ or $R_2^{(0)}$ with the maximum cardinality, in an attempt to guarantee similar dimensions for diagonal blocks. This row is moved to the other subset of rows, subsets $R_1^{(i)}, R_2^{(i)}$ are updated, and the row is locked. This is performed until all rows have been locked. This ends the inner iterations. The outer iterations repeat the above sequence of inner iterations, unlocking all rows at the beginning, until there is no improvement in the overall net-cut. The procedure records the best partitioning up to now obtained, which is returned as the solution.

If one needs more than two diagonal blocks (i.e., $k > 2$ in (1) or (2)), the KL algorithm can be recursively applied to any resulting submatrix defined by R_1 or R_2, thus eventually obtaining a row partition $R_1, R_2, \ldots, R_k$.

Uncoarsening phase. During this phase the partitioning of $A^{(r)}$ is projected to the original matrix through matrices $A^{(r-1)}$, $A^{(r-2)}$, …, $A^{(1)}$, $A^{(0)}$. Two rows i_1 and i_2 collapsed in a single one in A^h belonging to $R_p, 1 \leq p \leq k$, will appear as two different rows of R_p in A^{h-1} . Optionally, any new matrix A^h can be refined with the KL algorithm.

The KL algorithm itself can be applied to the matrix A (i.e., as if we had a single level algorithm, or equivalently with a coarsening phase with $r = 0$). However, as it will be shown in Section 3, it is computationally expensive, because of the large number of rows m of A, and can provide unsatisfactory partitionings. On the other hand, applied in combination with a multilevel approach (i.e, coarsening and uncoarsening phases) it is extremely efficient, and the partitioning is significantly improved. Note that the coarsening phase collapses rows with the largest edge-weight, and such rows are expected to be in the same subset R_p in a solution, which is exactly what the uncoarsening phase does.

The multilevel algorithm above described has been implemented in C, in a library named *UMOA* (unconstrained matrix ordering algorithm), which is roughly made of 3100 lines of code. The package can be freely obtained from the authors. It has been optimized using efficient data structures for the coarsening and uncoarsening phases (i.e., AVL trees), as well as for the computation and updating of gains and net-cuts in the KL algorithm. Among the several features of the package, we mention that: (i) when $k > 2$ the package allows full control to the user about how to recursively obtain the additional subsets R_p from R_1 and R_2; (ii) it can be used to obtain both primal and dual block-angular structures (1) and (2). (We note that the reordered matrix in primal block-angular form is not equivalent to the transpose of the reordered matrix in dual block-angular form.)

3 Reordering Tabular Data Instances

We applied the multilevel reordering algorithm of Section 2 to a subset of the CSPLIB test suite, a set of instances for tabular data protection (Fischetti and Salazar 2001), plus to additional large instances ("five20b", and "five20c"). CSPLIB can be freely obtained from `http://webpages.ull.es/users/casc/-#CSPlib:`. CSPLIB contains both low-dimensional artificially generated problems, and real-world highly structured ones. Some of the complex instances were contributed by National Statistical Agencies—as, e.g., Centraal Bureau voor de Statistiek (Netherlands), Energy Information Administration of the Department of Energy (U.S.), Office for National Statistics (United Kingdom) and Statistisches Bumdesant (Germany).

Table 1 shows the features of the instances considered. The small CSPLIB instances were omitted. Column "Name" shows the instance identifier. Columns "n", "m" and "N. coef" provide, respectively, the number of columns (cells),

Table 1. Dimensions of the largest CSPLIB instances

Name	n	m	N.coef
bts4	36570	36310	136912
five20b	34552	52983	208335
five20c	34501	58825	231345
hier13	2020	3313	11929
hier13x13x13a	2197	3549	11661
hier13x7x7d	637	525	2401
hier16	3564	5484	19996
hier16x16x16a	4096	5376	21504
jjtabeltest	3025	1650	7590
nine12	10399	11362	52624
nine5d	10733	17295	58135
ninenew	6546	7340	32920
targus	162	63	360
toy3dsarah	2890	1649	9690
two5in6	5681	9629	34310

rows (additivity constraints) and nonzero coefficients of the constraints matrix A to be reordered.

Tables 2, 3 and 4 show, respectively, the results obtained reordering the matrices in dual block-angular form for $k = 2$, in primal block-angular form for $k = 2$, and in dual-block angular form for $k = 4$ and 8. Last two columns of Tables 2 and 3 provide information for the single level algorithm (i.e., KL algorithm was applied to the whole matrix). Columns "m_1", "n_1", "m_2" and "n_2" of tables 2 and 3 provide the number of rows m_i and columns n_i of the two diagonal blocks. Columns "$100 \cdot l/n$" ("$100 \cdot l/m$") of Tables 2 and 4 (Table 3) give the relative size of the linking columns (constraints) block. Columns "CPU" of the three tables report the seconds of CPU time required to compute the reordering. The runs were carried on a standard PC running Linux with an AMD Athlon 1600+ at 1.4GHz and 320 MB of RAM. Therefore, the reordered matrix can be efficiently obtained without the need of sophisticated computational resources.

From Tables 2–4 it can be concluded that:

- The multilevel approach is instrumental in reordering tabular data constraints. It is not only one order of magnitude faster that the single level approach, but also provides much better reorderings (i.e., the linking block becomes much narrower). In particular, all the matrices could be reordered in few seconds on a desktop computer.
- The sizes of the diagonal blocks are similar, which may be a benefit for an optimization solver. This is also instrumental if parallel computations want to be exploited.
- The size of the linking block is instance dependent (from 7.4% in instance "bts4" of Table 2 to 59.8% in instance "hier16" of Table 3). Thus, in principle, not all the reordered matrices are appropriate for a specialized solver for structured problems.

Table 2. Results for dual block-angular ordering with $k = 2$

	Multilevel						Single level	
	A_1		A_2					
Name	m_1	n_1	m_2	n_2	$100 \cdot l/n$	CPU	$100 \cdot l/n$	CPU
bts4	17838	16898	18472	16937	7.4	2	21.4	53
five20b	25196	14343	27787	15586	13.3	6	48.4	597
five20c	28820	14034	30005	15093	15.5	7	54.3	416
hier13	1656	583	1657	563	43.2	0	45.0	0
hier13x13x13a	1890	774	1659	526	40.8	0	46.0	0
hier13x13x7d	792	419	651	218	46.1	0	37.8	0
hier16	2486	740	2998	1069	49.2	1	55.2	1
hier16x16x16a	2614	1074	2762	1138	45.9	0	45.2	0
jjtabeltest	849	1269	801	1383	12.3	0	15.1	0
ninenew	3007	2115	4333	3085	20.5	0	40.5	1
nine5d	9007	4568	8288	4303	17.3	1	47.1	9
nine12	5880	4249	5482	4098	19.7	3	35.5	3
targus	19	26	44	92	27.1	0	21.6	0
toy3dsarah	741	1118	908	1389	13.2	0	24.7	0
two5in6	5344	1749	4285	1502	42.7	0	56.7	3

Table 3. Results for primal block-angular ordering with $k = 2$

	Multilevel						Single level	
	A_1		A_2					
Name	m_1	n_1	m_2	n_2	$100 \cdot l/m$	CPU	$100 \cdot l/m$	CPU
bts4	16307	18389	16137	18181	10.5	2	21.7	37
five20b	26006	17735	22576	16817	12.7	4	57.9	50
five20c	26693	17286	26218	17215	17.1	5	60.3	41
hier13	1375	1091	1076	929	42.6	0	53.7	0
hier13x13x13a	1320	1075	1428	1122	36.4	0	51.6	0
hier13x13x7d	483	562	510	621	38.0	0	36.8	0
hier16	1674	1732	1676	1832	59.8	0	60.4	0
hier16x16x16a	1989	2118	1676	1978	41.7	0	44.6	0
jjtabeltest	737	1457	686	1568	7.5	0	16.8	0
ninenew	3181	3239	2721	3307	21.9	1	46.8	1
nine5d	7111	5383	7975	5350	20.5	1	40.5	2
nine12	4846	5206	4666	5193	17.7	1	45.6	2
targus	22	82	22	80	11.7	0	19.1	0
toy3dsarah	534	1446	529	1444	20.2	0	20.2	0
two5in6	3630	2816	3817	2865	38.4	0	37.1	0

– In general, the size of the linking block increases with k, the number of diagonal blocks. Although the more diagonal blocks, the more "decomposable" becomes the solution of linear systems of equations involving that matrix (see Section 4), the overall solution procedure can become very inefficient due to a large linking block. This tradeoff is usually optimized by small

Table 4. Results for dual block-angular ordering with $k = 4$ and $k = 8$

	$k = 4$		$k = 8$	
Name	$100 \cdot l/n$	CPU	$100 \cdot l/n$	CPU
bts4	15.4	3	22.9	3
five20b	34.0	7	43.8	9
five20c	26.8	9	44.9	11
hier13	66.9	0	82.8	0
hier13x13x13a	64.1	0	80.8	0
hier13x13x7d	71.7	0	88.6	0
hier16	73.2	1	85.4	0
hier16x16x16a	69.4	0	83.7	0
jjtabeltest	24.6	0	36.9	0
ninenew	34.1	0	54.3	0
nine5d	44.3	1	70.6	1
nine12	31.0	1	47.6	1
targus	53.7	0	54.3	0
toy3dsarah	45.6	0	68.9	0
two5in6	73.9	0	87.2	0

values of k (e.g., $k = 2$, 3 or 4), unless the information about the data allows specific larger ones.

- Primal and dual block-angular structures provide linking blocks of different size. The best (primal or dual) ordering is instance dependent, and it seems not to be a clear trend.

The above conclusions are consistent with those obtained in [16] for other types of matrices (for the dual block-angular structure, the only one considered in [16]).

Figures 2–5 of Appendix A show the original matrix, 2-blocks dual, 2-blocks primal, and 4-blocks dual reorderings for the four largest instances tested.

4 Using the Reordered Matrices

The numerical kernel of any optimization algorithm is to deal with linear systems of equations derived from the constraints matrix. If the block bordered structure of the constraints matrix is exploited, significant savings can be obtained. This is valid for both simplex and interior-point methods, which have been extensively used for cell-suppression, controlled tabular adjustment, and controlled rounding. We will focus on the use of primal block-angular matrices in interior-point methods (which have shown to be the most efficient option for tabular data), and will test them for the controlled tabular adjustment problem.

The main computational burden of an interior-point method [21] is to solve systems of equation with matrix $A\Theta A^T$, where Θ is a diagonal positive definite matrix. For our purposes, and without loss of generality, we will assume $\Theta = I$, thus, the system to be solved is

$$(AA^T)\Delta y = g. \tag{3}$$

This system is named the "normal equations" and it is usually solved by a sparse Cholesky factorization. If A has the structure of (1), we can recast the matrix of system (3) as

$$AA^T = \left[\begin{array}{ccc|c} A_1A_1^T & & & A_1L_1^T \\ & \ddots & & \vdots \\ & & A_kA_k^T & A_kL_k^T \\ \hline L_1A_1^T & \dots & L_kA_k^T & \sum_{i=1}^k L_iL_i^T \end{array}\right] = \begin{bmatrix} B & C \\ C^T & D \end{bmatrix}, \tag{4}$$

B, C and D being the blocks of AA^T. Appropriately partitioning g and Δy in (3), the normal equations can be written as

$$\begin{bmatrix} B & C \\ C^T & D \end{bmatrix} \begin{bmatrix} \Delta y_1 \\ \Delta y_2 \end{bmatrix} = \begin{bmatrix} g_1 \\ g_2 \end{bmatrix}. \tag{5}$$

By eliminating Δy_1 from the first group of equations of (5), we obtain

$$(D - C^TB^{-1}C)\Delta y_2 = (g_2 - C^TB^{-1}g_1) \tag{6}$$

$$B\Delta y_1 = (g_1 - C\Delta y_2). \tag{7}$$

Therefore we have reduced the solution of system (3) to the solution of systems with matrix B (which is made of k smallest subsystems $A_iA_i^T, i = 1, \dots, k$) and with one system with matrix $(D - C^TB^{-1}C)$, which is named the "Schur complement" of (3). In general, if the solution of systems with $A_iA_i^T, i = 1, \dots, k$ are not too expensive, and the sparsity (the number of nonzero elements) of the Schur complement is not degraded (i.e, decreased too much), we can obtain significant computational savings by solving (6)–(7) instead of (3). Moreover, there are efficient procedures for (6) based on iterative linear solvers [8].

The above procedure, using a preconditioned conjugate gradient (i.e., iterative solver) for (6) with a specialized preconditioner (see [5,8] for details), has been implemented in an interior-point package for primal block-angular problems [8]. Such procedure can be used for the efficient solution of controlled tabular adjustment (CTA) problems, once the tabular data constraints have been previously reordered as shown in Section 3. Table 5 reports preliminary computational results with an early implementation of CTA based on the specialized interior-point algorithm of [8]. For each of the instances of Table 1—but the two largest ones, that failed with the iterative solver—we solved systems (6)–(7) from some interior-point iterations with a sparse Cholesky factorization (the standard procedure used by general interior-point solvers, such as CPLEX) and with the specialized procedure of [8]. Column "Ratio time" of Table 5 show the ratio between both solution times, i.e., how many times faster is the specialized procedure compared to the standard one. We note that the problems were not

Table 5. Ratio time for the solution of CTA by an interior-point method without and with exploitation of structure

Name	Ratio time
bts4	1.5
hier13	12.7
hier13x13x13a	11.6
hier13x13x7d	3.9
hier16	43.5
hier16x16x16a	43.2
jjtabeltest	0.7
ninenew	7.5
nine5d	2.8
nine12	5.1
targus	1.0
toy3dsarah	6.0
two5in6	10.9

solved up to optimality with the approach of [8], since that procedure has still to be tuned for problems like CTA (which has, for instance, equality linking constraints instead of the inequality ones considered in [8]). However those figures are a good indicator of the expected overall performance in the solution of CTA by exploiting the constraints structure in an interior-point method.

5 Conclusions

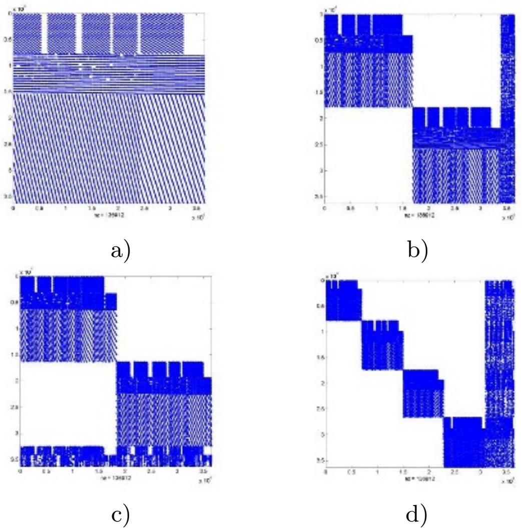
The structure detection tool used in this work for constraints of tabular data can provide significant computational savings for CTA, and, in general, for any tabular data protection method that has to solve a sequence of linear programming subproblems. Many additional tasks have still to be done. Among them, the main one is to tune the specialized approach of [8] for fully exploiting the reordered matrix, and for obtaining an optimal solution to CTA in a fraction of the time needed by a general solver.

References

1. Benders, J.F.: Partitioning procedures for solving mixed-variables programming problems. Computational Management Science **2** (2005) 3–19
2. Bixby, R.E.: Solving real-world linear programs: a decade and more of progress. Operations Research **50** (2002) 3–15
3. Bradley, S.P, Hax, A.C., Magnanti, T.L.: Applied Mathematical Programming. Addison-Wesley, Reading (1977).
4. Castro, J.: Network flows heuristics for complementary cell suppression: an empirical evaluation and extensions. Lect. Notes in Comp. Sci. **2316** (2002) 59–73. Volume Inference Control in Statistical Databases, ed. J. Domingo-Ferrer, Springer, Berlin

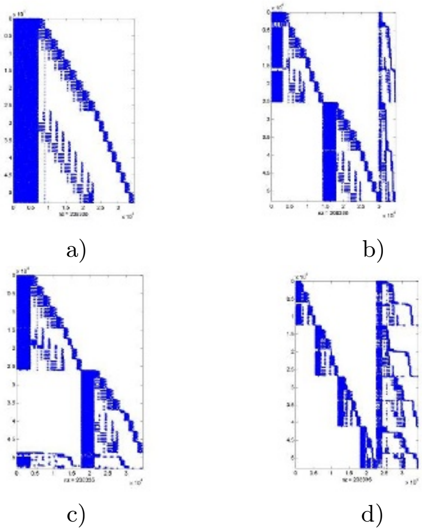
5. Castro, J.: Quadratic interior-point methods in statistical disclosure control. Computational Management Science **2** (2005) 107–121
6. Castro, J.: Minimum-distance controlled perturbation methods for large-scale tabular data protection. European Journal of Operational Research **171** (2006) 39–52
7. Castro, J.: A shortest paths heuristic for statistical disclosure control in positive tables. To appear in INFORMS Journal on Computing. Available as Research Report DR 2004/10 Dept. of Statistics and Operations Research, Universitat Politècnica de Catalunya, 2004
8. Castro, J.: An interior-point approach for primal block-angular problems. To appear in Computational Optimization and Applications (2007). Available as Research Report DR 2005/20 Dept. of Statistics and Operations Research, Universitat Politècnica de Catalunya, 2005
9. Cox, L.H.: Network models for complementary cell suppression. J. Am. Stat. Assoc. **90** (1995) 1453–1462
10. Cox, L. H., George, J. A.: Controlled rounding for tables with subtotals. Annals of Operations Research **20** (1989) 141–157
11. Dandekar, R.A.: personal communication (2005)
12. Dandekar, R.A., Cox, L.H.: Synthetic tabular Data: an alternative to complementary cell suppression, manuscript, Energy Information Administration, U.S.
13. Dantzig, G.B., Wolfe, P.: Decomposition principle for linear programs. Operations Research **8** (1960) 101–111
14. Fischetti, M., Salazar, J.J.: Solving the cell suppression problem on tabular data with linear constraints. Management Science **47** (2001) 1008–1026
15. Gondzio, J., Sarkissian, R.: Parallel interior point solver for structured linear programs. Mathematical Programming **96** (2003) 561–584
16. Hu, Y.F., Maguire, K.C.F., Blake, R.J.: A multilevel unsymmetric matrix ordering algorithm for parallel process simulation. Computers and Chemical Engineering **23** (2000) 1631–1647
17. Hundepool, A.: The CASC project. Lect. Notes in Comp. Sci. **2316** (2002) 172–180. Volume Inference Control in Statistical Databases, ed. J. Domingo-Ferrer, Springer, Berlin
18. Karypis, G., Kumar, V.: A fast and high quality multilevel scheme for partitioning irregular graphs. SIAM Journal on scientific Computing **20** (1999) 359–392.
19. Kernighan, B.W., Lin, S.: An efficient heuristic procedure for partitioning graphs. Bell Systems Technical Journal **49** (1970) 291–308
20. Salazar, J.J.: Controlled rounding and cell perturbation: statistical disclosure limitation methods for tabular data. Mathematical Programming **105** (2006) 583–603
21. Wright, S.J.: Primal-Dual Interior-Point Methods. SIAM, Philadelphia (1997).

A Sparsity Pattern of Original and Reordered Matrices



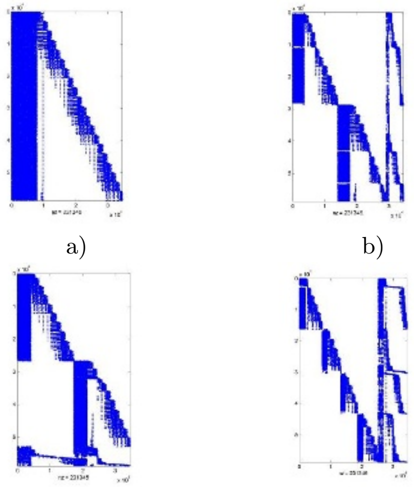
a) b) c) d)

Fig. 2. Results for instance bts4: a) original matrix; b) reordered matrix in dual 2-blocks-angular form; c) reordered matrix in primal 2-blocks-angular form; d) reordered matrix in dual 4-blocks-angular form



a) b) c) d)

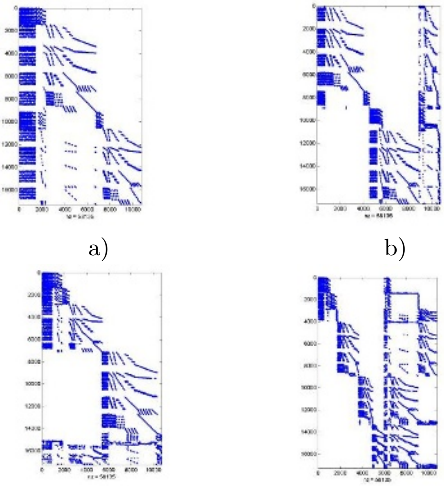
Fig. 3. Results for instance five20b: a) original matrix; b) reordered matrix in dual 2-blocks-angular form; c) reordered matrix in primal 2-blocks-angular form; d) reordered matrix in dual 4-blocks-angular form



a) b)

c) d)

Fig. 4. Results for instance five20c: a) original matrix; b) reordered matrix in dual 2-blocks-angular form; c) reordered matrix in primal 2-blocks-angular form; d) reordered matrix in dual 4-blocks-angular form



a) b)

c) d)

Fig. 5. Results for instance nine5d: a) original matrix; b) reordered matrix in dual 2-blocks-angular form; c) reordered matrix in primal 2-blocks-angular form; d) reordered matrix in dual 4-blocks-angular form

A New Approach to Round Tabular Data

Juan José Salazar González

DEIOC, Faculty of Mathematics, University of La Laguna
Av. Astrofísico Francisco Sánchez, s/n; 38271 La Laguna, Tenerife, Spain
Tel.: +34 922 318184; Fax: +34 922 318170
jjsalaza@ull.es

Abstract. Controlled Rounding is a technique to replace each cell value in a table with a multiple of a base number such that the new table satisfies the same equations as the original table. Statistical agencies prefer a solution where cell values already multiple of the base number remain unchanged, while the others are one of the two closest multiple of the base number (i.e., rounded up or rounded down). This solution is called zero-restricted rounding. Finding such a solution is a very complicated problems, and on some tables it may not exist. This paper presents a mathematical model and an algorithm to find a good-enough near-feasible solution for tables where a zero-restricted rounding is complicated. It also presents computational results showing the behavior of the proposal in practice.

1 Introduction

Statistical agencies publish information in tabular form called *tables*. A table is a collection of values located in *cells*. Some cell values are obtained by adding other value values, and are called *marginal cells*. The cells containing values which cannot be computed by adding other cell values in the table are called *internal cells*. Then, each marginal cell is associated to a mathematical equation defining its value by a sum of values in internal cells. Figure 1 shows an example of a table.

Before publication, in many cases, statistical agencies need to round each cell value to a multiple of a given number. This number is called *base number*. Standard examples of base numbers are 1, 3, 5 or 10. In particular, 1 may be used when the values in the original table are floating numbers, and the statistical agency desires to publish only integer values. This decision can be motivated for cosmetic reasons, as decimal values could be irrelevant. Another motivation for rounding values could be to limit the disclosure risk. Indeed, publishing too much details could reveal private information about the contributors of the data, which should be protected by national laws.

The difficulty of rounding values in a table is based on the existence of mathematical equations. The rounded values in marginal cells should coincide with the sum of rounded values of the associated internal cells. Depending on the table structure, this aim could lead to a complex mathematical problem, known

J. Domingo-Ferrer and L. Franconi (Eds.): PSD 2006, LNCS 4302, pp. 25–34, 2006.

Unrounded data	total	male	female	young	adult	thin	fat
North East	60593	29225	31368	13856	46737	34565	26028
North West	174414	78129	96285	25673	148741	3432	170982
Yorkshire and Humberside	108769	46119	62650	2342	106427	32223	76546
East Midlands	93346	43201	50145	23443	69903	23434	69912
West Midlands	131817	61046	70771	23878	107939	432	131385
East	107060	47376	59684	24532	82528	34233	72827
London	110811	49053	61758	17635	93176	3423	107388
South East	123359	50949	72410	34223	89136	4567	118792
South West	119863	44718	75145	35980	83883	56356	63507
England	1030032	449816	580216	201562	828470	192665	837367
Wales	95388	49579	45809	34989	60399	6454	88934
Scotland	124678	61327	63351	36789	87889	5643	119035
Great Britain	1250098	560722	689376	273340	976758	204762	1045336

Fig. 1. Number of inhabitants (dummy data)

as *Controlled Rounding Problem* (CRP). The rounded value should be as close as possible to the original (unrounded) value, but then the CRP is an optimization problem. To be more precise, the statistical offices use to require that values which are already multiple of the base number should remain unchanged; otherwise, the original value can be either rounded up or down to the closest multiple of the base number. Then the aim of CRP is to look for rounded values for all the cells such that the distance between the unrounded and the rounded tables is minimized. This distance is traditionally defined by the sum of the absolute difference between each rounded and unrounded cell value. Potentially, each term of this addition can be weighted with a different parameter to consider the importance of the cell in the tabular form. This restricted CRP problem is known as *zero-restricted CRP*. On tables with a 2-dimensional structure the zero-restricted CRP consists of solving a network flow problem on a bipartite graph, and it is always feasible. However, on more complex tables (including a 3-dimensional table with $2 \times 2 \times 2$ internal cells) a rounded table may not exists. Checking if a feasible solution exists is a NP-complete problem (see, e.g., Moravek and Vlach [9]), and therefore finding an optimal zero-restricted CRP is a very difficult combinatorial problem. Some early articles have addressed this problem; see e.g. Bacharach [1], Dalenius [6], Cox and Ernst [3], Cause, Cox and Earns [2], Fagan, Greenberg and Hemmig [7], and Cox and George [5].

To overcome the drawbacks of the zero-restricted CRP, some relaxations have been addressed. The most widely used by statistical offices is the so-called *Random Rounding* (see, e.g., Fellegi [8]). In this case, values which are multiple of the base number remain unchanged. Otherwise, each cell value is replaced by one of the two closest multiple of the base number, no matter if the output table satisfies or not the mathematical equations. The replacement is conducted by using a random procedure that takes into account the original value a_i of a cell and the two closest multiples of the base number, denoted by $\lfloor a_i \rfloor$ and $\lceil a_i \rceil$. More precisely, if the base number is β, the procedure

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

Fig. 2. Investment of enterprises by activity and region

selects $\lfloor a_i \rfloor$ with probability $(\lceil a_i \rceil - a_i)/\beta$, and $\lfloor a_i \rfloor$ with probability $(a_i - \lfloor a_i \rfloor)/\beta$. Let us denote the random value by $\xi(a_i)$. This methodology has the advantage of being easy to be applied on all type of tables. Another advantage of using random rounding is that the rounded values are, by construction, statistically unbiased. When solving the above defined CRP, due to the objective function, the probability distribution is not linearly proportional to $a_i - \lfloor a_i \rfloor$ as happens in the random rounding. This biased behavior is undesired by experts (see Cox [4]), and therefore support the use of random rounding. However, random rounding has the disadvantage of producing non-additive tables, and this is a fundamental drawback that motivates new research on CRP.

Salazar [11] has described a mathematical model and implemented an algorithm to find optimal solutions of the CRP on tables with general structures. The code has been able to find rounded tables to real-world instances with up to 1,000,000 cells. Still, it is possible to build small and artificial tables where this code cannot find even a feasible solution in a reasonable computing time. The current paper presents a new relaxed version of CRP, and discusses computational results.

In this paper we consider the CRP as above defined, without protection level requirements that make the optimization problem more difficult. This relaxation is still of interest in practice because experts in statistical office tend to accept rounded tables as protected by construction. We also discard here the concept of loss of information, which in Salazar [11] is the objective function to be minimized. A reason for discarding this concept is that a user can never measure the utility of the rounded values by the difference to the unrounded value, as he/she will never receive the unrounded value. The objective function in this paper will be used to conduct the approach to feasible solutions. More precisely, in this paper the CRP consists in choosing if an internal cell should be rounded up or down so the worst different between the rounded and the rounded values of a marginal cell is minimum. Because this combinatorial problem remains difficult, a near-optimal approach is proposed and tested.

2 The Optimization Problem

Most of the tabular data are created by initially having a set of cells containing unrelated numbers. These cells are called *internal cells*. Then, the values of some internal cells are aggregated and create the so-called *marginal cells*. Although some marginal cells can be seen as the aggregation of other marginal cells, the recursive definition shows that any marginal cell is always the sum of marginal cells. (For instance, the grand total is the sum of all the marginal cells.) In this case, each marginal cell corresponds to a mathematical equation where one cell of the table is equal to the some of other (internal) cells of the table. This is the case of, for example, the tiny table in Figure 2, where we find nine internal cells and seven marginal cells. In this simple table one could see eight mathematical equations because there are four rows and four columns, but one of the equations is redundant (it is linearly dependent of the other seven equations). Not all tables have this property, and the table in Figure 1 is an example. Indeed, the marginal cell with the value 60593 in North East is defined by three equations, and no one of the three is redundant. In other words, if forget one of the three equations defining this marginal cell during the rounding process, then the final table may not be additive. However, this type of tables where more than one equation is needed to define the same marginal cell is unusual in practice. This section sets the notation for a table with the above property and defines the rounding problem. An approach to solve this problem will be presented in the next section.

Let I be the index set of all the internal cells of a table, and J the index set of all the marginal cells. To write a mathematical model for the precise definition of the problem, we associate each cell $i \in I$ with a 0–1 variable x_i. The variable is 1 if the cell value a_i should be rounded up to $\lceil a_i \rceil_\beta$, and 0 if the cell value a_i should be rounded down to $\lfloor a_i \rfloor_\beta$. We keep the zero-restricted assumption on the internal cells, thus $x_i = 0$ when a_i is a multiple of the base number.

Regarding marginal cells, we do not associate variables. We associate constraints, instead. More precisely, if a marginal cell $j \in J$ contains a value a_j which is the sum of values related to a collection S_j of internal cells, i.e.

$$a_j = \sum_{i \in S_j} a_i$$

then an ideal CRP solution x^* for the internal cells would satisfy

$$\lfloor a_j \rfloor_\beta \leq \sum_{i \in S_j} \lfloor a_i \rfloor_\beta + \beta x_i^* \leq \lceil a_i \rceil_\beta .$$

Since such a solution may not exist, we introduce a new variable y to measure the violation of these constraints. This variable will the objective function to be minimized in our combinatorial problem, hence the aim is to find the "most feasible" CRP solution. It is possible to consider other objective functions to also select with a second criteria among all the most feasible solutions. For example, it could be the minimization of $\lambda y + \sum_{i \in I} w_i x_i$, where $\lambda := |I|\beta$ gives priority to the feasibility, and where $w_i := \lceil a_i \rceil_\beta + \lfloor a_i \rfloor_\beta - 2a_i$ measures the cost of rounding up instead of rounding down. Alternatively, as mentioned in Salazar [10], the parameter w_i could also be selected as the random value $\lceil a_i \rceil_\beta + \lfloor a_i \rfloor_\beta - 2\xi(a_i)$ to reduce the implicit bias of a deterministic procedure.

To simplify notation and without loss of generality, we can assume that $0 \leq a_i < \beta$ for all $i \in I$, and $\beta = 1$. This transformation replaces a_j by other numbers that will be denoted by a'_j, for all $j \in J$, whose closest integer values are denoted by $\lfloor a'_j \rfloor$ and $\lceil a'_j \rceil$. Then the CRP is equivalent to round fractional numbers to integer numbers, and a mathematical model is:

$$\text{minimize } |I|y + \sum_{i \in I} w_i x_i \tag{1}$$

subject to

$$\lfloor a'_j \rfloor - y \le \sum_{i \in S_j} x_i \le \lceil a'_j \rceil + y \qquad \text{for all } j \in J, \tag{2}$$

$$x_i = 0 \qquad \text{for all } i \in I : a_i = 0, \tag{3}$$

$$x_i \in \{0, 1\} \qquad \text{for all } i \in I : a_i > 0. \tag{4}$$

A solution with $y = 0$ defines a zero-restricted CRP output. Some statistical agencies would prefer to replace constraint (3) by $x_i \in \{0, 1\}$ even when $a_i = 0$ to increase the probability of finding a CRP solution x with $y = 0$. The model corresponds to an $\mathcal{NP}$-hard optimization problem because the particular case with $a'_j = 1$ for all $j \in J$ and appropriate values w_i can be used to solve instances of the *Set Partitioning Problem.*

Note that the problem presented in this section is different to the problem addressed in [11]. Both problems are related with the rounding methodology, but they do not aim at finding the same solutions. The problem in [11] looks for a rounded table taking into account protection level requirements and minimizing the loss of information. It is a complex combinatorial problem, not only in theory but also in practice. Indeed, for highly structured tables even finding a feasible solution may be difficult. As an alternative, this paper present a different problem where the protection levels are not considered, and where the aim is to provide a "good" rounded table. The goodness of the solution is based on ensure that internal cells are rounded either up or down the original values; then, the impact on the marginal cells is minimized.

The fact that the problem presented in this paper does not consider protection levels seems to be reasonable as the methodology is oriented to solve complex instances where the full problem presented in [11] does not work at all. These instances tend to be large and highly structured tables, and the first aim of the statistical agencies under this situation is to have at least a "good" rounded table.

3 Algorithms

Although the theoretical complexity of the problem introduced in the previous section is still $\mathcal{NP}$-hard, as the problem addressed in [11], it is in practice much simple to be solved. To support this claim, note that finding a feasible solution for the problem introduced in this paper is trivial: fixing the internal cells first, and then recomputing the marginal cells with the table structure, will always give a feasible solution. To this end y in (2) must be calculated as the last step, and the quality of the solutions depends directly on the value of y. Clearly, bad decisions when rounding the internal cells may lead to large values of y, and therefore bad solutions. However, the additivity of the solution is ensured in all instances where the property that each marginal cell is related only one mathematical equation hold. The same simple heuristic procedure cannot be applied to the problem in [11], and indeed finding a feasible solution is $\mathcal{NP}$-complete.

To solve the 0-1 Integer Linear Programming model (1)–(4) we use a branch-and-bound algorithm. At each node of the branch-and-bound search, the bound is computed by solving the Linear Programming relaxation of the model. When the solution is not integer, a branching is performed by fixing a fractional variable either down or up. This means that an internal cell is either decided to be rounded down or up. The optimal objective value of the Linear Programming relaxation is a lower bound of the objective value of the rounded tables with such decision on this internal cell. For that reason, the smallest lower bound is a global lower bound LB on the optimal integer solution. To get an upper bound, each fractional solution x^* is used to heuristically create an integer solution. To this end, a decision is taken by rounding up each internal cell i with probability x_i^*, and then the marginal cell are computed by using the equations defining the table. The largest difference between the rounded and unrounded marginal values is defined as the upper bound, and the smallest upper bound is saved on UB. The branch-and-bound search stops when $LB \geq UB$ or when the time limit is achieved. Further technical details on this algorithm are not explained due to the page limitation of this paper.

4 Computational Experiments

As mentioned in the introduction, the new proposal was motivated by the result of using the controlled rounding approach in [11] to some instances at the Office of National Statistics (U.K.). Table 1 shows the number of cells, the number of equations, the existence of a hierarchical variable defining the table and the type of final solution. When the solution type is ”Feasible” it means that the procedure was interrupted due to the time limit (10 minutes) and a zero-restricted solution was available. When the solution type is ”infeasible” it means that the procedure was interrupted due to the time limit and no zero-restricted solution was available.

To test the model and algorithm described in this paper we have implemented the approach in C programming language on a notebook Pentium Centrino 1.7 Ghz. This implementation is compared with the code implementation described in [11]. A fundamental difference between the two implementations is that the code in [11] considers protection levels and looks for a zero-restricted feasible solution minimizing the loss of information. Instead, the approach presented in this paper does not take into account protection levels in the set of constraints, and the loss of information is a secondary criterion in the objective function. The requirement of the approach presented in this work is the additivity and the fact that internal cells must be rounded either up or down. The main objective function of the approach is to minimize the worst distance between rounded and unrounded marginal values.

Table 1. Experiments using the code described in [11]

	dim	cells	equa	hierar	S-Type
cutTwoDimlarge50_AllGors_LSoa(100K)	2	102051	2052	No	Optimal
cutTwoDimlarge50_AllGors_LSoa(200K)	2	204051	4052	No	Optimal
cutTwoDimlarge50_AllGors_LSoa(400K)	2	408051	8052	No	Optimal
cutTwoDimlarge50_AllGors_LSoa(450K)	2	459051	9052	No	Optimal
cutTwoDimlarge50_AllGors_LSoa(500K)	2	510051	10052	No	Optimal
TwoDimLarge100_GorG_LSoa	2	437532	83314	Yes	Optimal
TwoDimLarge100_GorW_LSoa	2	118932	24568	Yes	Optimal
DWP_LSoas_GorA	3	46396	18255	No	Optimal
DWP_LSoas_GorB	3	124880	49088	No	Optimal
DWP_LSoas_GorG	3	99428	39089	No	Optimal
DWP_LSoas_GorH	3	133448	52454	No	Optimal
DWP_LSoas_GorJ	3	148960	58548	No	Optimal
cutDWP_LSoas_GorA	3	38922	15318	No	Optimal
cutDWP_LSoas_GorG	3	38922	15318	No	Optimal
DWP_LSoas_GorJ-hiera	3	181804	104295	Yes	Feasible
DWP_LSoas_GorW-hiera	3	65296	37860	Yes	Feasible
DWP_Oas_GorA-hiera	3	297388	173447	Yes	infeasible
DWP_Oas_GorB-hiera	3	787780	461385	Yes	infeasible

Table 1 shows the results of running the code described in [11] on a collection of benchmark instances from different statistical agencies. The base number for these experiments was 3. Column "dim" shows the dimension of the table, column "cells" shows the number of (internal and marginal) cells, "equa" is the number of marginal cells, "hierar" tells is the table has or not hierarchical structure, and "S-Type" is the type of solution found by the code described in [11]. Although 2-dimensional tables have always integer solutions, depending on the size of the table, an implementation may find or not a solution in a given time limit. An interesting observation from Table 1 is that the approach was able to find an optimal zero-restricted solution for all 2-dimensional tables, and for all the 3-dimensional tables without hierarchical variables. Only on two large tables the approach did not find a feasible zero-restricted table. Finding alternative and good rounded tables for these instances was the motivation of this paper.

Table 2 shows the results of running the code solving the mathematical model (1)–(4). From this table, one observe that the new implementation was able to find optimal zero-restricted solutions on three of the four unsolved instances. Only the last table remains unsolved in a time limit of 300,000 seconds, but the procedure found a rounded table where a modified cell value differs in at most 5 times the base number respect to the unrounded cell value. This is a very good solution for this table, as by simply rounding down the original internal cells in this table, a marginal cell value goes to 135,120 times the base number far from the original value. By rounding up all the original internal cells, a marginal cell value goes to 269,515 times the base number far from the unrounded value. Rounding each internal cell to the closest multiple of the base number in this

Table 2. Times on a Notebook Pentium Centrino 1.7 Ghz

name	time	nodes	UB
cutTwoDimlarge50_AllGors_LSoa(100K)	2.7	1	0
cutTwoDimlarge50_AllGors_LSoa(200K)	7.5	1	0
cutTwoDimlarge50_AllGors_LSoa(400K)	15.9	1	0
cutTwoDimlarge50_AllGors_LSoa(450K)	20.8	1	0
cutTwoDimlarge50_AllGors_LSoa(500K)	23.5	1	0
TwoDimLarge100_GorG_LSoa	39.5	1	0
TwoDimLarge50_GorW_LSoa	6.1	1	0
DWP_LSoas_GorA	8.2	1	0
DWP_LSoas_GorB	23.4	1	0
DWP_LSoas_GorG	33.7	1	0
DWP_LSoas_GorH	26.7	1	0
DWP_LSoas_GorJ	57.0	1	0
cutDWP_LSoas_GorA	6.0	1	0
cutDWP_LSoas_GorG	9.6	1	0
DWP_LSoas_GorJ-hiera	48700.5	280	0
DWP_LSoas_GorW-hiera	3655.8	119	0
DWP_Oas_GorA-hiera	206740.7	370	0
DWP_Oas_GorB-hiera	300000.0	1+	5

table will move a marginal cell value to 1,148 times the base number. Therefore, having a way of rounding the internal cells such that the worst marginal differs in at most 5 times the base number is a very good solution for this instance.

To further test the new approach, we have generated random instances. They are k-dimensional tables, and the base number has been fixed to $\beta = 5$. To control the effect of the density of a table, we have used a parameter $\delta \in \{25, 50, 75, 100\}$. A cell value is fixed to 0 with probability $1 - \delta/100$, and it is a random integer number in $[1, \beta - 1]$ with probability $\delta/100$. For each $k \in \{2, 3, 4, 5, 6, 7, 8\}$ we have created 10 tables. Table 3 summarizes some findings from our experiments, and each line gives the average result over the 10 generated tables. The column *type* gives the number of internal cells and the structure of the table (thus also the number of equations). The time are given on seconds on a notebook with a Pentium Centrino 1.7 Ghz., and the time limit was set to one hour. The column UB shows the value of y of the best found solution, and the column LB shows the value of the best LP-relaxation. All the instances associated to a raw of the table were optimally solved when both bounds coincides. A value 0 in these bounds means that an optimal zero-restricted solution was found. When the LB bound is 1, it means that a zero-restricted rounding has been proved to do not exist.

When considering 2-dimensional tables, the new approach trivially solves the optimization problem. Indeed, a zero-restricted solution always exists, and it can be found by solving the first linear-programming model with a network flow algorithm. Using the general purpose simplex algorithm available in CPLEX, it is done in less than 10 minutes when there are 1000 rows and 1000 columns. The difficulties of the problem appears on k-dimensional tables with $k > 2$. For instances, on average, the code cannot find in 1 hour an optimal zero-restricted

Table 3. Times on a Notebook Pentium Centrino 1.7 Ghz using base 5

type	δ	time	UB	LB
2000x1000	25	146	0	0
2000x1000	50	440	0	0
2000x1000	75	980	0	0
2000x1000	100	2047	0	0
30x30x30	25	340	0	0
30x30x30	50	1060	0	0
30x30x30	75	1646	0	0
30x30x30	100	2538	0	0
3x3x3x3x3	25	1	1	1
3x3x3x3x3	50	2	1	1
3x3x3x3x3	75	6	1	1
3x3x3x3x3	100	3600	1	0
2^8	25	1	1	1
2^8	50	10	1	1
2^8	75	49	1	1
2^8	100	288	1	1
3^6x2^2	25	3600	3	1
3^6x2^2	50	3600	4	0
3^6x2^2	75	3600	12	0
3^6x2^2	100	3600	15	0

table for a 8-dimensional table with structure 3^6x2^2 internal cells, all of them not multiple of the base number. The best that it can find is a solution where a marginal cell differs respect to the original cell in 15 times the base number. This is a good solution if one observes that rounding up (or down) all internal cells gives a marginal that differs in about 1,450 times the base number; rounding each internal cell to the closest multiple creates a marginal cell which differs from its unrounded value in about 30 times the base number. Very unfortunately, the lower bound remains zero for most of the instances, which means that the existence of a zero-restricted solution is not discard.

5 Conclusion

This paper has proposed a different mathematical model to find rounded tables on complex and large tables. It decides where to round each internal cell in order to reduce the impact on the marginal totals. An advantage of this approach is that it keeps the control on the internal cells, so all of them are either rounded up or down. A branch-and-bound algorithm has been implemented to solve the mathematical model, and it has been tested on benchmark and randomly-generated instances. The preliminary results show that the new code can find better solutions on larger tables than previous approaches. This work opens new research questions as finding new lower bounds to reduce the final gap when the problem is unsolved.

Acknowledgements

Work partially supported by "Ministerio de Educación y Ciencia", TIC2003-05982-C05-02.

References

1. M. Bacharach. Matrix rounding problem. *Management Science*, 9:732–742, 1966.
2. B. D. Causey, L. H. Cox, and L. R. Ernst. Applications of transportation theory to statistical problems. *Journal of the American Statistical Association*, 80:903–909, 1985.
3. L. H. Cox. Controlled rounding. *INFOR*, 20:423–432, 1982.
4. L. H. Cox. A constructive procedure for unbiased controlled rounding. *Journal of the American Statistical Association*, 82:520–524, 1987.
5. L.H. Cox and J.A. George. Controlled rounding for tables with subtotals. *Annals of Operations Research*, pages 141–157, 1989.
6. T. Dalenius. A simple procedure for controlled rounding. *Statistisk Tidskrift*, 3:202–208, 1981.
7. J.T. Fagan, B.V. Greenberg, and R. Hemmig. Controlled rounding of three dimensional tables. Technical report, Bureau of the Census, SRD/RR-88/02, 1988.
8. I. P. Fellegi. Controlled random rounding. *Survey Methodology*, pages 123–133, 1975.
9. J. Moravek and M. Vlach. On necessary conditions for the existence of the solution to the multi-index transportation problem. *Operations Research*, pages 542–545, 1967.
10. J. J. Salazar. A unified mathematical programming framework for different statistical disclosure limitation methods. *Operations Research*, 53(3):819–829, 2005.
11. J. J. Salazar. Controlled rounding and cell perturbation: Statistical disclosure limitation methods for tabular data. *Mathematical Programming*, 105(2–3):251–274, 2006.

Harmonizing Table Protection: Results of a Study

Sarah Giessing and Stefan Dittrich

Federal Statistical Office of Germany,
65180 Wiesbaden
{Sarah.Giessing, Stefan.Dittrich}@destatis.de

Abstract. The paper reports results of a study aimed at the development of recommendations for harmonization of table protection in the German statistical system. We compare the performance of a selection of algorithms for secondary cell suppression under four different models for co-ordination of cell suppressions across agencies of a distributed system for official statistics, like the German or the European statistical system. For the special case of decentralized across-agency co-ordination as used in the European Statistical System, the paper also suggests a strategy to protect the data on the top level of the regional breakdown by perturbative methods rather than cell suppression.

1 Introduction

Some cells of the tabulations released by official statistics contain information that chiefly relates to single, or very few respondents. In the case of establishment data, given the meta information provided along with the cell values (typically: industry, geography, size classes), those respondents could be easily identifiable. Therefore, traditionally, statistical agencies suppress a part of the data, hiding some table cells from publication. Efficient algorithms for cell suppression are offered f.i. by the software package τ-ARGUS [10].

When tables are linked through simple linear constraints, cell suppressions must obviously be coordinated between tables. This paper addresses problems that are connected to a situation where those tables are provided by different statistical agencies, as in the case of the European statistical system, and also in the case of the German statistical system.

We give an account on a study aimed at the development of recommendations for harmonization of table protection in the German statistical system, considering turnover tax statistics as a first, pilot instance. The study compared the performance of a selection of algorithms for secondary cell suppression under four different models for co-ordination of cell suppressions across the agencies. For a detailed report on the study see [9].

One of those models, the standard model adopted by Eurostat, leads to extremely many suppressions at the Union level. The situation might be improved substantially by replacing cell suppression on that level of aggregation by interval publication. We report on results of an experiment to protect data on the top level of the breakdown by region by publishing intervals as alternative to cell suppressions.

J. Domingo-Ferrer and L. Franconi (Eds.): PSD 2006, LNCS 4302, pp. 35 – 47, 2006.

2 Methodological Background

Cell suppression comprises two steps: In the first step 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 (p% rule). When cell suppression is used as disclosure limitation technique, 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 a 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. [5], 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 contribution that are too close.

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.

In section 2.1 we define relaxed standards for the concept of validity of a suppression pattern. In our study we considered only algorithms for secondary cell suppression that ensure validity of the suppression pattern at least with respect to the relaxed standards. Section 2.2 briefly describes those algorithms.

In a situation, where different agencies publish tables providing aggregate information of the same underlying data, these agencies must agree on parts of the suppression pattern. Section 2.3 outlines four different models for co-ordination of cell suppressions across tables provided by one Union or Federal level agency and a number of agencies on the level below.

2.1 Standards for Protection Requirements in Practice

For a table with hierarchical substructure, feasibility intervals computed on basis of the set of equations for the full table tend to be much closer than those computed on basis of separate sets of equations corresponding to sub-tables without any substructure. Moreover, making use of the additional knowledge of a respondent, who is the single respondent to a cell (a so called '*singleton*'), it is possible to derive intervals that are much closer than without this knowledge. Based on the assumption of a simple but not unlikely disclosure scenario where intruders will deduce feasibility intervals separately for each sub-table, rather than taking the effort to consider the full table, and that single respondents are more likely to reveal suppressed cell values within the same row or column using their special additional knowledge, we think the following minimum protection standards make sense.

(PS1) *Protection against exact disclosure*: With a valid suppression pattern, it is not possible to disclose the value of a sensitive cell exactly, if no additional knowledge (like that of a singleton) is considered, and if subsets of table equations are considered separately.

(PS2) *Protection against singleton disclosure*: A suppression pattern, with only two suppressed cells within a row (or column) of a table is not valid, if each of the two corresponds to a single respondent who are not identical to each other.

(PS1*) extension of (PS1) for inferential (instead of exact) disclosure,

(PS2*) extension of (PS2), covering the more general case where a single respondent can disclose another single respondent cell, not necessarily located within the same row (or column) of the table.

Note, that we define these protection standards in order to be able to classify current state-of-the-art cell suppression software with respect to the worst cases of disclosure risk, not because we think it would generally be enough to only consider those most likely disclosure risk scenarios.

2.2 Algorithms for Secondary Cell Suppression

The software package τ-ARGUS offers four algorithms for secondary cell suppression: Hypercube, Modular, Network and Optimal. In the current state of development, Network and Optimal tend to provide results violating the PS2 standard, i.e. they do not protect against singleton disclosure. Those algorithms have therefore been excluded from the study which provides results for Modular and for 3 variants of Hypercube. Both, Modular and Hypercube subdivide hierarchical tables into sets of linked, unstructured tables. The cell suppression problem is solved for each subtable separately. Secondary suppressions are co-ordinated for overlapping subtables. While for Modular [4], methods based on Fischetti/Salazar Linear Optimization tools [5] are used to select secondary suppressions, Hypercube is based on a hypercube heuristic [11] (see also [6]). Note, that use of the optimization tools requires a license for additional commercial software (LP-solver), whereas use of the hypercube method is for free. While Modular is only available for the Windows platform, of Hypercube there are also versions for Unix, IBM (OS 390) and SNI (BS2000).

Both, Modular and Hypercube provide sufficient protection according to standards PS1* (protection against inferential disclosure) and PS2 above. Regarding singleton disclosure, Hypercube even satisfies the extended standard PS2*. However, simplifications of the heuristic approach of Hypercube cause a certain tendency for over-suppression. The study therefore included a relaxed variant, referred to in the following as HyperSeq, where protection standard PS1* is slightly reduced for some of the subtables of a hierarchical table to the extent even of reduction to PS1 for some subtables. HyperSeq processes a table by processing a certain sequence of subtables as explained in [7, section 3]. The method is targeted at avoiding secondary cell suppressions on the higher levels of a table to some degree. A much simpler approach to reduce the tendency for over-suppression for the Hypercube method is to reduce the PS1* standard generally, for all subtables, to PS1, i.e. zero protection against inferential disclosure. We therefore refer to this method as Hyper0 in the following. Hyper0 processing can be achieved simply by inactivating the option "Protection against inferential disclosure required" when running Hypercube from out of τ-ARGUS.

Table 1. Algorithms for secondary cell suppression considered in the German harmonization study

Algorithm		**Modular**	**Hypercube**	**HyperSeq**	**Hyper0**
Procedure for secondary suppression		**Fischetti/Salazar optimization**	**Hypercube heuristic**		
Protection standard	**Interval disclosure**	PS1*	PS1*	PS1	PS1
	Singleton disclosure	PS2	PS2*	PS2*	PS2*

2.3 Approach for Co-ordination of Cell Suppressions Across Agencies

When different agencies publish tables providing aggregate information of the same underlying data, these agencies must agree on the suppression pattern for certain parts of those tables, in order to avoid that cells suppressed in the publication by one agency can be recalculated using the publications of the other agencies, referred to as '*disclosure across publications*' in the following. When there is a Federal or Union level agency and a number of agencies on the level below (the state level), we imagine basically two opposite approaches:

- Secondary suppressions are selected in a decentralized fashion: Agencies on the lower level select secondary suppressions independently. Secondary suppressions that are necessary to avoid disclosure across publications are allowed only on the federal (or Union, resp.) level.
- Secondary suppressions are selected in a centralized fashion. Disclosure across publications is avoided mostly by secondary suppressions at the state level.

The obvious advantage of a decentralized approach, which is the standard model of cooperation between Eurostat and the member states, is that state agencies can act (and publish) independently. The disadvantage is that it leads to rather much suppression at the Union level.

The German harmonization study has produced empirical results for tabulations of business turnover data by NACE (down to 5-digit level) and Region. We denote by Region^F the breakdown of the variable Region including the Federal level, and by Region^S a breakdown of the variable up to the state level only.

For empirical testing with tabulations of business turnover data the German harmonization study distinguished the following 4 variants for co-operation across agencies:

- **Centralized** (*central*)**:** Application of cell suppression to the tabulation by NACE and Region^{F}.
- **Decentralized** (*decentral*)**:** Independent application of cell suppression to the tabulation by NACE and Region^{S} for each state in a first step. In a second step, application of secondary suppression to the Federal level tabulation by NACE as to avoid the possibility of disclosure across publications.
- **Decentralized, weighted** (*dec-w*)**:** In a preparation step, we apply cell suppression to the tabulation of previous period data by NACE and Region^{F} . In the following independent application of cell suppression to the tabulation by NACE and Region^{S} making use of weighting options of the software, cells selected as secondary suppressions in the preparation step are preferred as secondary suppression. The idea of the approach is that by this procedure the final application of secondary suppression to the Federal level tabulation by NACE to avoid the possibility of disclosure across publications will lead to less suppressions on the Federal level, as compared to the simple decentralized approach.
- **Two blocks** (*block*)**:** States are grouped into two blocks: for each state of the first block, independent application of cell suppression to the tabulation by NACE and Region^{S}. Application of cell suppression to the tabulation by NACE and Region^{C} , where Region^{C} denotes the breakdown of the variable Region for the states of the second block including a new (artificial) Central level on top. Finally, application of secondary suppression to the Federal level tabulation by NACE to avoid the possibility of disclosure across publications. For efficiency, only those (four) states which did not cause any suppression on the 4-digit NACE level and above in the final step of the decentralized approach were admitted to the decentralized block.

3 Empirical Results

Within the study, any of the co-ordination models of 2.3 above has been tested in combination with one or more of the algorithms for secondary cell suppression presented in table 1. Our test data sets are tabulations of the variable ‘turnover’ of the ca. 2.9 mio records of the German turnover tax statistics by variables NACE, and Region. The NACE classification for the tabulations involves 1650[1] codes within a 7-level hierarchical structure. In the first phase of the study, for Region we used variable Region-1 involving 496 codes at a 4-level hierarchy down to the district level, while in a second phase the testing of the more promising approaches was extended using the detailed variable Region-2 down to the community level involving a total of 5412 codes.

In section 3.1 we report results of the study comparing the loss of information caused by secondary suppression, while section 3.2 provides results of the disclosure risk assessment carried out for some of the protected tables.

[1] 1133 after removing identities.

3.1 Information Loss

This section compares performances of the algorithms with respect to number and added values of the secondary suppressions, considering especially the hierarchical level of the suppressed entries. Suppressions on high levels of the table are considered undesirable.

Tabulations down to the district level

Table 2 below reports the number of secondary suppressions on the federal and state level and the percentage of the values of the secondary suppressions. Obviously, results differ considerably between methods. On the state level, the range is between 1314 and 3033, and on the federal level between 7 and 395. On the federal level, the result depends largely on the choice of the approach for across agency co-ordination (decentralized vs. centralized or two blocks approach). Decentralized approaches always caused a huge number of suppressions on the federal level. A first test with weights (the weighted decentralized method) to avoid some of the suppression on the federal level turned out to be not promising at all, therefore the approach was dropped from any further testing – although through more elaborate weighting schemes some improvement might eventually have been reached.

Table 2. Number and Total Values of Secondary Suppressions on Federal and State Level

Methodology		Secondary Suppressions				
		Number				Value
Co-ordination approach	Algorithm f. Sec.Supp.	State Level abs.	%	Federal Level abs.	%	(overall) %
decentral	Modular	1,314	7.8	189	16.7	5.17
dec-w		1,318	7.8	187	16.5	5.32
decentral	Hyper0	1,907	11.4	255	22.5	4.88
	Hypercube	2,285	13.6	395	34.9	10.23
	HyperSeq	1,938	11.5	193	17	6.95
central	Modular	1,675	10	7	0.6	2.73
	Hyper0	2,369	14.1	8	0.7	2.40
	Hypercube	2,930	17.4	22	1.9	5.69
	HyperSeq	3,033	18.1	9	0.8	7.42
block	Modular	1,621	9.7	10	0.9	2.65
	HyperSeq	3,016	18	7	0.6	6.13

On the state level, on the other hand, results are affected much stronger by the choice of the algorithm for secondary cell suppression, although of course with the same algorithm for secondary suppression there is always less suppression in the result of the decentralized variant. Best results are achieved by the modular method, while using a variant of the hypercube method leads to an increase in the percentage of secondary suppressions between about 4 (Hyper0) and 8 (HyperSeq). Note that, even in the centralized variant, the number of suppressions selected by Modular is still smaller than the number of suppressions selected by any variant of Hypercube with a decentralized co-ordination approach.

The same tendencies were observed in a more detailed analysis, c.f. table 3 of the appendix: Methods leading to more suppression overall also tend to lead to more suppression on the high levels of the NACE classification. Modular is the best performer, also on the lower regional levels. On these levels, choice of the approach for across agency co-ordination does not matter much.

For a comparison between the variants of across agency co-ordination, a look at the percentages of the values of the secondary suppressions presented in the rightmost column of table 2 is quite interesting. According to this analysis, the centralized approach generally preserves more information, overall, than the decentralized approach except for the HyperSeq algorithm. Apart from the result for Hyper0 in the centralized variant (2.4%) which is certainly affected by the low protection standard provided by this method, allowing for relatively small values to be picked as secondary suppressions, best results (about 2.7 %) are achieved by Modular in both, the two blocks and centralized variant, while Hypercube in the decentralized variant performs worst (10.2 %).

The conclusion from the results observed for the tabulations down to the department level was, neither to carry out any further experiments with the hypercube method in the variant that protects against inferential disclosure, nor with HyperSeq, because of the poor performance (on the state level and below, in case of HyperSeq) of these methods with respect to information loss. Although it was clearly only second best performer, testing with Hyper0 was to be continued, because of technical and cost advantages mentioned in 2.2 .

Tabulations down to the community level

Probably because of hardware restrictions, a run of Modular on the ca. 9 mio cells tabulation for Germany by NACE and Region-2 could not be carried out successfully, while a run of that table restricted to the 12 states or the block with centralized processing in the two blocks approach was no problem. Application of Hyper0 to the full table was no problem, either.

Using specific control options of Modular (so called ‘frozen cells’), a splitting of the application proved to be feasible: it turned out to be possible to process Region-2 tables for single states, while forcing the secondary suppressions on the state level to be identical to the secondary suppressions of the Region-1 table resulting from the centralized co-ordination approach. Such a partitioning approach is not possible for Hyper0.

Note, that the additional detail in the table caused an increase in suppression at the district level of about 70 - 80 % with Hyper0 and about 10 – 30 % with Modular.

3.2 Disclosure Risk Assessment

In order to assess the disclosure risk of a protected table, feasibility intervals were computed for all the sensitive cells (i.e. the primary suppressions). There are several options to compute those intervals, depending on the disclosure scenario that is considered. It makes a big difference, if we consider the full set of table equations 'simultaneously', or, in correspondence to the approach taken by the secondary cell suppression algorithms (c.f. 2.2), if we consider only a subset of table equations related to a subtable without hierarchical substructure, at a time. What also matters is the a priori information taken into account.

In the study, the experiments were generally based on the assumption that (external) users consider a lower bound of zero for all cell values and no upper bounds.

In principle, there is a risk of (inferential disclosure), when the bounds of the feasibility interval of a sensitive cell could be used to deduce bounds on an individual respondent contribution that are too close according to the method employed for primary disclosure risk assessment.

With this definition, and considering the full set of table equations we found between about 4 % (protection by Modular) and about 6 % (protection by Hyper0) of sensitive cells at risk looking at a tabulation by NACE and state[2], and at two more detailed tables (down to the district level) analyzed for two of the states. Note that, even if there is a risk of inferential disclosure for a cell (or, for respondents to that cell, in fact), it is rather a risk potential, not comparable to the risk for those respondents, if the cell, or the intervals were published. After all, the effort connected to the computation of feasibility intervals based on the full set of table equations is quite considerable, and moreover, in a part of the disclosure risk cases only insiders (other respondents) would actually be able to disclose an individual contribution.

The risk potential is certainly much higher for cells found to be at risk when only subsets of equations related to subtables without hierarchical substructure are considered for the interval computation because the effort for this kind of analysis is much lower. Because Modular protects according to PS1* standard, there are no cells at risk in tables protected by Modular. For Hyper0, we found about 0.4 % cells at risk, when computing feasibility intervals for equation subsets of a detailed tabulation.

In this tabulation, about 66 % of the sensitive cells are singletons. When protected by Modular, for about 0.08 % of those singletons, it turned out that another singleton would be able to disclose its contribution. Because Hyper0 satisfies PS2* standard, no such case was found, when the table was protected by Hyper0.

4 Protection Through Interval Publication

Experience by Eurostat, and also the results of the German study presented in section 3 prove that a decentralized approach leads to much suppression on the Union (or Federal) level. In a document prepared for the meeting of the steering group "Structural Business Statistics"[DOC 4.1, section 2a] in November 2005, Eurostat has

[2] Use of decentralized across agency co-ordination approach increases the number of cells at risk for Modular by 80 %.

proposed to modify European level data, protecting the data by slight adjustment, rather than to suppress them. The idea bears some resemblance to Controlled Tabular Adjustment (CTA) suggested for instance in [3], [1], [2], and [8]. CTA methods attempt to find the closest set of adjusted cell values that make the released table safe.

In the following, as another alternative, or rather, supplement, we discuss an approach to replace cell suppression by publication of intervals. A suitable, exact methodology for interval publication a.k.a. '*Partial Suppression*' has been suggested in [12]. Unfortunately, a software implementation of that methodology is not yet available. So, instead of an exact methodology, we use a simple heuristic approach applied to the federal level data of our tabulation by NACE and state. The general idea of the method is, to publish the original cell value of those cells that were not subject to adjustment in a first CTA step, while replacing in the publication the cell values of the other cells by the adjusted value, released together with an interval. Intervals should be small enough, on one hand, to be of interest to users of the publication, but large enough, on the other hand, to provide sufficient protection to individual respondent data contributing to sensitive cells at both, the federal, and the state level.

In a first step, the CTA method of [8] was applied to the federal level data, considering as sensitive cells those NACE positions, where either the federal level cell itself is sensitive, or which are suppressed in one, and only one, of the state publications[3], protected by Modular in decentralized fashion. NACE positions, where all the state positions are unsuppressed, were made ineligible for adjustment. Note, that this simple heuristic CTA method tends to provide results where some of the sensitive cells are not sufficiently adjusted. For our instance, it required in fact some post processing of the CTA result to get to a feasible solution. Table 4 of the appendix presents the number of cells by ranges of relative deviation and by level of NACE classification. Shaded cells of this table mark combinations of NACE level and range, where we consider the range as so wide that from the information loss point of view this cell is considered as lost, just as if it had been suppressed. The total of the frequencies presented in those shaded cells is 11, which is much less than the 189 secondary suppressions selected by Modular in the decentralized co-ordination approach. So our conclusion is that CTA has actually reduced information loss enormously compared to decentralized cell suppression.

In a second step, we compute intervals in such a way that

(1) the interval contains both, original, and adjusted value,
(2) both bounds are multiples of 10 raised to the same power, powers of ten serving as a kind of a flexible rounding base, computed as $\log_{10}$ of the distance between original and adjusted value,
(3) in about half of the cases, one of the interval bounds is the multiple of the corresponding rounding base closest to the original value.

Condition (1) ensures that the intervals satisfy the protection requirements imposed on the federal level data. Because, due to condition (2), the interval bounds are rounded figures, users of the data get an immediate, rough notion of the precision of

[3] In a real (production) application, of course some more positions have to be considered as sensitive, e.g. those, where the total of the suppressed state cells is sensitive.

the adjusted value, even if they don't care to look at the exact interval size. Finally, by forcing the intervals to be wider with some probability than 'necessary' according to (1) and (2), we avoid that intruders can use the information given by (1) and (2) to compute (with certainty) intervals for the original value that are closer than the distance between original and adjusted value.

The question is now, how much protection does this kind of interval publication on the federal (or European level) give to the state (or member state) level data? Considering the full set of table equations we found between about 4 % of sensitive cells at risk considering the tabulation by NACE and state, even when we consider *all* cells at the Federal level suppressed. Cell suppression with decentralized across agency co-ordination leads to an increase in this rate to about 7 %. When we computed the feasibility intervals considering those cells on the Federal level as suppressed that were subject to an adjustment, adding the intervals computed for publication to the set of constraints, we found about 7.5 % cells lacking sufficient protection. One might say, if agencies can put up with 7 % of under protected cells caused by cell suppression, they might as well be willing to put up with another 0.5 % cells at risk.

Alternatively, we suggest the following strategy for deciding on the publication of intervals according to PS1* standard accounting for the increase in risk potential for cells at risk caused by the release of intervals: In a first step, we audit the subtables separately, adding the corresponding intervals for publication to the set of constraints for each subtable. This step is followed by a final cell suppression step for the Federal level data, where those NACE positions, where a cell (from either the Federal or the state level) has been found to be under-protected in one of the subtables (we regard those cells as cells with a high risk potential for approximate disclosure), are considered as primary suppressions. In the final publication we would neither publish intervals nor adjusted values for cells suppressed in this final cell suppression step. At the time of writing, unfortunately we do not yet have any empirical experience with this strategy.

5 Summary and Final Conclusions

The paper has reported on results of a study aimed at the development of recommendations for harmonization of table protection in the German statistical system. The study has compared the performance of a selection of algorithms for secondary cell suppression (e.g. τ-ARGUS Modular, and three variants of the Hypercube method) under four different models for co-ordination of cell suppressions across the agencies. We considered only cell suppression algorithms that satisfy certain conditions related to a minimum protection standard defined in section 2.1.

Regarding the overall loss of information due to secondary suppression the Modular method of τ-ARGUS in the centralized approach for across agency co-operation gives the best results. It is, however, interesting to observe that also our 'two-blocks' approach, where centralized processing of secondary suppression is requested not for all states, but only for a sensibly defined block of some states, gave results that are not much different in quality.

Of the variants of the Hypercube method only one gave results that might be of acceptable quality. Auditing the results proved that the rate of cells with a (low) potential risk of inferential disclosure is about 50 % larger for tables protected by this variant of the Hypercube method, compared to the risk rates of tables protected by the modular optimization method. In a table protected by this variant of the Hypercube method, we also found a few cells (about 0.4 %) at a fairly high potential risk of inferential disclosure.

The conclusion from the study carried out so far is, first to await a decision for either of the approaches suggested for across-agency co-ordination of secondary suppressions, which we perceive to be a matter of agency policies in the first way. For the approach finally agreed upon by agency policy makers, the testing of the algorithms for secondary cell suppression might then be continued, with a focus on problems connected to the protection of higher dimensional and linked tables.

For the special case of decentralized across-agency co-ordination used by Eurostat and the EU member states, the paper has also suggested a strategy to protect the data on the top level of the regional breakdown (e.g. EU level data, or in the German case: Federal level data) by a mix of controlled tabular adjustment, interval publication, and cell suppression. The paper has presented empirical results at least for a part of the methodology suggested. In order to make it applicable for production purposes, however, a more advanced implementation of the software used for controlled tabular adjustment and auditing would be required.

References

1. Castro, J. (2003), ‘Minimum-Distance Controlled Perturbation Methods for Large-Scale Tabular Data Protection’, accepted subject to revision to *European Journal of Operational Research*, 2003
2. Castro, J., Giessing S. (2006). Testing variants of minimum distance controlled tabular adjustment, in *Monographs of Official Statistics*. Work session on Statistical Data Confidentiality, Eurostat-Office for Official Publications of the European Communities, Luxembourg, 2006, pp. 333-343
3. Cox, L., Dandekar, R.H., (2002), ‘Synthetic Tabular Data – an Alternative to Complementary Cell Suppression, unpublished manuscript
4. De Wolf, P.P. (2002), ‘HiTaS: A Heustic Approach to Cell Suppression in Hierarchical Tables’, In: ‘*Inference Control in Statistical Databases*’ Domingo-Ferrer (Ed.), Springer (Lecture notes in computer science; Vol. 2316)
5. Fischetti, M, Salazar Gonzales, J.J. (2000), ‘Models and Algorithms for Optimizing Cell Suppression in Tabular Data with Linear Constraints’, in *Journal of the American Statistical Association*, Vol. 95, pp 916
6. Giessing, S., Repsilber, D. (2002), ‘Tools and Strategies to Protect Multiple Tables with the GHQUAR Cell Suppression Engine’, In: ‘*Inference Control in Statistical Databases*’ Domingo-Ferrer (Ed.), Springer (Lecture notes in computer science; Vol. 2316)
7. Giessing, S. (2003), ‘Co-ordination of Cell Suppressions: strategies for use of GHM*ITER*’, paper presented at the Joint ECE/Eurostat Worksession on Statistical Confidentiality in Luxembourg, 7.-10. April 2003

8. Giessing, S.(2004), Survey on methods for tabular protection in ARGUS, Lecture Notes in Computer Science. Privacy in statistical databases 3050, 1-13. Volume Privacy in statistical databases, J. Domingo-Ferrer and V. Torra, Springer, Berlin.
9. Giessing, S., Dittrich, S., Gehrling, D., Krüger, A., Merz, F.J., Wirtz, H. (2006), , Bericht der Arbeitsgruppe „Geheimhaltungskonzept des statistischen Verbundes, Pilotanwendung: Umsatzsteuerstatistik“', document for the meeting of the „Ausschuss für Organisation und Umsetzung“, Mai 2006, in German
10. Hundepool, A., van de Wetering, A., Ramaswamy, R., de Wolf, P.P., Giessing, S., Fischetti, M., Salazar, J.J., Castro, J., Lowthian, P. (2006), τ-ARGUS users's manual, version 3.1.
11. Repsilber, D. (2002), 'Sicherung persönlicher Angaben in Tabellendaten' - in *Statistische Analysen und Studien Nordrhein-Westfalen, Landesamt für Datenverarbeitung und Statistik NRW,* Ausgabe 1/2002 (in German)
12. Salazar, J.J. (2003) "Partial Cell Suppression: a New Methodology for Statistical Disclosure Control", *Statistics and Computing*, 13, 13-21

Appendix

Table 3a. Number of district level secondary suppressions resulting from *decentralized* co-ordination approaches by levels of NACE classification

Level of NACE	Modular	Hyper0	Hypercube	HyperSeq
0	1	1	1	1
1	390	365	475	568
2	479	589	785	1055
3	2736	2996	3352	4009
4	8438	9476	10533	12609
5	12742	14283	15064	17328
6	16757	19268	19789	21351
Overall	41543	46978	49999	56921

Table 3b. Number of district level secondary suppressions resulting from *centralized* and *two blocks* co-ordination approaches by levels of NACE classification

Level of NACE	Centralized				Two blocks	
	Modular	Hyper0	Hypercube	HyperSeq	Modular	HyperSeq
0	1	1	1	2	1	2
1	381	362	478	561	394	564
2	495	584	788	1048	494	1049
3	2738	2998	3343	3864	2738	3926
4	8427	9482	10528	12355	8444	12466
5	12741	14295	15072	16970	12741	17161
6	16764	19281	19794	21165	16754	21275
Overall	41547	47003	50004	55965	41566	56443

Table 4. Number of cells by ranges of relative deviation and by level of NACE classification

Range	Level of NACE classification							Overall	
	0	1	2	3	4	5	6	abs.	%
0	1	11	13	36	155	329	388	933	81.4%
0 - 0,5	.	4	2	15	35	43	50	149	13.0%
0,5 - 1	.	.	1	.	3	3	7	14	1.2%
1 - 2	.	.	.	2	5	5	7	19	1.7%
2 - 3	.	.	.	.	2	2	1	5	0.4%
3 - 4	.	.	.	1	.	3	2	6	0.5%
4 - 5	.	.	.	.	3	.	.	3	0.3%
5 - 7	.	.	.	.	1	2	.	3	0.3%
7 - 10	.	.	.	3	3	1	2	9	0.8%
10 - 15	.	.	.	.	.	1	1	2	0.2%
15 - 20	.	.	.	.	.	.	.	.	-
20 - 30	.	.	.	.	.	.	1	1	0.1%
> 30	.	.	.	.	1	1	.	2	0.2%
Overall	1	15	16	57	208	390	459	1,146	100 %

Effects of Rounding on the Quality and Confidentiality of Statistical Data

Lawrence H. Cox and Jay J. Kim

National Center for Health Statistics, 3311 Toledo Road
Hyattsville, MD 20782 USA
{LCOX, JKIM5}@CDC.GOV

Abstract. Statistical data may be rounded to integer values for statistical disclosure limitation. The principal issues in evaluating a disclosure limitation method are: (1) Is the method effective for limiting disclosure? and (2) Are the effects of the method on data quality acceptable? We examine the first question in terms of the posterior probability distribution of original data given rounded data and the second by computing expected increase in total mean square error and expected difference between pre- and post-rounding distributions, as measured by a conditional chi-square statistic, for four rounding methods.

Keywords: conventional, unbiased, zero-restricted 50/50 rounding; Chi-square.

1 Introduction

Rounding (base B) amounts to replacing data values x with multiples R(x) of a predetermined positive integer rounding base B. Rounding to adjacent multiples of B is typically preferred, viz., $|x - R(x)| < B$, but not always possible, in which case suitable "neighboring" values are sought. Controlled rounding assures that additive relationships between data values are preserved ([1]). Data are rounded for purposes including elimination of unnecessary detail or visual clutter and for statistical disclosure limitation (SDL) ([2]).

The principal questions in evaluating an SDL method are: (1) Is the method effective for limiting disclosure? and (2) Are the effects of the method on data quality acceptable? We examine these questions for four rounding rules and present a general approach. We evaluate (1) in terms of the posterior probability of an original data value given its rounded value and (2) in terms expected increase in total mean squared error and expected difference between pre- and post-rounding distributions as measured by a conditional Chi-square statistic.

Section 2 describes the four rounding rules. We examine questions (1) and (2) in reverse order. Section 3 examines effects of rounding on total mean square error ([4]), and Section 4 examines its effects on the original x-distribution. Section 5 examines rounding's effectiveness for SDL in terms of $P[x|R(x)]$. Section 6 offers concluding comments.

J. Domingo-Ferrer and L. Franconi (Eds.): PSD 2006, LNCS 4302, pp. 48 – 56, 2006.

2 Four Rounding Rules

For a fixed integer (rounding base) B, any value can be expressed as $x = q_x B + r_x$, where q_x and r_x are the quotient and remainder. If x is a random variable, then q_x and r_x are random variables. For convenience, we assume x is an integer and, when the subscript x is not needed, we ignore it. We focus on adjacent rounding, so rounding x base B replaces x either by R(x) = qB or (q + 1)B, so that R(x) = qB + R(r), with R(r) = 0 or B.

We consider four rounding rules. The first is conventional rounding for which R(r) equals the multiple of B closest to r, for all r except: if B is even and r = B/2, then R(r) = B, viz., round B/2 "up". The second rounding rule is modified conventional rounding which is the same as conventional rounding, except that for B even B/2 rounds down or up each with probability ½. The third rule is zero-restricted 50/50 rounding for which r = 0 is rounded down and any other r is rounded down or up with probability ½. This rounding rule was suggested by the second author and Joe Fred Gonzalez, Jr. The last rule is unbiased rounding ([3]) for which r is rounded up with probability r/B and down with probability (B-r)/B.

From a stochastic viewpoint, we can distinguish two types of rounding. The first is unbiased for which E[R(r)lr] = r. The other is sum-unbiased for which E[R(r)] = E[r], and consequently the expected value of a sum of rounded entries equals the sum of the original entries. The second type is weaker than the first. All four rounding rules are sum-unbiased, except conventional rounding with B even. Only the fourth rule is unbiased.

Unbiased rounding can be performed in a controlled manner (viz., preserving additive structure) for two-dimensional tables and other tables that can be modeled by a mathematical network ([1], [3]). (Modified) conventional rounding cannot be performed in a controlled manner. The theorem below demonstrates that, whenever unbiased rounding can be controlled, so can zero-restricted 50/50 rounding.

Theorem. For any tabular structure that can be represented by a network (including two-dimensional tables), base B zero-restricted 50/50 rounding can be performed in a controlled manner.

Proof: Divide all cell values by B to reduce to the case of base 1 rounding of discrete values. The network consists of forward arcs (for increasing each value) and reverse arcs (for decreasing each value) connecting totals entries. Arcs representing any integer cell value have capacity zero; arcs representing non-integer cell values have reverse capacity equal to v and forward capacity equal to 1-v. Pick any non-integer cell value; denote its non-integer part v. Round it up or down with probability ½. Assume for convenience it was rounded down. Set the capacity on its forward arc to zero and compute a maximum flow through the reverse arc. The solution will round v down and change the values of other cells, some of which may become integers. Update capacities and continue in this manner. The algorithm terminates with a zero-restricted base 1 rounding. Re-multiply all values by B to obtain a zero-restricted base B rounding of the original. Values can be changed at intermediate steps, but under 50/50 probabilities these have no effect on the final rounded value—only the last probability step for that value affects its rounded value and that is a 50/50 step. Thus, we have described an algorithm for controlled base B zero-restricted 50/50 rounding. Q.E.D.

3 Effects of Rounding on Total Mean Squared Error

Throughout this paper we make two distributional assumptions. First, we assume that the r-distribution is uniform. Second, we assume that the r- and q-distributions are independent. Consequently, E[x] = BE[q] + E[r] and P(r) = P(r|q) = 1/B. The assumptions imply V(x) = B^2 V(q) + V(r), so that

$$E(r) = \frac{B-1}{2} \qquad \text{and} \qquad V(r) = \frac{B^2-1}{12} \tag{1}$$

$$E[x] = BE[q] + \frac{B-1}{2} \qquad \text{and} \qquad V(x) = B^2 V(q) + \frac{B^2-1}{12} \tag{2}$$

For conventional rounding with B even:

$$E\left[R(r)\right] = \frac{B}{2} \qquad \text{and} \qquad V\left[R(r)\right] = \frac{B^2}{4} \tag{3}$$

For conventional rounding with B odd:

$$E\left[R(r)\right] = \frac{B-1}{2} \qquad \text{and} \qquad V[R(r)] = \frac{B^2-1}{4} \tag{4}$$

Consequently, conventional rounding for B even introduces an absolute bias of ½, whereas for B odd the rule is sum-unbiased. For modified conventional rounding, $P\left[R(r)=0\right] = \frac{B+1}{2B}$ and $P\left[R(r)=B\right] = \frac{B-1}{2B}$, thus $E\left[R(r)\right] = \frac{B-1}{2}$ and $V[R(r)] = \frac{B^2-1}{4}$, as in (4). Consequently, modified conventional rounding is sum-unbiased. As zero-restricted 50/50 rounding exhibits the same rounding probabilities as modified conventional rounding, (4) again holds and zero-restricted 50/50 rounding is sum-biased.

For unbiased rounding, $P\left[R(r)=0\right] = \sum_{v=0}^{B-1} P(r)\,P[R(r)=0 \mid r] = \frac{B+1}{2B}$ and for r ≥ 1, $P\left[R(r)=B\right] = \sum_{v=1}^{B-1} P(r)\,P[R(r)=B \mid r] = \frac{B-1}{2B}$. These probabilities are the same for modified conventional rounding and (4) holds. Also, this rule is unbiased, not simply sum-unbiased. All results are summarized below.

Table 1. Expected Value of r and R(r)

Unrounded	Rounding Method				
	Conventional B Even	Conventional B Odd	Modifed Conv B even	50/50	Unbiased
$\frac{B\text{-}1}{2}$	$\frac{B}{2}$	$\frac{B\text{-}1}{2}$	$\frac{B\text{-}1}{2}$	$\frac{B\text{-}1}{2}$	$\frac{B\text{-}1}{2}$

Table 2. Variance of r and R(r)

Unrounded	Rounding Method				
	Conventional B Even	Conventional B Odd	Modified Conv B even	50/50	Unbiased
$\frac{B^2-1}{12}$	$\frac{B^2}{4}$	$\frac{B^2-1}{4}$	$\frac{B^2-1}{4}$	$\frac{B^2-1}{4}$	$\frac{B^2-1}{4}$

With the exception of conventional rounding for B even, all rounding methods have equivalent effects on mean and total mean square error.

4 Effects of Rounding on the Original x-Distribution

A statistic that can be used to decide whether a second distribution is significantly different from an original (first) distribution is the conditional Chi-square. If the x-data are at hand, then the first distribution is the x-distribution and the second is the R(x)-distribution, and the Chi-square statistic conditional on the x-data is: $\chi^2 = \sum_x U_x$ for

$$U_x = \frac{[R(x)-x]^2}{x} = \frac{[R(r_x)-r_x]^2}{x} \qquad \text{(when x = 0, U}_\text{x}\text{ = 0)} \qquad (5)$$

If the rounded data are also at hand, then the data released can compute the Chi-square statistic (5) directly, determine the degrees of freedom, and test for a significant difference between original x- and rounded R(x)-distributions. For tables, degrees of freedom *df* is determined by the tabular structure.

Our objective, and often that of the data releaser, is to test whether rounding is expected to alter the original distribution. There are two scenarios: the x-data are available, and, the x-data are not yet available but assumptions about the x-distribution are available.

In the first scenario, the conditional expectation given x of U$_\text{x}$ over the rounding process is:

$$E\ [U_x \mid x] = \sum_{r_x} \left(\left[\frac{r_x^{\ 2}}{x} \right] P[R(r_x)=0] + \left[\frac{(B-r_x)^2}{x} \right] P[R(r_x)=B] \right) \qquad (6)$$

In this situation, we develop a method for determining whether rounding base B is expected to alter the original x-distribution. We derive the method for unbiased rounding, as follows. Let *d* denote the number of x-observations and *e* the number of x-observations with value less than B, viz., zeroes and confidential values. As x ≥ qB and assumed independence of r- and q-distributions, we observe

$$
\begin{aligned}
E[U] = E\ [\sum_{x} U_x] &= E[\sum_{q=o} U_x + \sum_{q\geq 1} U_x] = \sum_{q=0} E[U_x \mid q=0] + E[\sum_{q\geq 1} U_x \mid q] \\
&= (0 + \sum_{q=0} E[\frac{(R(r)-r)^2}{r} \mid 1 \leq r \leq B-1]) + E[\sum_{q\geq 1} \frac{(R(r)-r)^2}{qB+r} \mid q] \\
&\leq \sum_{q=0} E[\frac{(R(r)-r)^2}{r} \mid 1 \leq r \leq B-1]) + E[\sum_{q\geq 1} (\frac{(R(r)-r)^2}{B})(\frac{1}{q}) \mid q] \\
&\leq \sum_{q=0} E[\frac{(R(r)-r)^2}{r} \mid 1 \leq r \leq B-1]) + (E[\frac{(R(r)-r)^2}{B}])\sum_{q\geq 1} E[\frac{1}{q} \mid q] \qquad (7)
\end{aligned}
$$

For unbiased rounding, $E[\frac{(R(r)-r)^2}{B}] = \frac{B^2-1}{6}$ and $E[\frac{(R(r)-r)^2}{r} \mid 1 \leq r \leq B-1] = \frac{B(B-1)}{2}$. We have

$$
E[U] \leq e\frac{B(B-1)}{2} + (d-e)\frac{B^2-1}{6}E[\frac{1}{q} \mid q \geq 1] \qquad (8)
$$

The data releaser can use the x-data to estimate the last term of (8) and test the hypothesis of no expected distributional change due to unbiased rounding base B. Table 3 provides terms for (7) for all four rounding methods (henceforth, conventional rounding for B even is replaced by modified conventional rounding, which is superior). An opposite inequality to (8) can be developed using x ≤ (q+1)B yielding bounds l ≤ E[U] ≤ u if needed.

Relationship (8) expresses quantitatively notions that are plausible intuitively, e.g., unbiased rounding base B should be avoided if the x-distribution is dominated by values small relative to B or if the full x-distribution is not sufficiently separated from B. In this manner, (8) provides a rule of thumb that a data releaser can apply prior to data collection based on assumptions about the q-distribution or after collection once the q-distribution is available.

In the second scenario, the x-data are not available but based on past experience, distributional assumptions, etc., the data releaser can make estimates concerning the x-distribution. An important example is a recurring (monthly, annual) survey. For all rounding methods (except conventional rounding with B even) the method is sum-unbiased and the Chi-square statistic can take the form $\chi^2 = \sum_{x} E[U_x]$ for

$$
U_x = \frac{(R(x)-E[x])^2}{E[x]} = \frac{(R(r_x)-E[r_x])^2}{E[x]} = \frac{(R(r_x)-E[r_x])^2}{BE[q]+E[r_x]} \quad (x = 0,\ U_x = 0) \qquad (9)
$$

Table 3. $E[\frac{(R(r)-r)^2}{r}]$ and $E[\frac{(R(r)-r)^2}{B}]$

	Modif. Conv Even B	Conventional Odd B	50/50	Unbiased
$E[\frac{(R(r)-r)^2}{r}]$	$B^2 \sum_{s=\frac{B}{2}+1}^{B-1} \frac{1}{s}$ $-\frac{B(B-3)}{2}$	$B^2 \sum_{s=\frac{B+1}{2}}^{B-1} (\frac{1}{s})$ $-\frac{B(B-1)}{2}$	$\frac{B^2}{2} \sum_{s=1}^{B-1} (\frac{1}{s})$ $-\frac{B(B-1)}{2}$	$\frac{B(B-1)}{2}$
$E[\frac{(R(r)-r)^2}{B}]$	$\frac{B^2+2}{12}$	$\frac{B^2-1}{12}$	$\frac{(B-1)(2B-1)}{6}$	$\frac{B^2-1}{6}$

Each expectation in (9) is known (Tables 1, 3), with the exception of E[q], for which the data releaser can estimate a (range of) value(s) from prior knowledge/experience or distributional assumptions. The releaser then can compute the Chi-square statistic and test in advance whether a particular form of rounding to a particular base is expected to alter the original distribution significantly. Note that because the Chi-square test is distribution-free, so are our methods.

5 Effectiveness of Rounding for Disclosure Limitation

In this section, we evaluate the effectiveness of rounding base B for statistical disclosure limitation. Specifically, we are interested in determining posterior probabilities: P[x=qB+r|R(x)=mB]. These probabilities equal zero except for m = q or q+1, so we may focus on P[x=qB+r|R(r)=0] and P[x=qB+r|R(r)=B]. We report results for all four rounding rules, but derive the method only for unbiased rounding.

Except for m=q=0, a rounded value mB can be achieved in two ways under unbiased rounding: by rounding down values in the range {mB, mB+1,…., (m+1)B-1} or by rounding up values in the range {(m-1)B+1, …, mB-1}. Thus, there are 2B-1 possible values for x|R(x). Now

$$\mathrm{P[r \mid R(r) = B]} = \frac{\mathrm{P(r)\,P[R(r) = B|r]}}{\sum_{\mathrm{r}} \mathrm{P(r)\,P[R(r) = B|r]}} \tag{10}$$

Consider first rounding up. As $\mathrm{P(r)} = \frac{1}{\mathrm{B}}$ and $\mathrm{P[R(r) = B \mid r]} = \frac{\mathrm{r}}{\mathrm{B}}$, for $\mathrm{r} \geq 1$, then

$$P[R(r) = B] = \sum_{r=1}^{B-1} P(r)\,P[R(r) = B|r] = \sum_{1}^{B-1} \frac{1}{B}\frac{r}{B} = \frac{B-1}{2B} \text{ and}$$

$$P[r \mid R(r) = B] = \frac{r/B^2}{(B-1)/2B} = \frac{2r}{B(B-1)} \tag{11}$$

Probabilities depend on r. For B=10, the range is $\frac{1}{45}$ (r = 1) to $\frac{1}{5}$ (r = 9).

These probabilities are the same as that for zero restricted 50/50 rounding.

For rounding down, P(r) = $\frac{1}{B}$ and P[R(r) = 0 | r] = $\frac{B-r}{B}$ for all r. Thus,

$$P[R(r) = 0] = \sum_{r=0}^{B-1} P(r)\,P[R(r) = 0|r] = \sum_{0}^{B-1} \frac{1}{B}\frac{B-r}{B} = \frac{B+1}{2B}$$

Hence,

$$P[r \mid R(r) = 0) = \frac{(B-r)/B^2}{(B+1)/2B} = \frac{2(B-r)}{B(B+1)} \tag{12}$$

For B = 10, this probability ranges from $\frac{2}{11}$ (r=1) to $\frac{1}{55}$ (r=9).

We now evaluate P[x | R(x) = qB, q>0] for unbiased rounding. The rounded value R(x) can result from rounding r up when x = (q-1)B + r or rounding r down when x = qB + r. As this depends only on r, which we continue to assume is uniformly distributed, then P(x) = $\frac{1}{2B-1}$. When R(x)=qB results from rounding up, q > 0 and

$$P[R(x) = qB|x = (q-1)B + r] = \frac{r}{B}.$$

When R(x)=qB results from rounding down, $P[R(x) = qB|x = qB + r] = \frac{B-r}{B}$.

Hence,

$$P[R(x) = qB] = \frac{1}{2B-1}\left[\sum_{(q-1)B+1}^{(q-1)B+B-1} \frac{r}{B} + \sum_{qB}^{qB+B-1} \frac{B-r}{B}\right] = \frac{B}{2B-1} \tag{13}$$

Thus,

$$P[x \mid R(x) = qB] = \frac{1/(2B-1)\; r/B}{(B-1)/(2B-1)} = \frac{r}{B(B-1)} \quad \text{if x is rounded up}$$

viz.,

$$\text{for}\,(q-1)B + 1 \le x \le (q-1)B + (B-1) \;\; , q > 0 \qquad (14)$$

These probabilities range from $\frac{1}{B(B\text{-}1)}$ (r=1) to $\frac{1}{B}$ (r=B-1)

e.g., for B=10, q=1, P[x=19|R(x)=20] = $\frac{9}{B(B\text{-}1)} = \frac{1}{10}$.

Similarly, for rounding down

$$P[x\ |R(x) = qB] = \frac{\frac{1}{(2B\text{-}1)} \; \frac{(B\text{-}r)}{B}}{\frac{B}{(2B\text{-}1)}} = \frac{B\text{-}r}{B(B\text{-}1)} \;\; \text{if x is rounded down}$$

$$\text{viz., for } qB \le x \le qB + (B-1) \qquad (15)$$

These probabilities range from $\frac{1}{B(B\text{-}1)}$ (r=B-1) to $\frac{1}{B\text{-}1}$ (r=0)

e.g., in the preceding example, P[x = 19|R(x) = 10] = 1/90 and P[x=10|R(x)=10] = 1/9.

Our analysis of P[x|R(x)] is summarized in Tables 4 and 5.

Table 4. P[x|R(x)=mB] for m>0

Rounding Direction for x=qB+r	Conventional Odd B	Modified B even $\{r = \frac{B}{2}\}$	50/50 $\{r = 0\}$	Unbiased
Up R(x)=(q+1)B	$\frac{1}{B}$, 0	$\frac{1}{B}$ $\{\frac{1}{2B}\}$	$\frac{1}{2B}$ $\{\frac{1}{B}\}$	$\frac{r}{B(B-1)}$
Down R(x)=qB	$\frac{1}{B}$, 0	$\frac{1}{B}$ $\{\frac{1}{2B}\}$	$\frac{1}{2B}$ $\{\frac{1}{B}\}$	$\frac{B\text{-}r}{B(B-1)}$

Table 5. P[x|R(x)=0]

Rounding Direction for x=r	Conventional Odd B	Modified B even $\{r = \frac{B}{2}\}$	50/50 $\{r = 0\}$	Unbiased
Down R(x)=0	$\frac{2}{B+1}$, 0	$\frac{2}{B-1}$ $\{\frac{1}{B-1}\}$	$\frac{1}{B+1}$ $\{\frac{2}{B+1}\}$	$\frac{2(B-r)}{B(B+1)}$

Table 5 is the focus of interest from a confidentiality standpoint for count data. Given our assumption that the prior r-probabilities were uniform over the set of B possible r-values {0, 1, …, B-1}, then from a confidentiality standpoint the optimal posterior probabilities would be uniform probabilities over this set or, as typically x=r=0 is not considered a confidential value, uniform over its nonzero values, because this means that rounding has provided no additional information about an original confidential value x=r. For B even, conventional rounding probabilities are uniform over {0, 1, …, B/2} and modified conventional rounding probabilities are nearly the same. For B odd, conventional rounding probabilities are uniform over {0, 1, …, (B+1)/2}. In each case, instead of B equi-probable choices for r, there are about half as many. For unbiased rounding, posterior probabilities are not at all uniform, and rounding has provided additional information about original confidential values x=r. For zero-restricted 50/50 rounding, posterior probabilities are uniform over {1, …, B-1}. So, if x=r=0 is not confidential, the posterior probabilities are uniform over the confidential values, and nearly so in any case. We conclude that zero-restricted 50/50 rounding is the preferred rounding method from a confidentiality standpoint.

6 Concluding Comments

We analyzed the effects of rounding on data quality and the effectiveness of rounding for statistical disclosure limitation. We provided quantitative expressions (8), (9) for deciding whether or how to round data based on statistical hypothesis testing for the conditional Chi-square statistic.

In terms of three data quality measures: total mean square error, preserving the original x-distribution, and preserving additive structure, unbiased rounding and zero-restricted 50/50 rounding perform well overall. In terms of statistical disclosure limitation, zero-restricted 50/50 rounding performs best.

We conclude that zero-restricted 50/50 rounding is the superior rounding method for balancing quality and confidentiality in statistical data and tabulations.

Disclaimer. This paper represents the views of the authors and should not be interpreted as representing the views, policies or practices of the Centers for Disease Control and Prevention.

References

1. Cox, L.H. and Ernst, L.R.: Controlled rounding. INFOR: Canadian Journal of Operations Research and Information Processing, Volume 20 (1982) 423-432.
2. Cox, L.H., Fagan, J.,Greenberg, B.and Hemmig, R.: Research at the Census Bureau into disclosure avoidance techniques for tabular data. Proceedings of the Section on Survey Research Methods, American Statistical Association, Alexandria, VA (1986) 388-393.
3. Cox, L.H.: A constructive procedure for unbiased controlled rounding. Journal of the American Statistical Association, Volume 82 (1987) 520-524.
4. Kim, J. J., Cox, L.H., Gonzalez, J.F. and Katzoff, M.J.: Effects of rounding on data quality. Monographs of Official Statistics, Work Session on Statistical Data Confidentiality, Geneva, 9-11 November 2005. United Nations Economic Commission for Europe and EUROSTAT, Luxembourg (2005) 255 - 265.

Disclosure Analysis for Two-Way Contingency Tables*

Haibing Lu[1], Yingjiu Li[1], and Xintao Wu[2]

[1] Singapore Management University, 80 Stamford Road, Singapore 178902
[2] University of North Carolina at Charlotte, 9201 Univ City Blvd,
Charlotte, NC 28223
{hblu, yjli}@smu.edu.sg, xwu@uncc.edu

Abstract. Disclosure analysis in two-way contingency tables is important in categorical data analysis. The disclosure analysis concerns whether a data snooper can infer any protected cell values, which contain privacy sensitive information, from available marginal totals (i.e., row sums and column sums) in a two-way contingency table. Previous research has been targeted on this problem from various perspectives. However, there is a lack of systematic definitions on the disclosure of cell values. Also, no previous study has been focused on the distribution of the cells that are subject to various types of disclosure. In this paper, we define four types of possible disclosure based on the exact upper bound and/or the lower bound of each cell that can be computed from the marginal totals. For each type of disclosure, we discover the distribution pattern of the cells subject to disclosure. Based on the distribution patterns discovered, we can speed up the search for all cells subject to disclosure.

1 Introduction

In this paper, we focus on the disclosure problem for two-way contingency tables. The traditional disclosure problem in two-way contingency tables, which has been formulated before (e.g., in [40, 11]), asks whether a data snooper can infer accurate information about any protected cell values given the marginal totals. In this context, the internal cells of a contingency table provide privacy sensitive information, which should be protected, while the marginal totals are the sums of cell values in a row or column, which can be released to the public if they lead to no disclosure of any cell values. This problem has many practical applications such as medical/health statistics, national census, and student records management. In health insurance data, for example, it is important to protect a cell value, which represents how many times a patient undergoes a certain treatment, against being inferred from the marginal totals, which are aggregate statistics on the total number of each treatment being taken or the total number of each patient visiting doctors. For another example, in an agent-stock

* Work of Lu and Li was supported in part by SMU Research Office. Work of Wu was supported in part by USA National Science Foundation Grant IIS-0546027.

J. Domingo-Ferrer and L. Franconi (Eds.): PSD 2006, LNCS 4302, pp. 57–67, 2006.

table, where each cell indicates the volume of a stock in which an agent invests, a commercial secret may be revealed if a snooper infers from the released marginal totals that the agent buys more (or less) than certain amount of the stock.

Previous study on this problem has identified that the disclosure of any cell value depends on the upper bound and lower bound of the cell value which a snooper can derive from the available marginal information (e.g., see [21,22,6]). If the upper bound is the same as the lower bound, the cell value is exposed. Likewise, if the difference between the upper bound and the lower bound is very small, the security of the table is also considered to be compromised [40]. However, there is a lack of systematic definitions on the disclosure of cell values. Also, no previous study has been focused on the distribution of the cells that are subject to various types of disclosure. In this paper, we define four types of possible disclosure based on the exact upper bound and/or the lower bound of each cell that can be computed from the marginal totals. For each type of disclosure, we discover the distribution pattern of the cells subject to disclosure. Based on the distribution patterns discovered, we propose two efficient methods to speed up the search for all cells subject to disclosure.

The rest of this paper is organized as follows. Section 2 presents the preliminaries for the research of disclosure analysis. Section 3 defines various types of disclosure that are commonly used in practice. Section 4 reveals in a contingency table the distribution patterns of the cells that are subject to different types of disclosure. Based on the distribution patterns discovered, Section 5 investigates how to efficiently detect all cells subject to disclosure. Section 6 reviews the related work. Finally, Section 7 concludes the paper.

2 Preliminaries

A two-way contingency table A with m rows and n columns is denoted by $\{a_{ij} \mid 1 \leq i \leq m, 1 \leq j \leq n\}$, where $a_{ij} \geq 0$. In tradition, the cell values in a contingency table are usually assumed to be nonnegative integers (e.g., counts). We extend this assumption such that the cell values can be nonnegative real numbers. The results given in this paper hold in both integer domain and real domain.

Denote $a_{+j} = \sum_{i=1}^{m} a_{ij}$, $a_{i+} = \sum_{j=1}^{n} a_{ij}$, and $a_{++} = \sum_{ij} a_{ij}$, where a_{+j} and a_{i+} are *marginal totals* and a_{++} is the *grand total*. The marginal totals satisfy $\sum_{j=1}^{n} a_{+j} = \sum_{i=1}^{m} a_{i+} = a_{++}$, which is called the *consistency condition*.

The marginal totals of a two-way contingency table can be released while the cell values are protected. A traditional disclosure analysis question in a two-way contingency table asks [40,11]: *Can any information about protected cells be inferred from the released marginal totals?* The answer to this question depends on the bounds that a snooper can obtain about a protected cell from the marginal totals that are given [21,22,6].

A nonnegative value $\underline{a}_{ij}$ is said to be a *lower bound* of cell value a_{ij} if, for any contingency table $\{a'_{ij}\}$ that has the same marginal totals as A, the inequality $\underline{a}_{ij} \leq a'_{ij}$ holds. A value $\underline{a}_{ij}$ is said to be the *exact lower bound* of a_{ij} if (i) it is a lower bound; and (ii) there exists a contingency table $\{a'_{ij}\}$ such that a) the marginal totals of A' are the same as those of A, and b) $a'_{ij} = \underline{a}_{ij}$. An upper bound or the exact upper bound $\overline{a}_{ij}$ can be defined similarly.

Definition 2.1. (Fréchet bounds) *Given marginal totals $\{a_{+j}\}$ and $\{a_{i+}\}$ of a two-way contingency table A, the Fréchet bounds for any cell value a_{ij} are*

$$\max\{0, a_{i+} + a_{+j} - a_{++}\} \leq a_{ij} \leq \min\{a_{i+}, a_{+j}\}$$

The Fréchet bounds are exact bounds as proven in [13]. Therefore, the Fréchet bounds give a data snooper the "best" estimate of a protected cell from the marginal totals.

3 Disclosure Types

Based on the exact bounds, we define four types of information disclosure in two-way contingency tables: existence disclosure, threshold upward disclosure, threshold downward disclosure, and approximation disclosure.

Definition 3.1. Existence disclosure: *The exact lower bound of a protected cell is positive.*

The concept of existence disclosure can be illustrated using a patient-treatment table. In such table, each cell shows the number of times that a patient undergoes a particular treatment. To protect each patient's privacy, only the marginal totals are released. However, from the marginal totals, a snooper can easily calculate the exact lower bound of each cell. If an exact lower bound is positive, the snooper may infer that a patient has suffered from certain disease. This type of disclosure is common in privacy protection of statistical data.

Definition 3.2. Threshold upward disclosure: *The exact lower bound of a protected cell is greater than a positive threshold.*

Definition 3.3. Threshold downward disclosure: *The exact upper bound of a protected cell is less than a positive threshold.*

The threshold upward disclosure is similar to the existence disclosure with the difference that the threshold is a positive value rather than zero, while the threshold downward disclosure is a dual to the threshold upward disclosure. In certain applications, knowing that a cell value is positive is not harmful, while knowing that the cell value is greater or less than certain threshold is dangerous. For example, in an agent-stock contingency table, where each cell indicates the volume of certain stock in which an agent invests, it is often trivial if a snooper deduces that an agent invests in certain stock, but a commercial secret may be revealed if the snooper infers that the agent buys more (or less) than certain amount of the stock. These types of disclosure often occur in business and wealth related tables.

Definition 3.4. Approximation disclosure: *The difference between the exact lower bound and the exact upper bound of a protected is less than a positive threshold.*

This type of disclosure is defined based on not only the exact lower bound but also the exact upper bound. If the difference between the two exact bounds for a protected cell is small enough, one can estimate the cell's value with a high precision. For example, if one knows that a professor's salary is between 90K and 92K, then the actual salary amount is largely revealed.

Among the four types of disclosure, the definitions of existence disclosure and approximation disclosure summarize the similar concepts discussed in some previous papers (e.g., [40,11,34]). To be more systematic, we extend these concepts to threshold upward disclosure and threshold downward disclosure.

4 Distribution of Cells Subject to Disclosure

In the previous section, we have defined four types of information disclosure in a contingency table. In this section, we study the distribution of the cells that are subject to various types of disclosure. For the first time we discover that the cells subject to disclosure demonstrate some regular patterns.

Theorem 4.1. *Consider existence disclosure or threshold upward disclosure with a fixed threshold in a two-way contingency table. The cells subject to disclosure, if exist, must appear in the same row or column, but not both.*

Proof. Prove by contradiction. Assume there exist two cells $a_{i_1j_1}$ and $a_{i_2j_2}$ subject to existence disclosure and $i_1 \neq i_2, j_1 \neq j_2$. Then

$$a_{i_1+} + a_{+j_1} - a_{++} > 0$$
$$a_{i_2+} + a_{+j_2} - a_{++} > 0$$

These two inequalities lead to $a_{i_1+} + a_{i_2+} - a_{++} - \sum_{j \neq j_1,j_2} a_{+j} > 0$. A contradiction is committed as $a_{i_1+} + a_{i_2+} - a_{++} - \sum_{j \neq j_1,j_2} a_{+j} \leq 0$ must hold (note that $a_{i_1+} + a_{i_2+} - a_{++} \leq 0$). Thus, the theorem is proven for the existence disclosure. Since any cell subject to threshold upward disclosure must also be subject to existence disclosure, the theorem is proven for the threshold upward disclosure. ♢

The above theorem reveals the distribution pattern for the cells that are subject to existence disclosure or threshold upward disclosure. This pattern can be used to limit the search for cells subject to existence disclosure or threshold disclosure, which we will discuss in the next section.

Now consider the distribution of the cells that are subject to threshold downward disclosure or approximation disclosure. The following lemma compares the difference of the exact bounds for any cells that are subject to existence disclosure with that for any cells that are not subject to existence disclosure.

Lemma 4.1. *The difference of the exact bounds for any cell that is subject to existence disclosure is no less than that for any cell that is not subject to existence disclosure in a two-way contingency table.*

Proof. Assume $a_{i_1j_1}$ is subject to existence disclosure. The difference of its exact bounds is

$$min\{a_{i_1+}, a_{+j_1}\} - (a_{i_1+} + a_{+j_1} - a_{++}) = min\{\sum_{i \neq i_1} a_{i+}, \sum_{j \neq j_1} a_{+j}\}$$

Consider any other cell $a_{i_2j_2}$ that is not subject to existence disclosure. Because the exact lower bound of $a_{i_2j_2}$ is zero, the difference of its exact bounds is $min\{a_{i_2+}, a_{+j_2}\}$. To prove the theorem, we need to prove

$$min\{a_{i_2+}, a_{+j_2}\} \leq min\{\sum_{i \neq i_1} a_{i+}, \sum_{j \neq j_1} a_{+j}\}$$

We prove this in three possible cases: (i) $i_2 \neq i_1, j_2 \neq j_1$, (ii) $i_2 \neq i_1, j_2 = j_1$, and (iii) $i_2 = i_1, j_2 \neq j_1$. Clearly, the inequality holds for case (i). In the following, we prove the theorem for case (ii) only. The proof for case (iii) is similar to case (ii).

In case (ii), let $j_1 = j_2 = j'$. Since $i_1 \neq i_2$, we have $a_{+j'} = a_{i_1j'} + a_{i_2j'} + \sum_{i \neq i_1, i_2} a_{ij'}$ and $a_{i_2+} = a_{i_2j'} + \sum_{j \neq j'} a_{i_2j}$. Because $a_{i_1j'}$ is subject to existence disclosure, we have $a_{i_1+} + a_{+j'} - a_{++} = a_{i_1j'} - \sum_{i \neq i_1, j \neq j'} a_{ij} > 0$; then, we have $a_{i_1j'} > \sum_{i \neq i_1, j \neq j'} a_{ij} \geq \sum_{j \neq j'} a_{i_2j}$. Therefore, we have $a_{+j'} > a_{i_2+}$. Since $a_{i_2j'}$ is not subject to existence inference, the difference of the exact bounds for $a_{i_2j'}$ is a_{i_2+}. To prove the theorem, we need to prove $a_{i_2+} \leq min\{\sum_{i \neq i_1} a_{i+}, \sum_{j \neq j'} a_{+j}\}$.

On the one hand, it is clear $a_{i_2+} \leq \sum_{i \neq i_1} a_{i+}$. On the other hand, since $a_{i_2j'}$ is not subject to existence disclosure, we have $a_{i_2j'} \leq \sum_{i \neq i_2, j \neq j'} a_{ij}$. Adding $\sum_{j \neq j'} a_{i_2j}$ to both sides of this inequality, we have $a_{i_2+} \leq \sum_{j \neq j'} a_{+j}$. The theorem is proven. ♢

From this lemma, one can easily derive the following

Lemma 4.2. *The exact upper bound for any cell that is subject to existence disclosure is no less than that for any cell that is not subject to existence disclosure in a two-way contingency table.*

According to the above lemmas, we have the following theorem regarding the distribution of cells subject to approximation disclosure or threshold downward disclosure.

Theorem 4.2. *Consider approximation disclosure or threshold downward disclosure with a fixed threshold in a two-way contingency table. If a cell is subject to disclosure, the other cells in the same row or column must also be subject to disclosure.*

Proof. First, consider approximation disclosure with a fixed threshold $\tau > 0$. If a cell $a_{i'j'}$ is subject to approximation disclosure, the difference of its exact bounds is less than τ. The theorem is proven in the following two cases.

Case (i): $a_{i'j'}$ is also subject to existence disclosure. From Theorem 4.1, we know that all of the cells that are subject to existence disclosure must be in row i' or column j', but not both. Without loss of generality, we assume that these cells are in row i. Therefore, all of the cells in the column j' except $a_{i'j'}$ are not subject to existence disclosure. According to the Lemma 4.1, the differences of the exact bounds for these cells in column j' are smaller than or equal to the difference of the exact bounds for a_{ij}, which is less than τ. Therefore, all of the cells in column j' are subject to approximation disclosure. The theorem is proven.

Case (ii): $a_{i'j'}$ is not subject to existence disclosure. According to Theorem 4.1, all of the cells in row i' and column j' are not subject to existence disclosure. Since $a_{i'j'}$ is subject to approximation disclosure, we have $min\{a_{i'+}, a_{+j'}\} < \tau$. To prove the theorem, we prove that all of the cells in either row i' or column j' are subject to approximation disclosure. If $min\{a_{i'+}, a_{+j'}\} = a_{i'+} < \tau$, then for any cell $a_{i'j}$ where $j \neq j'$, the difference of the exact bounds for $a_{i'j}$ is $min\{a_{i'+}, a_{+j}\} \leq a_{i'+} < \tau$. Thus, all of the cells in row i' is subject to approximation disclosure. Similarly, if $min\{a_{i'+}, a_{+j'}\} = a_{+j'}$, all of the cells in column j' are subject to approximation disclosure.

Then consider the threshold downward disclosure with a fixed threshold. The theorem can be proven similarly as in the case of approximation disclosure. The only difference is that one needs to replace the phrase "approximation disclosure" with "threshold downward disclosure", "the difference of the exact bounds" with "the exact upper bound", and "lemma 4.1" with "lemma 4.2" in the proof. $\diamondsuit$

Note that the distribution pattern for the cells that are subject to approximation disclosure or threshold downward disclosure is different from that for the cells that are subject to existence disclosure or threshold upward disclosure. The former pattern is a single row or column, but not both, while the latter must "fill" some rows or columns.

5 Disclosure Detection

An important task in contingency table protection is to detect all cells that are subject to disclosure before one can eliminate such disclosure using some disclosure limitation method. We consider disclosure detection in this section, while disclosure limitation will be summarized in the related work section.

A naive approach to disclosure detection is to check all cells one by one. To check whether a cell is subject to disclosure, one needs to compute its Fréchet lower bound (two plus/minus operations and one comparison operation) and/or Fréchet upper bound (one comparison operation), depending on what type of disclosure is of concern. This naive approach requires checking all mn cells in an $m \times n$ contingency table.

We improve this naive approach by reducing its time complexity from $O(mn)$ to $O(m+n)$. Such an improvement is meaningful in practice especially for some information organizations (e.g., statistical offices) which routinely process a large number of sizable contingency tables.

First, consider the existence disclosure and the threshold upward disclosure. According to Theorem 4.1, the cells subject to disclosure must exist in a single row or column, but not both. Based on this distribution pattern, we propose the following

Procedure 1. (Disclosure detection for existence disclosure or threshold upward disclosure)

1. Discover all i' and j' such that $a_{i'+} = \max_i\{a_{i+}\}$ and $a_{+j'} = \max_j\{a_{+j}\}$; proceed to step (2) if $a_{i'j'}$ is subject to disclosure; otherwise, output no cell subject to disclosure.
2. Check all cells in row i'. If no cell is subject to disclosure, continue checking all cells in column j'. Output all cells subject to disclosure that are discovered in both step (1) and step (2).

If there exists at least one cell subject to existence disclosure or threshold upward disclosure, $a_{i'j'}$ must be one of such cells since the exact upper bound of any other cell is less than or equal to the exact upper bound of $a_{i'j'}$. According to this fact and the distribution pattern, it is easy to know that this procedure outputs all and only the cells that are subject to existence disclosure or threshold upward disclosure.

Second, consider the threshold downward disclosure and the approximation disclosure. According to Theorem 4.2, the cells subject to disclosure must "fill" some rows or columns. Based on this distribution pattern, we propose the following

Procedure 2. (Disclosure detection for threshold downward disclosure)

1. Discover all i' and j' such that $a_{i'+} < \tau$ and $a_{+j'} < \tau$.
2. Output all cells in the discovered rows i' and columns j' to be subject to disclosure.

For threshold downward disclosure with threshold τ, a cell $a_{i'j'}$ is subject to disclosure if and only if its marginal total $a_{i'+}$ or $a_{+j'}$ is less than τ. According to this fact and the distribution pattern, it is easy to know that the above procedure outputs all and only the cells that are subject to threshold downward disclosure.

For approximation disclosure with threshold τ, one can classify those cells that are subject to disclosure into two categories: (i) cells that are subject to threshold downward disclosure with threshold τ, and (ii) cells that are not subject to threshold downward disclosure with threshold τ. It is clear that the cells in category (ii) must be subject to existence disclosure (and approximation disclosure). Procedure 2 can be used to discover all and only the cells in category (i), while procedure 1 can be easily extended to discover all and only the cells in category (ii). The union of the cells discovered in categories (i) and (ii) is the set of cells subject to approximation disclosure.

6 Related Work

The problem of protecting sensitive data (e.g., privacy related information) against disclosure from nonsensitive data (e.g., aggregations) has long been a focus in statistical database research [1, 19, 46, 24, 26]. The proposed techniques can be roughly classified into restriction-based and perturbation-based. The restriction-based techniques limit the disclosure of privacy information by posing restrictions on queries [5, 45, 44], including the number of values aggregated in each query [19], the common values aggregated in different queries [20], and the rank of a matrix representing answered queries [10]. Other restriction-based techniques include partition [9,39], microaggregation [25,46], suppression and generalization [14, 13, 31, 43], and k-anonymity privacy protection [37, 41, 47]. The perturbation-based techniques protect/distort sensitive private data by adding random noises without affecting the use of data significantly. The random noises can be added to data structures [38], query answers [4], or source data [42,2,3,35,8,36]. Recently, however, people have discovered that the original sensitive data can be estimated accurately from the perturbed data [32,30], indicating that the perturbation-based techniques should be examined carefully in practice so as to protect sensitive data effectively.

For protecting contingency tables, people have developed various techniques including cell suppression, controlled rounding, and controlled tabular adjustment. Cell suppression is applied to suppress any sensitive cells as well as other appropriately selected cells so as to prevent inference to sensitive cells from marginal totals [14,17,28,29]. The challenge is to provide sufficient protection while minimizing the amount of information loss due to suppression [27].

Controlled rounding is another disclosure limitation method which rounds each cell value in a contingency table to adjacent integer multiples of a positive integer base [16,15,7]. It requires that the sum of the rounded values for any row or column be equal to the rounded value of the corresponding marginal total. The controlled round can be customized for limiting various types of disclosure.

Controlled tabular adjustment (or synthetic substitution) [18] uses threshold rules to determine how cells should be modified. It replaces a sensitive cell value by a “safe” value (e.g., either zero or a threshold value) and uses linear programming to make small adjustments to other cells so as to restore the tabular structure. Similar to the controlled rounding method, this method requires that some cell values be modified, thus introducing errors to the protected data.

Our study on the disclosure analysis is complementary to the previous study on disclosure limitation. To apply any disclosure limitation method, one needs to first discover all cells that are subject to disclosure. Rather than applying a naive brute-force approach, we investigate the distribution patterns for the cells subject to disclosure and, based on the patterns, propose efficient methods to speedup the searching process significantly.

Parallel to the development of data protection techniques for two-way tables, an active line of research deals with protecting multiway contingency tables or “cubes.” It has been known that the Fréchet bounds, after being extended to high-dimensional space, may not necessarily be the exact bounds [13]. Recent studies have been focused on estimating the exact bounds [13,12,6,33] or giving the exact bounds in some special cases [21, 22, 23]. Once the exact bounds are given, our definitions on various types of disclosure can be easily extended to multiway contingency tables. The challenge is that the distribution patterns discovered in two-way contingency tables may not hold in high dimensions. Therefore, it deserves further study on multiway contingency tables.

7 Conclusion

The major contribution of this paper can be summarized as follows. Firstly, we defined four types of disclosure for evaluating the disclosure of cell values in contingency tables. Secondly, for each type of disclosure, we discovered the distribution patterns for the cells subject to disclosure in a two-way contingency table. The discovery of the distribution patterns is important as it enables us to speed up the search for all cells subject to disclosure. In the future, we plan to extend our study to multiway contingency tables. The major challenge in multiway contingency tables is that the Fréchet bounds may not be exact bounds in general. Some recent efforts have been made to approach the exact bounds beyond two-dimensions [21, 22, 33].

References

1. N. R. Adam and J. C. Wortmann. Security-control methods for statistical databases: a comparative study. *ACM Computing Surveys*, 21(4):515–556, 1989.
2. Dakshi Agrawal and Charu C. Aggarwal. On the design and quantification of privacy preserving data mining algorithms. In *PODS*, 2001.
3. Rakesh Agrawal and Ramakrishnan Srikant. Privacy-preserving data mining. In *SIGMOD Conference*, pages 439–450, 2000.
4. Leland L. Beck. A security mechanism for statistical databases. *ACM Trans. Database Syst.*, 5(3):316–338, 1980.
5. Alexander Brodsky, Csilla Farkas, and Sushil Jajodia. Secure databases: Constraints, inference channels, and monitoring disclosures. *IEEE Trans. Knowl. Data Eng.*, 12(6):900–919, 2000.
6. L. Buzzigoli and A. Giusti. An algorithm to calculate the lower and upper bounds of the elements of an array given its marginals. In *Proceedings of the conference for statistical data protection*, pages 131–147, 1999.
7. B. D. Causey, L. H. Cox, and L. R. Ernst. Applications of transportation theory to statistical problems. *Journal of the American Statistical Association*, 80:903–909, 1985.
8. Keke Chen and Ling Liu. Privacy preserving data classification with rotation perturbation. In *ICDM*, pages 589–592, 2005.
9. Francis Y. L. Chin and Gultekin Özsoyoglu. Statistical database design. *ACM Trans. Database Syst.*, 6(1):113–139, 1981.
10. Francis Y. L. Chin and Gultekin Özsoyoglu. Auditing and inference control in statistical databases. *IEEE Trans. Software Eng.*, 8(6):574–582, 1982.
11. S. Chowdhury, G. Duncan, R. Krishnan, S. Roehrig, and S. Mukherjee. Disclosure detection in multivariate categorical databases: auditing confidentiality protection through two new matrix operators. *Management Sciences*, 45:1710–1723, 1999.
12. L. Cox. Bounding entries in 3-dimensional contingency tables. In *SDC: From Theory to Practice*, 2001. http://vneumann.etse.urv.es/amrads/papers/coxlux.pdf.

13. L. Cox. On properties of multi-dimensional statistical tables. *Journal of Statistical Planning and Inference*, 117(2):251–273, 2003.
14. L. H. Cox. Suppression methodology and statistical disclosure control. *Journal of American Statistical Association*, 75:377–385, 1980.
15. L. H. Cox. A constructive procedure for unbiased controlled rounding. *Journal of the American Statistical Association*, 82:520–524, 1987.
16. L. H. Cox and J. A. George. Controlled rounding for tables with subtotals. *Annuals of operations research*, 20(1-4):141–157, 1989.
17. Lawrence H. Cox. Network models for complementary cell suppression. *Journal of the American Statistical Association*, 90:1453–1462, 1995.
18. R. A. Dandekar and L. H. Cox. Synthetic tabular data: An alternative to complementary cell suppression. Manuscript available from URL http://mysite.verizon.net/vze7w8vk/.
19. D. E. Denning and J. Schlorer. Inference controls for statistical databases. *IEEE Computer*, 16(7):69–82, 1983.
20. David P. Dobkin, Anita K. Jones, and Richard J. Lipton. Secure databases: Protection against user influence. *ACM Trans. Database Syst.*, 4(1):97–106, 1979.
21. A. Dobra and S. E. Fienberg. Bounds for cell entries in contingency tables given fixed marginal totals and decomposable graphs. *Proceedings of the National Academy of Sciences of the United States of America*, 97(22):11885–11892, 2000.
22. A. Dobra and S. E. Fienberg. Bounds for cell entries in contingency tables induced by fixed marginal totals with applications to disclosure limitation. *Statistical journal of the united states*, 18:363–371, 2001.
23. A. Dobra, A. Karr, and A. Sanil. Preserving confidentiality of high-dimensional tabulated data: Statistical and computational issues. *Statistics and Computing*, 13:363–370, 2003.
24. Josep Domingo-Ferrer. Advances in inference control in statistical databases: An overview. In *Inference Control in Statistical Databases*, pages 1–7, 2002.
25. Josep Domingo-Ferrer and Josep Maria Mateo-Sanz. Practical data-oriented microaggregation for statistical disclosure control. *IEEE Trans. Knowl. Data Eng.*, 14(1):189–201, 2002.
26. Csilla Farkas and Sushil Jajodia. The inference problem: A survey. *SIGKDD Explorations*, 4(2):6–11, 2002.
27. M. Fischetti and J. Salazar. Solving the cell suppression problem on tabular data with linear constraints. *Management sciences*, 47(7):1008–1027, 2001.
28. M. Fischetti and J. J. Salazar. Solving the cell suppression problem on tabular data with linear constraints. *Management Sciences*, 47:1008–1026, 2000.
29. M. Fischetti and J. J. Salazar. Partial cell suppression: a new methodology for statistical disclosure control. *Statistics and Computing*, 13:13–21, 2003.
30. Zhengli Huang, Wenliang Du, and Biao Chen. Deriving private information from randomized data. In *SIGMOD Conference*, pages 37–48, 2005.
31. Vijay S. Iyengar. Transforming data to satisfy privacy constraints. In *KDD*, pages 279–288, 2002.
32. Hillol Kargupta, Souptik Datta, Qi Wang, and Krishnamoorthy Sivakumar. On the privacy preserving properties of random data perturbation techniques. In *ICDM*, pages 99–106, 2003.
33. Yingjiu Li, Haibing Lu, and Robert H. Deng. Practical inference control for data cubes (extended abstract). In *IEEE Symposium on Security and Privacy*, 2006.
34. Yingjiu Li, Lingyu Wang, and Sushil Jajodia. Preventing interval-based inference by random data perturbation. In *Privacy Enhancing Technologies*, pages 160–170, 2002.

35. Kun Liu, Hillol Kargupta, and Jessica Ryan. Random projection-based multiplicative data perturbation for privacy preserving distributed data mining. *IEEE Trans. Knowl. Data Eng.*, 18(1):92–106, 2006.
36. K. Muralidhar and R. Sarathy. A general aditive data perturbation method for database security. *Management Sciences*, 45:1399–1415, 2002.
37. P. Samarati and L. Sweeney. *Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression.* Technical report, SRI International. 1998.
38. Jan Schlörer. Security of statistical databases: Multidimensional transformation. *ACM Trans. Database Syst.*, 6(1):95–112, 1981.
39. Jan Schlörer. Information loss in partitioned statistical databases. *Comput. J.*, 26(3):218–223, 1983.
40. Bernd Sturmfels. Week 1: Two-way contingency tables, 2003. John von Neumann Lectures 2003 at the Technical University München. http://www-m10.mathematik.tu-muenchen.de/neumann/lecturenotes/neumann_week1.pdf.
41. L. Sweeney. Achieving k-anonymity privacy protection using generalization and suppression. *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems*, 10(5):571–588, 2002.
42. J. F. Traub, Yechiam Yemini, and Henryk Wozniakowski. The statistical security of a statistical database. *ACM Trans. Database Syst.*, 9(4):672–679, 1984.
43. Ke Wang, Philip S. Yu, and Sourav Chakraborty. Bottom-up generalization: A data mining solution to privacy protection. In *ICDM*, pages 249–256, 2004.
44. Lingyu Wang, Sushil Jajodia, and Duminda Wijesekera. Securing olap data cubes against privacy breaches. In *IEEE Symposium on Security and Privacy*, pages 161–175, 2004.
45. Lingyu Wang, Yingjiu Li, Duminda Wijesekera, and Sushil Jajodia. Precisely answering multi-dimensional range queries without privacy breaches. In *ESORICS*, pages 100–115, 2003.
46. L. Willenborg and T. de Walal. *Statistical Disclosure Control in Practice.* Springer Verlag, 1996.
47. Chao Yao, Xiaoyang Sean Wang, and Sushil Jajodia. Checking for k-anonymity violation by views. In *VLDB*, pages 910–921, 2005.

Statistical Disclosure Control Methods Through a Risk-Utility Framework

Natalie Shlomo[1] and Caroline Young[2]

[1] Southampton Statistical Sciences Research Institute, University of Southampton, UK and the Department of Statistics, Hebrew University, Mt. Scopus, Jerusalem, Israel
[2] School of Social Sciences, University of Southampton, Southampton, UK and the Office for National Statistics, Segensworth Road, Fareham, UK
{N.Shlomo, CJY}@soton.ac.uk

Abstract. This paper discusses a disclosure risk – data utility framework for assessing statistical disclosure control (SDC) methods on statistical data. Disclosure risk is defined in terms of identifying individuals in small cells in the data which then leads to attribute disclosure of other sensitive variables. Information Loss measures are defined for assessing the impact of the SDC method on the utility of the data and its effects when carrying out standard statistical analysis tools. The quantitative disclosure risk and information loss measures can be plotted onto an R-U confidentiality map for determining optimal SDC methods. A user-friendly software application has been developed and implemented at the UK Office for National Statistics (ONS) to enable data suppliers to compare original and disclosure controlled statistical data and to make informed decisions on best methods for protecting their statistical data.

Keywords: Disclosure risk, Information loss, R-U Confidentiality Map.

1 Introduction

In order to preserve the privacy and confidentiality of statistical units, data suppliers need to assess the disclosure risk in statistical outputs and if required choose appropriate Statistical Disclosure Control (SDC) methods to apply to the data. The most common forms of statistical outputs are tables containing frequency counts or aggregates (for example, total turnover) and microdata from social surveys usually released under special licence agreements. In the future, more flexibility in statistical outputs is envisioned through web-based user defined table generating software and online query systems. This will require more research and development for online applications of SDC methods. Statistical outputs can contain whole population counts from Censuses, administrative data and registers or weighted counts from survey samples. Census outputs are harder to protect against disclosure risk than outputs derived from sample data since the sampling mechanism introduces ambiguity into the counts.

SDC methods perturb, modify, or summarize the data depending on the format for releasing the statistical outputs. Higher levels of protection through SDC methods will impact negatively on the utility and quality of the data. Therefore, choosing optimal SDC methods involves solving a decision problem where a balance is sought between

J. Domingo-Ferrer and L. Franconi (Eds.): PSD 2006, LNCS 4302, pp. 68 – 81, 2006.

managing and minimizing disclosure risk to tolerable risk thresholds and maintaining utility and quality in the statistical data. A third dimension is sometimes added to the SDC problem which involves the feasibility of implementing SDC methods in a production line.

Examples of common SDC methods for statistical data are:

- Pre-tabular methods (implemented on microdata prior to its tabulation or when releasing sample microdata) such as recoding, coarsening and eliminating variables, sub-sampling, micro-aggregation, record swapping or other probabilistic perturbation processes,
- Post-tabular methods (implemented on the tables themselves) such as table redesign (coarsening and recoding), cell suppression, rounding or other perturbation processes.

Note that some methods provide protection against disclosure risk by limiting the amount of information that is released (non-perturbative methods) while other methods actually alter the data (perturbative methods).

Measuring disclosure risk for the SDC decision problem involves assessing and evaluating numerically the risk of re-identifying statistical units. For microdata, disclosure risk arises from attribute disclosure where small counts on cross-classified indirectly identifying key variables (such as: age, sex, place of residence, marital status, occupation, etc.) can be used to identify an individual and further confidential information may be learnt. For tabular data, disclosure risk arises from small counts in the tables, the position of the zeros and whether they are structural or random zeros. In addition, when multiple tables are generated from one dataset (such as in a Census context) this raises the risk of being able to reveal original counts protected by SDC methods through linking and differencing multiple tables. Also, releasing many tables from one dataset increases the risk of attribute disclosure from an identification of a unique cell in another table.

Measuring information loss and utility for the SDC decision problem is more subjective. This depends on the users, the purpose of the data and the required statistical analysis, and the type and format of the statistical data. Therefore it is useful to have a wide range of information loss measures for assessing the impact of the SDC methods on the statistical data. These measures include:

- effects on the bias and variance of point estimates and other sufficient statistics,
- distortions to the rankings of variables, and univariate and joint distributions between variables,
- changes to model parameters and goodness of fit criteria when carrying out statistical analysis.

When assessing SDC methods and their parameters for statistical outputs, an iterative process is carried out. For each method and its parameters, quantitative disclosure risk and information loss measures are calculated. These points can then be plotted on a Disclosure Risk - Data Utility (R-U) Confidentiality Map (see Duncan, et.al., 2001). An optimal SDC method is chosen which reduces the disclosure risk to tolerable risk thresholds while ensuring high quality data that is fit for purpose.

Information loss measures can be classed into two research areas: information loss measures for data suppliers in order to enable informed decisions about optimal SDC methods, and information loss measures for users in order to enable adjustments to statistical analysis on modified disclosure controlled statistical data. For example, indication of the extra variance that is added to a variable due to a stochastic rounding procedure would allow users to adjust their model parameters when carrying out a regression analysis. The difference between the two classes of information loss measures is that users do not have access to the original data for comparisons. Therefore it is the responsibility of data suppliers to inform users of the quality of the disclosure controlled data and the impact on statistical analysis without revealing confidential parameters which may be used to reveal original values. In this paper, we focus only on information loss measures for data suppliers who have access to the original statistical data and can use them to make judgements on optimal SDC strategies.

In the United Kingdom, a key mechanism for disseminating statistics for small areas (usually tables containing Census and administrative data) is the Neighbourhood Statistics website (NeSS). Data suppliers are both internal to the Office for National Statistics (ONS) and external in other departments of the Government Statistical Service. In order to allow data suppliers to make informed decisions about optimal disclosure control methods, we have developed a user-friendly software application that calculates both a simple disclosure risk measure and a wide range of information loss measures for disclosure controlled statistical data. The software application also outputs R-U Confidentiality Maps.

Section 2 outlines information loss measures for data that have undergone SDC methods. They include distance metrics which assess distortions to distributions on internal counts and marginal totals, the impact on statistical inference based on the variance, goodness of fit criteria for statistical modelling, and the order and rankings of the counts. Section 3 describes a basic disclosure risk measure that is used for whole population counts and Section 4 an example of the measures on a table from the 2001 UK Census. Section 5 provides a brief overview of the software application that has been developed at the UK ONS. Section 6 concludes with a discussion and future developments.

2 Information Loss Measures

The initial focus for developing information loss measures was on tables containing frequency counts. The information loss measures are easily adapted to microdata since these are frequency tables with cells of size one. Moreover, when assessing the impact of SDC methods on microdata it is useful to carry out the analysis by tabulations and examining univariate and joint distributions. Magnitude or weighted sample tables have an additional element which is the number of contributors to each cell of the table. In the following descriptions of the information loss measures we refer to the statistical data as a table.

2.1 Basic Statistics and Information About the Statistical Data

Summary statistics include the number of rows and columns, number of small cells and zeros, the total number of contributing units to the table and the total information content in the table (for a frequency table, the total information is the number of contributors and for a magnitude or weighted sample data, the total information is obtained by summing the aggregates of the cell values). Average, minimum and maximum cell sizes and their standard errors are presented for the whole table, columns and rows. These measures give an indication of the skewness and sparsity of the table.

For a table that has been suppressed, the number of suppressed cells (internal cells and marginal totals) and the number of contributors and total information that have been suppressed are calculated. In order to assess utility, the data supplier needs to decide on an imputation method for replacing suppressed cells similar to what one would expect a user to do prior to analyzing the data. A naive user might enter zeros in place of the suppressed cells whereas a more sophisticated user might replace suppressed cells by some form of averaging of the total information that was suppressed in a row (or a column by transposing the table).

In this application we carry out three common types of imputation:

- Replace each suppressed cell by a zero
- For each row, replace suppressed cells by the simple average of the total information suppressed: Let m_{kj} be a cell count in a two way table $k = 1,...,K$ rows and $j = 1,...J$ columns. Let marginal totals be defined as: $m_{k.}$ and $m_{.j}$. The margins appear in the table without perturbation unless they have a small value and are primary suppressed. In that case, we define the margin to take a value of 1 for the following imputation schemes. Let z_{kj} be an indicator taking on the value of 1 if the cell was suppressed (primary or secondary) and a 0 otherwise. Each suppressed cell in row k is replaced by the average of the total information that was suppressed, i.e. $\dfrac{m_{k.} - \sum_{j=1}^{J} m_{kj}(1 - z_{kj})}{\sum_{j=1}^{J} z_{kj}}$. Note that the row total is preserved.

  Example: Two cells are suppressed in a row where the known marginal total is 500. The total obtained by adding up non-suppressed cells is 400, and therefore the total loss of information in the row is 100. Each of the 2 suppressed cells is replaced with a value of 50.
- For each row, replace suppressed cells by the weighted average of the total information suppressed where the weights are obtained by the unsuppressed column averages. In other words, the weights are calculated as the average cell size of the columns j: $w_j = \dfrac{m_{.j}}{J}$. A low frequency column will result in a smaller imputed cell frequency and a high frequency column will result in a larger imputed cell frequency. Each suppressed cell in row k is replaced by: $\dfrac{w_j \times (m_{k.} - \sum_{j=1}^{J} m_{kj}(1 - z_{kj}))}{\sum_{j=1}^{J} w_j z_{kj}}$. Note that the row total is preserved.

Example: Based on the above example, the column average for Cell 1 is 50 and the column average for Cell 2 is 10. Cell 1 is imputed by: 100*50/60 =83.33 and Cell 2 is imputed by 100*10/60=16.67.

2.2 Statistical Hypothesis Tests for Bias

We carry out an exact Binomial Hypothesis Test to check if the realization of a random stochastic perturbation scheme follows the expected probabilities. For example, for a random rounding to base 3, the null hypothesis is: $H_0 : p = 2/3$. The test is carried out using a PROC FREQ SAS procedure. Small p-values mean that we reject the null hypothesis and the stochastic procedure is biased.

For other SDC methods, we use a non-parametric signed rank test in the PROC UNIVARIATE SAS procedure to check whether the location of the empirical distribution has changed. The null hypothesis for the test is no change. The test statistic is based on the rankings of the original minus perturbed cells. If there is a large deviation (small p-value), then the location of the distribution has shifted.

2.3 Distance Metrics

We calculate distance metrics between original and disclosure controlled internal cells of a distribution in a table k. When combining several tables we may want to calculate an overall average distance metric across the different tables. This format is particularly useful in the case of Census or Register based tables where the rows represent a geographical area and the columns define the categories of a specific table or distribution. For example, each row of the Employment table as described in Section 4 is a geographical area within which a separate 3-dimensional table is defined by cross classifying sex, long term illness and primary economic activity.

Let D^k represent a row (i.e., a distribution) k in a table and let $D^k(c)$ be the cell frequency c in the row. Let n_r be the number of rows in the comparison. The distance metrics are:

- Hellinger's Distance:

$$HD(D_{pert}, D_{orig}) = \frac{1}{n_r}\sum_{k=1}^{n_r}\sqrt{\sum_{c\in k}\frac{1}{2}\left(\sqrt{D_{pert}^k(c)} - \sqrt{D_{orig}^k(c)}\right)^2}$$

- Relative Absolute Distance:

$$RAD(D_{pert}, D_{orig}) = \frac{1}{n_r}\sum_{k=1}^{n_r}\sum_{c\in k}\frac{|D_{pert}^k(c) - D_{orig}^k(c)|}{D_{orig}^k(c)}$$

- Average Absolute Distance per Cell:

$$AAD(D_{pert}, D_{orig}) = \frac{1}{n_r}\sum_{k=1}^{n_r}\frac{\sum_{c\in k}|D_{pert}^k(c) - D_{orig}^k(c)|}{n_k}$$ where

$n_k = \sum_c I(c \in k)$ the number of cells in the k^{th} table.

These distance metrics can also be calculated for totals or sub-totals of the tables:

- Relative Absolute Distance for a Total:

$$RAT(N_{pert},N_{orig})=\sum_{k=1}^{n_r}\frac{|N^k_{pert}(C')-N^k_{orig}(C')|}{N^k_{orig}(C')}$$

- Average Absolute Distance per Total:

$$AT(N_{pert},N_{orig})=\sum_{k=1}^{n_r}\frac{|N^k_{pert}(C')-N^k_{orig}(C')|}{n_r}$$

where $N^k(C')=\sum_{c\in C'}D^k(c)$ is a sub- total for group C'.

2.4 Variance of the Cell Counts

We examine the variance of the cell counts for each row (distribution) k in the table and take the average across all of the rows before and after applying SDC methods as follows: For each row k, we calculate: $V(D^k_{orig})=\frac{1}{n_k-1}\sum_{c\in k}(D^k_{orig}(c)-\overline{D}^k_{orig})^2$ where $\overline{D}^k_{orig}=\frac{\sum_{c\in k}D^k_{orig}(c)}{n_k}$ and $n_k=\sum_c I(c\in k)$ the number of cells in the k^{th} row. Next we calculate the average of the variances for the original table: $AV(D_{orig})=\frac{1}{n_r}\sum_{k=1}^{n_r}V(D^k_{orig})$ as well as the perturbed table. The final information loss measure is: $RAV(D_{pert},D_{orig})=100\times\frac{AV(D_{pert})-AV(D_{orig})}{AV(D_{orig})}$.

2.5 Impact on Measures of Association

Another statistical analysis that is frequently carried out on tabular data are tests for independence between categorical variables that span a table. The test for independence for a two-way table is based on a Pearson Chi-Squared Statistic $\chi^2=\sum_i\sum_j\frac{(o_{ij}-e_{ij})^2}{e_{ij}}$ where o_{ij} is the observed count and $e_{ij}=\frac{n_{i.}\times n_{.j}}{n}$ is the expected count for row *i* and column *j*. If the row and column are independent then χ^2 has an asymptotic chi-square distribution with (R-1)(C-1)and for large values the test rejects the null hypothesis in favour of the alternative hypothesis of association. We use the measure of association, Cramer's V:

$CV = \sqrt{\frac{\chi^2 / n}{\min(R-1),(C-1)}}$ and define the information loss measure by the percent relative difference between the original and perturbed table: $RCV(D_{pert}, D_{orig}) = 100 \times \frac{CV(D_{pert}) - CV(D_{orig})}{CV(D_{orig})}$.

We can calculate the same for the Pearson Statistic.

The entropy for a row k in a table is calculated as follows: $E(D^k_{orig}) = \frac{D^k_{orig}}{\sum\limits_{c \in k} D^k_{orig}} \log(\frac{D^k_{orig}}{\sum\limits_{c \in k} D^k_{orig}})$. The total entropy across all the rows of a table is: $ENT(D_{orig}) = \sum\limits_{k=1}^{n_r} E(D^k_{orig})$. The information loss measure is defined as:

$$ENTT(D_{pert}, D_{orig}) = 100 \times \frac{ENT(D_{pert}) - ENT(D_{orig})}{ENT(D_{orig})}.$$

2.6 Impact on Rank Correlation

One statistical tool for inference is the Spearman's Rank Correlation. This is a non-parametric statistic that tests the direction and strength of a relationship between two variables. The statistic is based on ranking both variables from the highest to the lowest and testing for association. There are several other statistical analysis tools which are based on rankings and empirical distribution functions. Therefore, an important assessment on the impact of an SDC method is to determine how much the method distorts the rankings of the variables.

We sort the original cell values according to their size and define groupings of size 10 or size 20 denoted by $v^{orig}(c)$. This is repeated for the disclosure controlled cell values which are sorted according to both their size and the original order in order to maintain consistency for the tied variables. The groupings for the disclosure controlled cells c are denoted by $v^{pert}(c)$. The information loss measure is the percent of cells that have changed groupings: $RC = \frac{100 \times \sum\limits_{c \in k} I(v^{orig}_k(c) \neq v^{pert}_k(c))}{n_k}$ where I is the indicator function and is 1 if the statement is true and 0 otherwise, and n_k is the number of cells.

2.7 Impact on a Regression Analysis

For continuous variables, we assess the impact on the correlation and goodness of fit criterion R^2 of a regression (or ANOVA) analysis. For example, in an ANOVA, we test whether a continuous dependent variable has the same means within groupings defined by categorical explanatory variables. The goodness of fit criterion R^2 is based on a decomposition of the variance of the mean of the dependent variable. The total sum of squares *SST* can be broken down into two components: the "within" sum of squares *SSW* which measures the variance of the mean of the target variable within groupings which are defined by combining explanatory variables and the "between" sum of squares *SSB* which measures the variance of the mean of the target variable between the groupings. R^2 is the ratio of *SSB* to *SST*. By perturbing the statistical data, the groupings may lose their homogeneity, *SSB* becomes smaller, and *SSW* becomes larger. In other words, the proportions within each of the groupings shrink towards the overall mean. On the other hand, *SSB* may become artificially larger showing more association within the groupings than in the original variable.

We define information loss based on the "between" variance of a proportion: Let $P^k_{orig}(c)$ be a target proportion for a cell *c* in row *k*, i.e. $P^k_{orig}(c) = \frac{D^k_{orig}(c)}{\sum_{c \in k} D^k_{orig}(c)}$

and let $P^k_{orig}(c) = \frac{\sum_{k=1}^{n_r} D^k_{orig}(c)}{\sum_{k=1}^{n_r} \sum_{c \in k} D^k_{orig}(c)}$ be the overall proportion across all the rows of the table. The "between" variance is defined as: $BV(P_{orig}) = \frac{1}{n_r - 1} \sum_{k=1}^{n_r} (P^k_{orig}(c) - P_{orig})^2$ and the information loss measure is:

$$BVR(P_{pert}, P_{orig}) = 100 \times \frac{BV(P_{pert}) - BV(P_{orig})}{BV(P_{orig})} .$$

2.8 Impact on Goodness of Fit Criterion for a Log Linear Model

Another type of statistical analysis frequently carried out on a complete contingency table is log linear modelling. For a 2-way table this narrows down to the test for independence and the Cramer's V statistic as described in Section 2.5. For more variables in a contingency table, one can examine conditional dependencies and calculate expected cell frequencies based on the theory of log-linear models. The goodness of fit test for assessing the best fitting parsimonious model is the deviance or log-likelihood ratio L^2. This is the statistic that is minimized when calculating maximum likelihood estimates for the parameters of the model. The information loss measure is defined as the ratio of the deviance between the disclosure controlled table and the original table for a given log-linear model as specified by the data supplier: $LR_m = \frac{L^2_{m,pert}}{L^2_{m,orig}}$.

3 Disclosure Risk Measures

When statistical data contain whole population counts, disclosure risk can be assessed by counting the number of small cells and calculating the proportion of those cells that were targeted for disclosure control, i.e., the probability that a small value in a cell of the table is the true value. Other methods for assessing disclosure risk for whole population counts include probabilistic record linkage techniques where protected datasets are matched back to the original datasets and the disclosure risk measure is based on the proportion of correct matches (Yancy, et.al. 2002). These probabilities can also be obtained using statistical models taking into account misclassification and perturbation probabilities.

When the statistical data contain sample counts, global file-level disclosure risk measures are typically defined as follows: the number of sample uniques that are population uniques on a set of cross-classified indirectly identifying key variables (i.e., a key) of size $k = 1,...,K$: $\sum_{k=1}^{K} I(F_k = 1, f_k = 1)$ where F_k is the population size in cell k of the key, f_k is the sample size and I is the indicator function obtaining a value of 1 if the statement is true and 0 if not; or the expected number of correct matches of the sample uniques to a population: $\sum_{k=1}^{K} I(f_k = 1)\frac{1}{F_k}$.

If the population from which the sample is drawn is known, these measures can be calculated directly. If the population is unknown, we use sophisticated statistical modelling techniques to estimate the disclosure risk measures by inferring from the sample counts in the contingency table spanned by the key variables.

4 Example on a Census Table

The information loss and disclosure risk measures will be calculated and compared for a table from the 2001 UK Census containing employed individuals between the ages of 16 and 74. The rows consist of small geographical areas, Output Areas (OA) in one Estimation Area of the UK (1,472 OAs), each OA with an average size of about 215 individuals between the ages of 16 and 74. The columns consist of the internal cells for a cross classified variable defined by sex (2), long-term illness (2) and primary economic activity (9). To demonstrate the comparison of SDC methods, the table underwent two methods of disclosure control:

- Semi-controlled random rounding to base 3 – Each cell is stochastically rounded as follows: counts with a residual of one from the base are rounded up to the nearest base with a probability of 1/3 and down with a probability of 2/3. Counts with a residual of two from the base are rounded up to the nearest base with a probability of 2/3 and rounded down with a probability of 1/3. The process is semi-controlled which means that the expected number of cells to be rounded up are selected without replacement from the table and only those cells are rounded up, the remaining cells are rounded down. This ensures that there is no bias in the total since the sum of the perturbations equals zero.

- 10% random record swapping - The microdata underlying the table is perturbed by random record swapping as follows: within each broad geographical area, a 10% random selection of households are paired with other households having the same household size (households of 8 persons and over are banded), broad age sex distribution and a "hard-to-count" index. The geographical variables are swapped between the households.

Table 1 presents results for the measures that were detailed in Sections 2 and 3 for the Census employment table. Note that these measures were obtained as output from the SDC software application described in Section 5.

As observed in Table 1, the two SDC methods behave quite differently on the statistical data in the Census employment table. The main conclusions are:

- Disclosure risk is much greater for the 10% random record swapping with only 38.4% of the small cells actually being targeted for disclosure control. Rounding on the other hand masks all of the small cells.
- Because of the benchmarking and the control placed on the totals for the rounding and swapping procedures, these are not distorted. This is evident also in the results of the statistical tests for bias. On the other hand, the rounding had greater distance metrics between the original and protected distributions of the internal cells than the record swapping.
- Rounding introduces more association between the variables (by placing more zeros in the table), as can be seen by the variance of the cell counts, the between variance, the impact on the deviance of the log linear model and Cramer's V. Those measures are all positive or greater than one.
- Record swapping attenuates the relationship between variables. This is seen in the flattening of the cells counts through the variance, the between variance, and less association in Cramer's V and the deviance of the log linear model. Those measures are all negative or less than one.
- Because the table is sparse, the impact on the rankings of the cell counts was great. Every column in the employment table had more than 20% movements between the 20 groupings after carrying out the record swapping, where only 25 (69%) of the columns were affected by the rounding.

Based on the results of Table 1, the needs of the users and the purpose of the data, data suppliers need to choose which SDC method is preferable. R-U confidentiality maps for comparing SDC methods may be useful for this purpose and these are outputted in the software application described in Section 5. For example, if we compared the SDC methods for the employment table on an R-U confidentiality map where the risk measure is the percent of small cells that are unprotected and the utility measures is the HD distance metric, then the optimal SDC method would depend on the tolerable risk threshold set by the data supplier. If the tolerable risk threshold is above 65% then record swapping with the least distortion to the distribution of internal cells would be preferred. However if the tolerable risk threshold is below 65%, then rounding would be preferred.

Table 1. Results of a Disclosure Risk-Data Utility Analysis for two SDC Methods on the 2001 UK Census Employment Table in an Estimation Area

Summary Statistics		
Number of cells	52,992	
Number of small cells (one,twos)	14,684 (27.7%)	
Number of zeros	17,413 (32.9%)	
Total information (no. of individuals)	317,064	
Average cell size	6.0 (± 0.12)	
Range of average cell size in row	1.0 – 13.2	
Range of average cell size in column	0.1 – 67.1	
	10% Random Record Swap across Geography	**Semi-Controlled Random Rounding to Base 3**
Disclosure Risk Measure		
Small numbers changed	5,713	14,686
Small numbers not changed	8,973	0
Percent small numbers not changed	61.1%	0%
Information Loss Measures		
Statistical tests for bias (p-values)	0.338	0.494
Distance metrics		
Internal cells AAD	0.662	0.678
HD	1.405	2.043
Totals AAD	0.001	0.003
Change in average variance RAV	-1.31%	0.51%
Change in Cramer's V RCV	-3.66%	11.66%
No. of columns with more than 20% movements between groupings of 20	36 (100.0%)	25 (69.4%)
Ratio of between variance BVR (proportion of males with no long term illness who are retired)	0.959	1.061
Ratio of deviance LR (model: 2 interactions: sex*OA , long term illness*economic activity)	0.963	1.172

5 Software Application for the SDC Problem

5.1 Brief Overview

The software application is written in SAS and computes information loss measures by comparing two output tables (or microdata), the original output and the disclosure controlled output. The program has been designed to deal with tables that have been protected with pre-tabular methods or post-tabular methods such as: cell suppression, any form of rounding and small cell adjustments. When the tables are protected using cell suppression, the data supplier must define a method for imputing suppressed cells as described in Section 2.1.

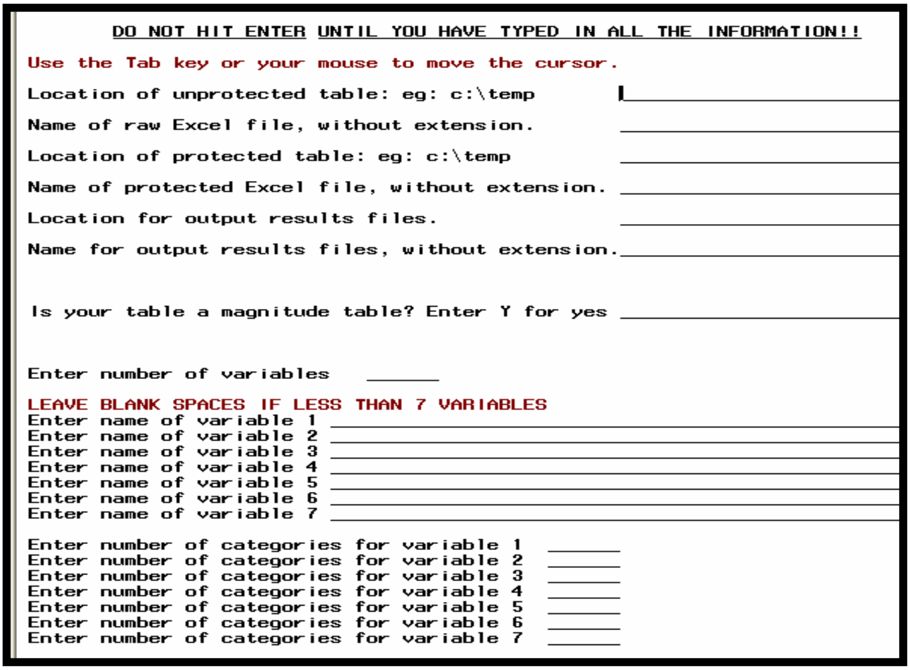
5.2 Preparing Input and Running the Program

Both the original and the protected outputs are imported from Excel files in the application. Data suppliers need to prepare the tables in a standard format, leaving only internal cells for the original table and the internal cells and margins for the disclosure controlled table. This is because margins in disclosure control tables can be protected separately from the internal cells and tables may not be additive. When defining the variables that span the tables, data suppliers must understand the hierarchies and input the variables in that order, starting with the column variables and then the row variables.

The program is run through a batch file. Users do not need to know how to use SAS to be able to run the program. Error messages pop up at all stages of the program to inform the data supplier if an error has been detected in one of the procedures. For example, if the tables are incorrectly imported into SAS from Excel, an error message will pop up immediately specifying the problem causing the error and the program will be stopped.

5.3 The Windows of the Program

Windows pop up during the course of running the program where the data suppliers are asked to fill in details. The first window asks about the location of the datasets and the variables of the table.



DO NOT HIT ENTER UNTIL YOU HAVE TYPED IN ALL THE INFORMATION!!

Use the Tab key or your mouse to move the cursor.

Location of unprotected table: eg: c:\temp

Name of raw Excel file, without extension.

Location of protected table: eg: c:\temp

Name of protected Excel file, without extension.

Location for output results files.

Name for output results files, without extension.

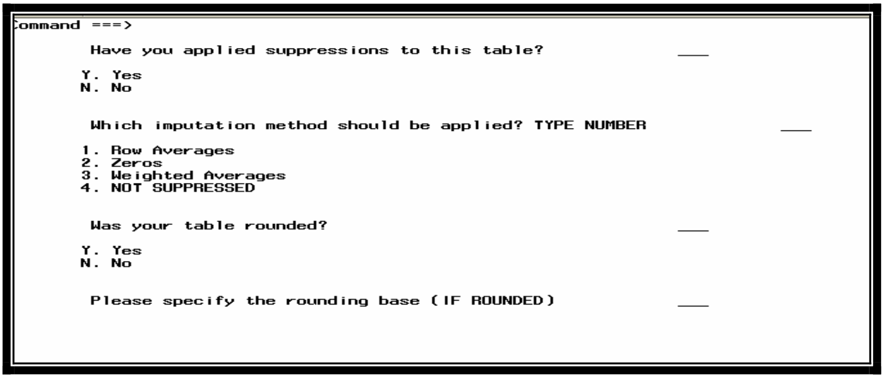
Is your table a magnitude table? Enter Y for yes

Enter number of variables

LEAVE BLANK SPACES IF LESS THAN 7 VARIABLES
Enter name of variable 1
Enter name of variable 2
Enter name of variable 3
Enter name of variable 4
Enter name of variable 5
Enter name of variable 6
Enter name of variable 7

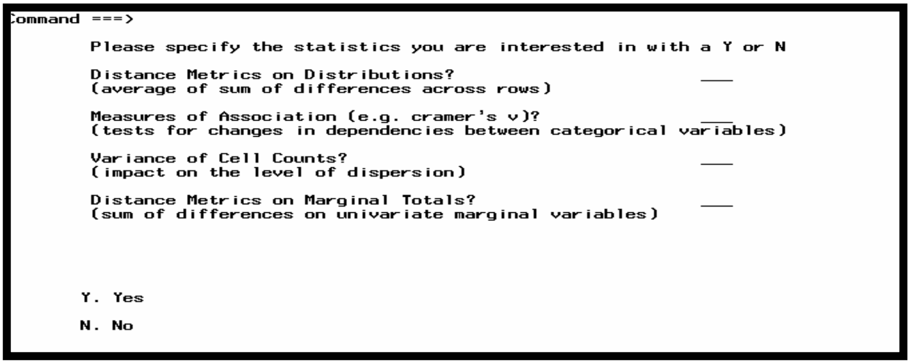
Enter number of categories for variable 1
Enter number of categories for variable 2
Enter number of categories for variable 3
Enter number of categories for variable 4
Enter number of categories for variable 5
Enter number of categories for variable 6
Enter number of categories for variable 7

The second window will open a dialogue and ask if the table was protected by cell suppression and to define the imputation method, or rounding and to specify the rounding base.

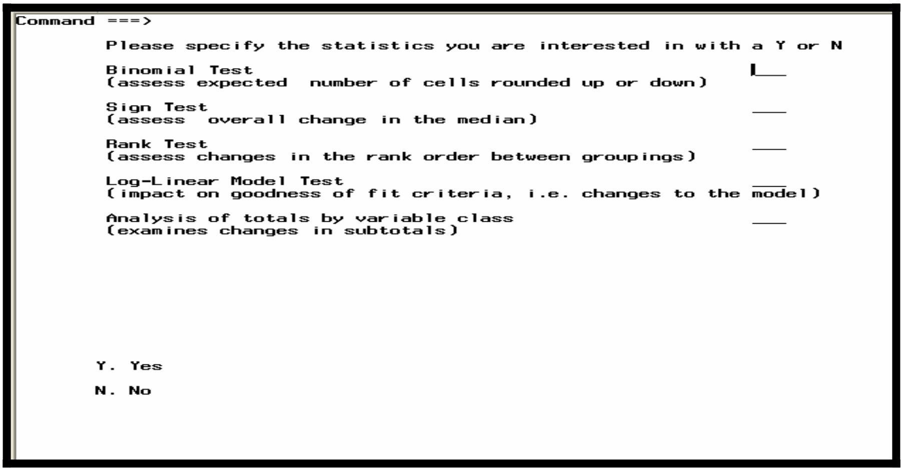


```
Command ===>

       Have you applied suppressions to this table?              ___

      Y. Yes
      N. No

       Which imputation method should be applied? TYPE NUMBER              ___

      1. Row Averages
      2. Zeros
      3. Weighted Averages
      4. NOT SUPPRESSED

       Was your table rounded?                                   ___

      Y. Yes
      N. No

       Please specify the rounding base (IF ROUNDED)             ___
```

The next windows that pop up ask which information loss measures will be calculated, what are the variables that will undergo analysis and whether to calculate disclosure risk measures. Note that although the data is inputted as a two-dimensional table, data suppliers define flexible distributions and choose the variables that they are interested in. For example, the windows that pop up for the information loss measures are the following:



```
Command ===>

       Please specify the statistics you are interested in with a Y or N

       Distance Metrics on Distributions?                          ___
       (average of sum of differences across rows)

       Measures of Association (e.g. cramer's v)?                  ___
       (tests for changes in dependencies between categorical variables)

       Variance of Cell Counts?                                    ___
       (impact on the level of dispersion)

       Distance Metrics on Marginal Totals?                        ___
       (sum of differences on univariate marginal variables)



      Y. Yes

      N. No
```



```
Command ===>

       Please specify the statistics you are interested in with a Y or N

       Binomial Test                                               ___
       (assess expected  number of cells rounded up or down)

       Sign Test                                                   ___
       (assess  overall change in the median)

       Rank Test                                                   ___
       (assess changes in the rank order between groupings)

       Log-Linear Model Test                                       ___
       (impact on goodness of fit criteria, i.e. changes to the model)

       Analysis of totals by variable class                        ___
       (examines changes in subtotals)



      Y. Yes

      N. No
```

If all the information has been entered correctly, a final window will appear to inform the data supplier that the program has ended. An 'html' file is produced under the name that was specified by the user in the first window. In addition, the program outputs R-U confidentiality maps to provide data suppliers with a visual tool for choosing optimal SDC methods.

6 Future Developments

The software application has completed its first phase of development and is undergoing testing. New information loss measures and the probabilistic disclosure risk assessment for sample data with unknown populations will be included in the next version. More flexibility will be introduced into the program with respect to different rounding bases, choosing which measures to plot on the R-U confidentiality map and more.

As mentioned, the SDC software tool is designed for data suppliers who have access to the original data and want to make informed decisions on best SDC methods for their statistical data. However, the key next stage of this project is to develop and disseminate information loss measures for the users who do not have access to the original data. These information loss measures should allow the users to take into account the impact of the SDC methods on the statistical data and to make adjustments when carrying out statistical analysis and inference.

References

1. Duncan, G., Keller-McNulty, S., and Stokes, S.: Disclosure Risk vs. Data Utility: the R-U Confidentiality Map, Technical Report LA-UR-01-6428, Statistical Sciences Group,Los Alamos, N.M.:Los Alamos National Laboratory (2001)
2. Gomatan, S. and Karr, A.: Distortion Measures for Categorical Data Swapping, Technical Report Number 131, National Institute of Statistical Sciences (2003)
3. Yancey, W., Winkler, W., and Creecy, R.: Disclosure Risk Assessment in Perturbative Microdata Protection," in (J. Domingo-Ferrer, ed.) Inference Control in Statistical Databases, New York: Springer, (2002) 135-151.

A Generalized Negative Binomial Smoothing Model for Sample Disclosure Risk Estimation

Yosef Rinott[1] and Natalie Shlomo[2,⋆]

[1] The Hebrew University of Jerusalem, Israel
[2] The Hebrew University of Jerusalem, Israel and Southampton Statistical Sciences Research Institute, University of Southampton, UK

Abstract. We deal with the issue of risk estimation in a sample frequency table to be released by an agency. Risk arises from non-empty sample cells which represent small population cells and from population uniques in particular. Therefore risk estimation requires assessing which of the relevant population cells are indeed small. Various methods have been proposed for this task, and we present a new method in which estimation of a population cell frequency is based on smoothing using a local neighborhood of this cell, that is, cells having similar or close values in all attributes.

The statistical model we use is a *generalized Negative Binomial* model which subsumes the Poisson and Negative Binomial models. We provide some preliminary results and experiments with this method.

Comparisons of the new approach are made to a method based on *Poisson regression log-linear hierarchical model*, in which inference on a given cell is based on classical models of contingency tables. Such models connect each cell to a 'neighborhood' of cells with one or several common attributes, but some other attributes may differ significantly. We also compare to the *Argus* Negative Binomial method in which inference on a given cell is based only on sampling weights, without learning from any type of 'neighborhood' of the given cell and without making use of the structure of the table.

1 Introduction

Let $\mathbf{f} = \{f_k\}$ denote an m-way frequency table, which is a sample from a population table $\mathbf{F} = \{F_k\}$, where $k = (k_1, ..., k_m)$ indicates a cell, and f_k and F_k denote the frequency in cell k in the sample and population, respectively, and the number of cells is denoted by K. Formally, the sample and population sizes in our models are random and their expectations are denoted by n and N respectively. We formally assume that n and N are known, but in practice they are usually replaced by their natural estimators: the actual sample and population sizes, assumed to be known, and without further comment.

[⋆] Both authors' research was partially supported by the Israel Science Foundation grant No. 473/04.

J. Domingo-Ferrer and L. Franconi (Eds.): PSD 2006, LNCS 4302, pp. 82–93, 2006.

The m attributes in the table are considered to be *key variables*, that is, variables which are to some extent accessible to the public or to potential intruders. Disclosure risk arises from cells in which both f_k and F_k are positive and small, and in particular when $f_k = F_k = 1$ (sample and population *uniques*). An intruder who locates a sample unique in cell k, say, and is aware of the fact that in the population the combination of values $k = (k_1, ..., k_m)$ is unique ($F_k = 1$) or rare (F_k small) but matches an individual of interest, can identify this individual on the basis of these m attributes. If the sample contains information on the values of other attributes, then these can now be inferred for the individual in question, and his privacy is violated.

Individual risk measures will be briefly discussed in Section 2 and we start with *global risk measures* which quantify an aspect of the total risk in the file by aggregating risk over the individual cells. For simplicity we shall focus here only on two global measure, which are based on sample uniques:

$$\tau_1 = \sum_k \mathbf{I}(f_k = 1, F_k = 1)\,, \qquad \tau_2 = \sum_k \mathbf{I}(f_k = 1)\frac{1}{F_k}\,,$$

where $\mathbf{I}$ denotes the indicator function. Note that τ_1 counts the number of *sample uniques* which are also *population uniques*, and τ_2 is the expected number of correct guesses if each sample unique is matched to a randomly chosen individual from the same population cell. These measures are somewhat arbitrary, and one could consider measures which reflect matching of individuals that are not sample uniques, possibly with some restrictions on cell sizes. Also, it may make sense to normalize these measures by some measure of the total size of the table, by the number of sample uniques, or by some measure of the information value of the data.

Various individual and global risk measures have been proposed in the literature, see e.g. Franconi, and Polettini (2004) and references therein, Skinner and Holmes (1998), Elamir and Skinner (2006), Rinott (2003).

In Sections 2 and 3 we propose and explain a new method of estimation of quantities like τ_1 and τ_2, using a *generalized Negative Binomial model*, and *local smoothing* of frequency tables, Simonoff (1998). The method is based on the idea that one can learn about a given population cell from neighboring cells, if a suitable definition of closeness or neighbors is possible, by standard smoothing techniques, without relying on complex dependence structure modeling. This method differs from that of Elamir and Skinner (2006), in which one uses classical hierarchical log-linear models, which means inferring on a given cell by using cells that could be very different in many attribute values. For example, in the independence model, inference on a cell uses all cells which have at least one common attribute with the given cell, but all others may be very different. Thus neighborhoods formed by the classical log-linear model theory seem to be too large for our purposes. This point is explained in detail in Rinott and Shlomo (2005). On the other hand, the Argus approach, see, e.g., Franconi, and Polettini (2004), uses no neighborhoods at all and ignores the table structure. We consider the smoothing approach simple conceptually but not necessarily in terms of the computations required.

In this paper it is assumed that $\mathbf{f}$ is known, and $\mathbf{F}$ is an unknown parameter (on which there may be some partial information) and the quantities τ_1 and τ_2 should be estimated. Note that they are not proper parameters, since they involve both the sample $\mathbf{f}$ and the parameter $\mathbf{F}$.

The methods discussed in this paper consist of modeling the conditional distribution of $\mathbf{F}|\mathbf{f}$, estimating parameters in this distribution and then using estimates of the form

$$\hat{\tau}_1 = \sum_k \mathbf{I}(f_k = 1)\hat{P}(F_k = 1|f_k = 1), \quad \hat{\tau}_2 = \sum_k \mathbf{I}(f_k = 1)\hat{E}[\frac{1}{F_k}|f_k = 1], \tag{1}$$

where $\hat{P}$ and $\hat{E}$ denote estimates of the relevant conditional probability and expectation. For a general theory of estimates of this type see Zhang (2005) and references therein. Some direct variance estimates appear in Rinott (2003).

2 The Model

For completeness we briefly introduce the Poisson and Negative Binomial models. More details can be found, for example, in Bethlehem et al. (1990), Cameron and Trivedi (1998), Rinott (2003).

We assume $F_k \sim$ Poisson$(N\gamma_k)$, independently, with $\sum \gamma_k = 1$. Binomial (or Poisson) sampling from F_k means that $f_k|F_k \sim Bin(F_k, \pi_k)$, π_k being the (known) sampling fraction in cell k. These are common assumptions in the frequency table literature, where it is convenient for log-linear modeling to assume that all π_k's are equal, an assumption not made here. However, we assume that the inclusion probabilities π_k are fixed within cells. In certain cases such an assumption may not hold, and more complex models may be required.

By standard calculations we then have

$$f_k \sim \text{ Poisson}(N\gamma_k\pi_k) \text{ and } F_k \,|\, f_k \;\sim\; f_k + \text{ Poisson}(N\gamma_k(1-\pi_k))\,, \tag{2}$$

leading to the Poisson model (see references below).

We now add the Bayesian assumption $\gamma_k \sim$ Gamma(α, β) independently. (Later we assume a common value for α and β in some neighborhoods of cells, rather than the whole table.)

Then

$$f_k \sim NB(\alpha, p_k = \frac{1}{1 + N\pi_k\beta}), \tag{3}$$

the *generalized Negative Binomial distribution*, defined for any $\alpha > 0$ by

$$X \sim NB(\alpha, p) \text{ if } P(X = x) = \frac{\Gamma(x+\alpha)}{\Gamma(x+1)\Gamma(\alpha)}(1-p)^x p^\alpha, \; x = 0, 1, 2, \ldots, \tag{4}$$

which for a natural α counts the number of *failures* until α successes occur in independent Bernoulli trials with probability of success p. For this distribution we have $\mu = EX = \alpha(1-p)/p$, $\text{Var}X = \alpha(1-p)/p^2 = \mu + \mu^2/\alpha$, and the probability generating function is $g(t) = Et^X = p^\alpha/[1-(1-p)t]^\alpha$, see Cameron and Trivedi (1998, p 375).

With the above parametrization $\mu_k \equiv Ef_k = N\pi_k\alpha\beta$, and for $b > 0$

$$E[1/(b+X)] = \int_0^1 t^{b-1}g(t)dt. \tag{5}$$

Further calculations yield

$$F_k \mid f_k \sim f_k + NB(\,\alpha + f_k,\ \rho_k = \frac{N\pi_k\beta + 1}{N\beta + 1}\,), \tag{6}$$

and clearly $F_k \geq f_k$.

This is the generalized Negative Binomial model used in this paper.

As $\alpha \to 0$ and $\beta \to \infty$ we obtain $F_k \mid f_k \sim f_k + NB(f_k, \pi_k)$, which is exactly the Negative Binomial assumption used in the *Argus* method. See Franconi and Polettini (2004) and references therein for details. If $\alpha \to \infty$ and $\alpha\beta \to$ constant, the Poisson model used in this context by Skinner and Holmes (1998) and Elamir and Skinner (2006) is obtained. Therefore the generalized Negative Binomial subsumes both models.

Using (5), (6) and setting $\rho_k = (N\pi_k\beta + 1)/(N\beta + 1)$, it is easy to compute *individual risk measures* for cell k, defined by

$$P(F_k = 1|f_k = 1) = \rho_k^{1+\alpha}, \qquad E[\frac{1}{F_k}|f_k = 1] = \frac{\rho_k(1-\rho_k^\alpha)}{\alpha(1-\rho_k)}. \tag{7}$$

3 Smoothing Polynomials and Local Neighborhoods

Our goal in this section is to estimate the parameters of the model so that we can estimate the quantities in (7). The global risk measures will then be estimated as indicated in (1).

The estimation question here is essentially the following: given, say, a sample unique, how likely is it to be also a population unique, or arise from a small population cell. If a sample unique is found in a part of the sample table where neighboring cells (by some reasonable metric, to be discussed later) are small or empty, then it seems reasonable to believe that it is more likely to have arisen from a small population cell. This motivates our attempt to study local neighborhoods, and compare the results to those obtained by using model-driven neighborhood arising in hierarchical log-linear models, where it seems that the neighborhoods may be too large, and the Argus method which uses no neighborhoods.

Consider frequency tables in which some of the attributes are ordinal, and define closeness between categories of an attribute in terms of the order, or more generally, suppose that for a certain attribute one can say that some values of the attribute are closer to a given value than others. For example, Age and number of Years of Education are ordinal attributes, and naturally the age of 16 is closer to 15 than to 20, say, while Occupation is not ordinal, but one can try to define reasonable notions of closeness between different occupations. The attribute values of variables which are purely categorical will be kept fixed within a neighborhood, and ordinal variables will vary within a range that defines the neighborhood.

Classical log-linear models do not take such closeness into account, and therefore, when such models are used for individual cell parameter estimation, the estimates involve data in cells which may be rather remote from the estimated cell. On the other hand, as mentioned above, the *Argus* method bases its estimation only on the sampling weights in the estimated population cell. There is no learning from other cells, the structure of the table plays no role, and each cell's parameter is estimated separately.

Our approach consists of using local neighborhood smoothing which will be described in (10) below, along with the generalized Negative Binomial model of (3)-(6). We thus assume that $f_k \sim NB(\alpha, p_k = \frac{1}{1+N\pi_k\beta})$, and therefore $\mu_k \equiv Ef_k = \alpha(1-p_k)/p_k = N\pi_k\alpha\beta$, see (3) and the subsequent relations.

We describe the proposed estimation method for μ and α. These estimates will be transformed to estimates of the parameters appearing in the individual risk measures (7), which in turn lead to estimates of the global risk measures using (1).

For each fixed cell k we define a neighborhood of cells $M = M_k$ (where $k \in M$) and estimate the values of μ_k and α_k using neighboring cells $k' \in M_k$ and the assumption

$$f_{k'} \sim NB(\alpha_k, p_{k'} = \frac{1}{1+N\pi_{k'}\beta_k}), \tag{8}$$

where α_k and β_k are fixed in the neighborhood and do not depend on k', while $p_{k'}$ actually depends also on k. Since we now fix k we suppress it as an index in α, β or $p_{k'}$, and write $Ef_{k'} = \mu_{k'} = \alpha(1-p_{k'})/p_{k'}$. For the fixed k, set $\boldsymbol{\mu} = \{\mu_{k'} : k' \in M\}$, so the index k is suppressed also in μ. We consider the likelihood of the observations $\{f_{k'} : k' \in M\}$ in a neighborhood $M = M_k$ of k based on (8), and using different parameterizations which include μ and $a = 1/\alpha$

$$\begin{aligned} L(a, \boldsymbol{\mu}) \equiv L(a, \boldsymbol{\mu}; \{f_{k'} : k' \in M\}) &= \prod_{k' \in M} \frac{\Gamma(x+\alpha)}{\Gamma(x+1)\Gamma(\alpha)}(1-p_{k'})^{f_{k'}} p_{k'}^{\alpha} \\ &= \prod_{k' \in M} \frac{\Gamma(x+\alpha)}{\Gamma(x+1)\Gamma(\alpha)}[1-\alpha/(\mu_{k'}+\alpha)]^{f_{k'}}[\alpha/(\mu_{k'}+\alpha)]^{\alpha}. \end{aligned} \tag{9}$$

We emphasize again that although in the above formulas only dependence on k' is shown, it should be noted that α, β and μ depend on k, and therefore $p_{k'}$ and $\mu_{k'}$ depend both on k and k'.

For each k we will estimate $\alpha = \alpha_k$ and $\mu_{k'}$ for $k' \in M = M_k$ using the likelihood (9) and a smoothing model described next, and then use the estimates of α_k and μ_k (not using the $\mu_{k'}$ estimates for $k' \neq k$) for further risk estimates, as discussed below.

Following Simonoff (1998), see also references therein, we use a local smoothing polynomial model.

For convenience of notation we now assume $m = 2$ (a two-way table); the extension to any m is straightforward. For each fixed $k = (k_1, k_2)$ separately, we write the log-linear model below for $\mu_{k'}$ in terms of the parameters $\boldsymbol{\theta}=(\theta_0, \theta_1, \vartheta_1, \ldots, \theta_t, \vartheta_t)$, with $k' = (k'_1, k'_2)$ varying in the neighborhood $M = M_k$ of k :

$$\log \mu_{k'}(\boldsymbol{\theta}) = \theta_0 + \theta_1(k'_1 - k_1) + \vartheta_1(k'_2 - k_2) + \ldots + \theta_t(k'_1 - k_1)^t + \vartheta_t(k'_2 - k_2)^t, \tag{10}$$

for some natural number t. One can hope that such a polynomial, with a suitable t, provides a reasonable approximation to $\log \mu_{k'}$ if $k' = (k'_1, k'_2)$ is in a small neighborhood of $k = (k_1, k_2)$. Substituting (10) into the likelihood function (9) using the relations between parameterizations as described above we obtain the likelihood function $L(a, \boldsymbol{\theta})$.

Our next goal is to maximize it as a function of $a = 1/\alpha$ and $\boldsymbol{\theta}$. This maximization takes place in principle for each cell k (although it may suffice for our purposes to carry it out for sample uniques only, that is, for cells such that $f_k = 1$). A source of difficulty here is that $\log L(a, \boldsymbol{\theta})$ is concave in $\boldsymbol{\theta}$, but not jointly in $(a, \boldsymbol{\theta})$, and therefore local maxima may occur, Hessians are not necessarily positive definite, and standard algorithms may not converge to the real MLE. This difficulty does not arise in the Poisson case of log-linear models of this type, where the log-likelihood is concave, see Rinott and Shlomo (2006), for a detailed discussion of the Poisson model. There are several options for maximization. SAS uses a *Newton-Raphson Ridge Optimization* (NRRIDG) which adds a multiple of the identity matrix to the Hessian when the latter is not positive definite, and also the *Fisher Scoring Algorithm* which replaces the Hessian by its expectation which is the information matrix, (using the parameter estimates of the current iteration), thus making it positive definite. We used our own program of the latter algorithm.

The components of the gradient of the log-likelihood function are obtained by differentiation and some manipulations as in Cameron and Trivedi (1998 p. 71), taking the form:

$$\frac{\partial \log L(a, \boldsymbol{\theta})}{\partial a} = \sum_{k' \in M} \left\{ \frac{1}{a^2} \left(\log(1 + a\mu_{k'}) - \sum_{v=0}^{f_{k'}-1} \frac{1}{v + a^{-1}} \right) + \frac{f_{k'} - \mu_{k'}}{a(1 + a\mu_{k'})} \right\}$$

$$\frac{\partial \log L(a, \boldsymbol{\theta})}{\partial \theta_\ell} = \sum_{k' \in M} \frac{f_{k'} - \mu_{k'}}{(1 + a\mu_{k'})}(k'_1 - k_1)^\ell, \;\; \ell = 0, \ldots, t$$

$$\frac{\partial \log L(a, \boldsymbol{\theta})}{\partial \vartheta_\ell} = \sum_{k' \in M} \frac{f_{k'} - \mu_{k'}}{(1 + a\mu_{k'})}(k'_2 - k_2)^\ell, \;\; \ell = 1, \ldots, t.$$

Note that in the solution to the related normal equations, the resulting vector $(a, \boldsymbol{\theta})$ depends on k.

The Hessian is calculated as follows:

$$\frac{\partial^2 \log L(a,\boldsymbol{\theta})}{\partial a^2} = \sum_{k' \in M} \left\{ \frac{-2}{a^3} \left(\log(1 + a\mu_{k'}) - \sum_{\upsilon=0}^{f_{k'}-1} \frac{1}{\upsilon + a^{-1}} \right) + \frac{1}{a^2} \left(\frac{\mu_{k'}}{1 + a\mu_{k'}} - \sum_{\upsilon=0}^{f_{k'}-1} \frac{1}{(a\upsilon + 1)^2} \right) - \frac{(f_{k'} - \mu_{k'})(1 + 2a\mu_{k'})}{a^2(1 + a\mu_{k'})^2} \right\},$$

$$\frac{\partial^2 \log L(a,\boldsymbol{\theta})}{\partial \theta_\ell \partial a} = -\sum_{k' \in M} \frac{(f_{k'} - \mu_{k'})\mu_{k'}}{(1 + a\mu_{k'})^2} (k'_1 - k_1)^\ell, \ \ \ell = 0, \dots, t,$$

and

$$\frac{\partial^2 \log L(a,\boldsymbol{\theta})}{\partial \theta_i \partial \theta_j} = -\sum_{k' \in M} \frac{(1 + af_{k'})\mu_{k'}}{(1 + a\mu_{k'})^2} (k'_1 - k_1)^{i+j} \ \ i,j = 0, \dots, t.$$

$$\frac{\partial^2 \log L(a,\boldsymbol{\theta})}{\partial \theta_i \partial \vartheta_j} = -\sum_{k' \in M} \frac{(1 + af_{k'})\mu_{k'}}{(1 + a\mu_{k'})^2} (k'_1 - k_1)^i (k'_2 - k_2)^j, \ \ i = 0, \dots, t, \ j = 1, \dots, t.$$

$$\frac{\partial^2 \log L(a,\boldsymbol{\theta})}{\partial \vartheta_i \partial \vartheta_j} = -\sum_{k' \in M} \frac{(1 + af_{k'})\mu_{k'}}{(1 + a\mu_{k'})^2} (k'_2 - k_2)^{i+j}, \ \ i,j = 1, \dots, t.$$

With $\arg\max L(a,\boldsymbol{\theta}) = (\hat{a}, \hat{\boldsymbol{\theta}})$, and $\hat{\theta}_0$ denoting the first component of $\hat{\boldsymbol{\theta}}$, we finally obtain our estimate of $\mu_k = \mu_{(k_1,k_2)}$ in the form

$$\hat{\mu}_k \equiv \mu_k(\hat{\boldsymbol{\theta}}) = \exp(\hat{\theta}_0), \tag{11}$$

where the second equality is explained by taking $k' = k = (k_1, k_2)$ in (10).

To summarize, we obtain the estimates $\hat{a}_k$, $\hat{\boldsymbol{\theta}}$ both depending on k by a separate maximization for each k as explained above, leading to the estimates $\hat{a}_k$, and $\hat{\mu}_k$ of (11). For the risk measure discussed in this paper, it suffices to compute these estimates for cells k which are sample uniques, that is, $f_k = 1$

Having estimated $\hat{a}_k, \hat{\mu}_k$ for each cell k separately on the basis of a neighborhood M_k, we use them to estimate the quantities ρ_k and $\alpha = \alpha_k$ which are obtained by tracing back the reparameterizations. Using the relations $\rho_k = \frac{N\pi_k\beta+1}{N\beta+1}$, and $\mu_k = N\pi_k\alpha_k\beta_k$ we readily obtain

$$\rho_k = \frac{\mu_k + \alpha_k}{\mu_k/\pi_k + \alpha_k}, \qquad \alpha_k = 1/a_k.$$

We plug our estimates $\hat{a}_k, \hat{\mu}_k$ in the latter formula, and then plug the resulting estimates of α_k and ρ_k into (7), to obtain the individual risk estimates. The global risk measures are estimated as indicated in (1).

4 Experiments with Neighborhoods

We present a few experiments. Our results are preliminary as already mentioned and more work is needed on the approach itself and on classifying types of data for which it might work.

For the computations we used our versions of the *Argus* and log-linear models methods, programmed on the SAS system. The weights w_i for the *Argus* method in all our examples were computed by post-stratification on Sex by Age by Geographical location (the latter is not one of the attributes in any of the tables, but it was used for post-stratification). These variables are commonly used for post-stratification, other strata may give different, and perhaps better results.

In the experiments below we compare results of our NB smoothing method with the Argus estimates and with Poisson hierarchical log-linear models (Elamir and Skinner 2006), with two log-linear models: one of independence, the other including all two-way interactions.

We defined neighborhoods M of k by varying around k coordinates corresponding to attributes that are ordinal, allowing in each coordinate a fixed maximal distance, which is equivalent to using a ball in the sup-norm, or intersection of sup-norm and ℓ_1 balls (see below). In principle we would use close values in non-ordinal attributes when possible (e.g., in Occupation). Attributes in which closeness of values cannot be defined, such as Sex remain constant in the whole neighborhood and therefore in our experiments neighborhoods always consist of individuals of the same Sex.

In all experiments we took a real population data file of size N given in the form of a contingency table with K cells, and from it we took a simple random sample of size n, so that always $\pi_k = n/N$. Our approach and formulas have the advantage of allowing for variable π_k's, but taking them all equal enables us to compare to the log-linear models method, where equal π_k's are required. Since the population and the sample are known to us, we can compute the *true values* of τ_1 and τ_2 and their estimates by the different methods, and compare.

Example 1. Population : an extract from the 1995 Israeli Census. $N = 37,586$, $n = 3,759$, $K = 11,648$. Attributes (with number of levels in parentheses): Sex(2) * Age Groups (32) * Income Groups(14) * Years of Study (13).

In this small experiment we tried our proposed smoothing polynomial model of (10) for $t = 2$. We considered one type of neighborhood here, constructed by fixing Sex and varying each of the other attribute value in k by at most c values up or down, that is, the neighborhood of each cell k (with a fixed Sex value) is of the type

$$M = \{k' : k'_1 = k_1,\ \max_{2 \le i \le m} |k'_i - k_i| \le c\}. \tag{12}$$

With $m = 4$ and one variables fixed we vary three variables, each over a range of five values for $c = 2$, , so we have $|M| = 5^3 = 125$, and taking $c = 3$ we have $|M| = 7^3 = 343$.

For cells near the boundaries some of the cells in their neighborhoods do not exist; here we set non-existing cells' frequencies to be zero, but other possibilities can be considered.

The table below presents the true τ values and their estimates by the methods described above.

Model	τ_1	τ_2
True Values	187	452.0
Argus	137.2	346.4
Log Linear Model: Independence	217.3	518.0
Log Linear Model: 2-Way Interactions	167.2	432.8
NB Smoothing $t = 2$ $\|M\| = 125$	181.9	461.3
NB Smoothing $t = 2$ $\|M\| = 343$	179.6	449.8

Example 2. Population : an extract from the 1995 Israeli Census. $N = 746,949$, $n = 14,939$, $K = 337,920$. Attributes: Sex (2) * Age Groups (16) * Years of Study (10) * Number of Years in Israel (11) * Income Groups (12) * Number of Persons in Household (8). Note that this is a very sparse table.

We applied the smoothing polynomial of (10) for $t = 2$ and neighborhoods obtained by varying all attributes except for Sex which was fixed. Neighborhoods are of the type

$$M = \{k' : k'_1 = k_1, \max_{2 \le i \le m} |k'_i - k_i| \le c, \sum_i |k'_i - k_i| \le d\}, \tag{13}$$

with $c = 2$; $d = 4$ and 6, and $|M| = 581$ and $1,893$, respectively. The results are given in the table below.

Model	τ_1	τ_2
True Values	430	1,125.8
Argus	114.5	456.0
Log Linear Model: Independence	773.8	1,774.1
Log Linear Model: 2-Way Interactions	470.0	1,178.1
NB Smoothing $t = 2$ $\|M\| = 581$	300.7	999.4
NB Smoothing $t = 2$ $\|M\| = 1,893$	461.9	1,179.6

Example 3. Population : an extract from the 1995 Israeli Census. $N = 746,949$, $n = 7,470$, $K = 42,240$. Attributes: Sex (2) * Age Groups (16) * Years of Study (10) * Number of Years in Israel (11) * Income Groups (12).

We applied the smoothing polynomial of (10) for $t = 2$ and neighborhoods obtained by varying all attributes except for Sex which was fixed. Neighborhoods are as in (13) with $c = 2$ and $d = 4$ and $|M| = 257$; $c = 2$, $d = 6$, and $|M| = 545$, and $c = 2$, $d = 8$ and $|M| = 625$. Smaller neighborhoods did not yield good estimates. The results are given in the table below.

Model	τ_1	τ_2
True Values	42	171.2
Argus	20.7	95.4
Log Linear Model: Independence	28.8	191.5
Log Linear Model: 2-Way Interactions	35.8	164.1
NB Smoothing $t = 2$ $\|M\| = 257$	24.7	147.5
NB Smoothing $t = 2$ $\|M\| = 545$	39.3	174.8
NB Smoothing $t = 2$ $\|M\| = 625$	45.8	184.4

Discussion of examples. The log-linear model method was tested in Skinner and Shlomo (2005, 2006) and references therein, and based on model selection techniques and goodness of fit criteria, yields good estimates for disclosure risk measures for the types of experiments done here. Di Consiglio et al. (2003) presented experiments for individual risk assessment with Argus, which seems to perform less well than the log-linear method in many of our experiments with global risk measures. Our new method still requires fine-tuning. At present the results seem comparable or somewhat better than the Poisson hierarchical log-linear method. In Rinott and Shlomo (2006) we performed experiments of this kind on a smoothing method based on the Poisson rather than the Negative Binomial distribution. So far the present Negative Binomial model improves all the results, and seems potentially promising.

Naturally, more variables and sparse data sets with a large number of cells are typical and need to be tested. Such files will cause difficulties to any method, and this is where the different methods should be compared. In sparse multi-way tables, model selection will be crucial but difficult for the log-linear method, and perhaps simpler for the smoothing approach.

Our proposed method is at a preliminary stage and requires more work. Particular directions are the following:

1. Adjust the parameter estimates to fit known population marginals obtained from prior knowledge and sampling weights, and vary the sampling fractions π_k. In all our experiments so far we used constant π_k's but unlike methods based on log-linear models, the formulas given here allow for variables π_k's, and we intend to try variable π_k's obtained by sampling design or post-stratification.
2. Use goodness of fit measures and information on population marginals and sampling weights to select the type and size of the neighborhoods, and the degree of the smoothing polynomial in (10).

The examples show a typical monotonicity phenomenon discussed also in the papers of Rinott and Shlomo (2005, 2006): the risk measure estimates decrease as a function of the size of the log-linear model (that is, with one exception of τ_1 in Example 3, the two-way models always yield lower estimates than the independence model). In the present smoothing approach the risk estimates always decrease with the size of the neighborhood. These two facts can be explained in the same way: the better the fit to the sample data, the smaller the risk estimates. A larger log-linear model or a smaller smoothing neighborhood correspond to a better fit and therefore yield smaller risk estimates. In the presence of such monotonicity, a study of suitable goodness of fit measures to choose the right model is critical.

3. We intend to test this method also for individual risk measure estimates, which are important in themselves, and may also shed more light on efficient neighborhood and model selection. Our preliminary experiments suggest that the smoothing approach performs relatively well in estimating individual risk.

References

1. Benedetti, R., Franconi, L. and Piersimoni, F.: Per-record risk of disclosure in dependent data. *Proceedings of the Conference on Statistical Data Protection, Lisbon March 1998.* (1999) European Communities, Luxembourg.
2. Bethlehem, J., Keller, W., and Pannekoek, J.: Disclosure Control of Microdata. *J. Amer. Statist. soc.* **8**5 (1990) 38–45.
3. Cameron, A. C., and Trivedi, P. K.: *Regression analysis of count data*, Econometric Society Monographs **3**0 (1998) Cambridge University Press.
4. Di Consiglio, L., Franconi, L., and Seri, G.: Assessing individual risk of disclosure: an experiment. *Proceedings of the Joint ECE/Eurostat Work Session on Statistical Data Confidentiality* (2003) Luxemburg 286–298.
5. Elamir, E. and Skinner, C.: Record-level measures of disclosure risk for survey microdata, *J. Official Statist* **2**2 (2006) to appear.
6. Franconi, L. and Polettini, S.: Individual risk estimation in mu-argus: a review. In J. Domingo-Ferrer and V. Torra, editors, *Privacy in Statistical Databases*, volume **3**050 of LNCS, Springer Berlin Heidelberg (2004) 262-272.
7. Polettini, S. and Seri, G.: Guidelines for the protection of social micro-data using individual risk methodology - Application within mu-argus version 3.2. *CASC Project Deliverable No. 1.2-D3* (2003). http://neon.vb.cbs.nl/casc/
8. Rinott, Y.: On models for statistical disclosure risk estimation. *Proceedings of the Joint ECE/Eurostat Work Session on Statistical Data Confidentiality*, Luxemburg (2003) 275-285.
9. Rinott, Y. and Shlomo, N.: A neighborhood regression model for sample disclosure risk estimation. *Proceedings of the Joint UNECE/Eurostat work session on statistical data confidentiality* Geneva, Switzerland (2005) 79-87.
10. Rinott, Y. and Shlomo, N.: A smoothing model for sample disclosure risk estimation. (2006) Submitted.
11. Simonoff S. J.: Three sides of smoothing: categorical Data smoothing, nonparametric regression, and density estimation, *International Statistical Review*, **66** (1998) 137-156.
12. Skinner, C. and Holmes, D.: Estimating the Re-identification Risk Per Record in Microdata, *J. Official Statist.* **14** (1998) 361-372.

13. Skinner, C. and Shlomo, N.: Assessing disclosure risk in microdata using record-level measures. In *Proceedings of the Joint UNECE/Eurostat work session on statistical data confidentiality* Geneva, Switzerland (2005) 69-78.
14. Skinner, C. and Shlomo, N.: Assessing identification risk in survey microdata using log-linear models. (2006) Submitted.
15. Willenborg, L. and de Waal T.: *Elements of Statistical Disclosure Control*, Lecture Notes in Statistics, **155** (2001) Springer, New York.
16. Zhang C.-H.: Estimation of sums of random variables: examples and information bounds. *Ann. Statist.* **33** (2005) 2022-2041.

Entry Uniqueness in Margined Tables

Shmuel Onn[*]

Technion - Israel Institute of Technology, 32000 Haifa, Israel

Abstract. We consider a problem in secure disclosure of multiway table margins. If the value of an entry in all tables having the same margins as those released from a source table in a data base is unique, then the value of that entry can be exposed and disclosure is insecure. We settle the computational complexity of detecting whether this situation occurs. In particular, for multiway tables where one category is significantly richer than the others, that is, when each sample point can take many values in one category and only few values in the other categories, we provide, for the first time, a polynomial time algorithm for checking uniqueness, allowing disclosing agencies to check entry uniqueness and make learned decisions on secure disclosure. Our proofs use our recent results on universality of 3-way tables and on n-fold integer programming, which we survey on the way.

1 Introduction

It is a common practice in the disclosure of a multiway table containing sensitive data to release some of the table margins rather than the table itself (see e.g. [2,3,9] and references therein). Once the margins are released, the security of any specific entry of the table is related to the structure of the set of possible values that can occur in that entry in any table having the same margins as those of the source table in the data base. In particular, if this set consists of a unique value, that of the source table, then this entry can be exposed and security can be violated.

This raises the following algorithmic problem.

Entry Uniqueness in Margined Tables. Given consistent disclosed margin values, and a specific entry index, is the value that can occur in that entry in any table having these margins, unique ?

An efficient algorithm for this problem allows the disclosing agency to check possible collections of margins before disclosure: if an entry value is not unique then disclosure may be assumed secure, whereas if the value is unique then disclosure may be risky and fewer margins should be released.

A less desirable situation occurs when this problem is computationally intractable: then the disclosing agency may not be able to check for uniqueness; however, in this situation, some consolation is in that an adversary will be computationally unable to identify and retrieve a unique entry either.

[*] Supported in part by a grant from ISF - the Israel Science Foundation and by an S. Langberg Research Fund.

J. Domingo-Ferrer and L. Franconi (Eds.): PSD 2006, LNCS 4302, pp. 94–101, 2006.

In this article we show the following contrasting statements, settling the complexity of the problem. The first theorem concerns the intractability of "short" 3-way tables, that is, tables of size $r \times c \times 3$.

Theorem 1. *It is coNP-complete to decide, given r, c and consistent 2-margins (line-sums) for 3-way tables of size $r \times c \times 3$, if the value of the entry $x_{1,1,1}$ is the same in all tables with these margins.*

This statement concerns tables with two large sides - r rows and c columns - which are considered varying part of the input. This result is sharpest possible: 2-way tables are easy; 3-way tables of size $r \times c \times 2$ with fixed line-sums are equivalent to 2-way tables and hence are easy as well; and, as follows from our second theorem below, d-way tables with only one side varying can be efficiently treated. This strengthens an earlier result of [10] on tables of format $r \times c \times n$ with all sides variable, see also [4]. Related work on so-called Markov bases for random walks on tables can be found in [1,6].

The second theorem concerns the efficient treatment of "long" tables of any dimension d, that is, d-way tables of size $m_1 \times \cdots \times m_{d-1} \times n$ with one long, variable, side n, and all other sides fixed.

Theorem 2. *For every fixed $d, m_1, \ldots, m_{d-1}$, there is a polynomial time algorithm that, given any n, any hierarchical collection of margins for $m_1 \times \cdots \times m_{d-1} \times n$ tables, and any entry index $(i_1, \ldots, i_d)$, determines whether or not the value of the entry $x_{i_1,\ldots,i_d}$ is the same in all tables with these margins.*

The precise definition of a *hierarchical* collection of margins is given below; in particular, for any k, the collection of all k-margins is hierarchical. Note again that this result is best possible: if two sides of a d-way table with $d \geq 3$ are large and varying then, by Theorem 1, the uniqueness problem is coNP-complete. Theorem 2 is especially reassuring in situations where one of the d categories generating tables is significantly richer than the others, that is, when each sample point can take many values in one category and only few values in the other categories. Then the algorithm underlying Theorem 2 allows disclosing agencies to check entry uniqueness and make learned decisions on secure disclosure.

In the next two sections we establish Theorems 1 and 2 respectively. Theorem 1 is proved using our recent results in [5,7] on the universality of short 3-way tables. Theorem 2 is proved using our recent results in [8] on the polynomial time solvability of the broad class of n-fold integer programming problems in variable dimension. Indeed, on the way, as a secondary goal of this article, we briefly survey our work on universality and n-fold integer programming, which provide powerful tools for treating multiway tables (see e.g. [6] for applications to Markov bases and walks on tables). The final section contains some concluding discussion of approximate versus accurate bounds computation.

Before proceeding to the proofs, we give some definitions on multiway polytopes, margins and hierarchical margin collections. A d-way polytope is the set of all $m_1 \times \cdots \times m_d$ nonnegative arrays $x = (x_{i_1,\ldots,i_d})$ such that the sums of the entries over some of their lower dimensional sub-arrays (margins) are specified. More precisely, for any tuple $(i_1,\ldots,i_d)$ with $i_j \in \{1,\ldots,m_j\} \cup \{+\}$, the corresponding *margin* $x_{i_1,\ldots,i_d}$ is the sum of entries of x over all coordinates j with $i_j = +$. The *support* of $(i_1,\ldots,i_d)$ and of $x_{i_1,\ldots,i_d}$ is the set $\mathrm{supp}(i_1,\ldots,i_d) := \{j : i_j \neq +\}$ of non-summed coordinates. For instance, if x is a $4 \times 5 \times 3 \times 2$ array then it has 12 margins with support $F = \{1,3\}$ such as $x_{3,+,2,+} = \sum_{i_2=1}^{5} \sum_{i_4=1}^{2} x_{3,i_2,2,i_4}$. A collection of margins is *hierarchical* if, for some family $\mathcal{F}$ of subsets of $\{1,\ldots,d\}$, it consists of all margins $u_{i_1,\ldots,i_d}$ with support in $\mathcal{F}$. In particular, for any $0 \leq k \leq d$, the collection of all k-margins of d-tables is hierarchical with $\mathcal{F}$ the family of all k-subsets of $\{1,\ldots,d\}$. Given a hierarchical collection of margins $u_{i_1,\ldots,i_d}$ supported on a family $\mathcal{F}$ of subsets of $\{1,\ldots,d\}$, the corresponding *d-way polytope* is the set of nonnegative arrays with these margins,

$$T_{\mathcal{F}} \;=\; \left\{ x \in \mathbf{R}_{+}^{m_1 \times \cdots \times m_d} \;:\; x_{i_1,\ldots,i_d} = u_{i_1,\ldots,i_d}\,,\;\; \mathrm{supp}(i_1,\ldots,i_d) \in \mathcal{F} \right\} \,.$$

The integer points in this polytope are precisely the d-way tables with the specified (disclosed) margins.

2 Intractability for Short 3-Way Tables: The Universality Theorem

Consider 3-way polytopes of $r \times c \times 3$ arrays with all line-sums fixed, that is, the hierarchical collection of all 2-margins, supported on the family $\mathcal{F} = \{\{1,2\},\{1,3\},\{2,3\}\}$, and their integer points, which are precisely the corresponding tables with all line-sums fixed (disclosed). The following striking universality of such 3-way polytopes and tables was very recently shown in [5,7].

Theorem 3. *Any rational polytope $P = \{y \in \mathbf{R}_{+}^{m} : Ay = b\}$ is polynomial time representable as a 3-way line-sum polytope of size $r \times c \times 3$ for some (polynomially bounded) r and c,*

$$T \;=\; \{\, x \in \mathbf{R}_{+}^{r \times c \times 3} \;:\; \sum_i x_{i,j,k} = w_{j,k}\,,\; \sum_j x_{i,j,k} = v_{i,k}\,,\; \sum_k x_{i,j,k} = u_{i,j} \,\}\,.$$

Here representable means that there is a coordinate-erasing projection from $\mathbf{R}^{r \times c \times 3}$ onto $\mathbf{R}^{m}$ providing a bijection between T and P and between the sets of integer points $T \cap \mathbf{Z}^{r \times c \times 3}$ and $P \cap \mathbf{Z}^{m}$. Thus, *any* rational polytope is an $r \times c \times 3$ line-sum polytope, and *any* integer (respectively, linear) programming problem is equivalent to an integer (respectively, linear) $r \times c \times 3$ line-sum transportation problem.

This result solved several open problems from [11,12], and had several implications on the complexity of Markov bases for hierarchical margins and related issues; in particular, it implied the following surprising statement, see [5,6].

Proposition 4. *For any finite set S of nonnegative integers, there are r, c and 2-margins such that the set of values occurring in the entry $x_{1,1,1}$ in all $r \times c \times 3$ tables with these margins is precisely S.*

The margins that realize any desired set of values S can be automatically computed using the construction underlying the universality Theorem 3. Applying this to the set $S = \{0, 2\}$, we obtain the following example where the set of values occurring in $x_{1,1,1}$ in all tables is $S = \{0, 2\}$ and has a gap.

Example 1. **Gap in 2-margined 3-tables:** There are precisely two 3-way tables of size $6 \times 4 \times 3$ with the 2-margins below; in one table $x_{1,1,1} = 0$ while in the other table $x_{1,1,1} = 2$.

$$\begin{pmatrix} 2&1&2&0&2&0 \\ 1&0&2&0&0&2 \\ 1&0&0&2&2&0 \\ 0&1&0&2&0&2 \end{pmatrix}, \quad \begin{pmatrix} 2&1&2&3&0&0 \\ 2&1&0&0&2&1 \\ 0&0&2&1&2&3 \end{pmatrix}, \quad \begin{pmatrix} 2&3&2 \\ 2&1&2 \\ 2&1&2 \\ 2&1&2 \end{pmatrix} .$$

We proceed to prove Theorem 1 using Theorem 3.

Theorem 1. *It is coNP-complete to decide, given r, c and consistent 2-margins (line-sums) for 3-way tables of size $r \times c \times 3$, if the value of the entry $x_{1,1,1}$ is the same in all tables with these margins.*

Proof. The *subset-sum* problem, well known to be NP-complete, is the following: given positive integers $a_0, a_1, \ldots, a_m$, decide if there is an $I \subseteq \{1, \ldots, m\}$ with $a_0 = \sum_{i \in I} a_i$. We reduce its complement to ours. Given $a_0, a_1, \ldots, a_m$, consider the polytope in $2(m+1)$ variables $y_0, y_1 \ldots, y_m, z_0, z_1, \ldots, z_m$,

$$P \ := \ \{(y, z) \in \mathbf{R}_+^{2(m+1)} \ : \ a_0 y_0 - \sum_{i=1}^{m} a_i y_i = 0\,,\ y_i + z_i = 1\,,\ i = 0, 1 \ldots, m\}\ .$$

First, note that it always has one integer point with $y_0 = 0$, given by $y_i = 0$ and $z_i = 1$ for all i. Second, note that it has an integer point with $y_0 \neq 0$ if and only if there is an $I \subseteq \{1, \ldots, m\}$ with $a_0 = \sum_{i \in I} a_i$, given by $y_0 = 1$, $y_i = 1$ for $i \in I$, $y_i = 0$ for $i \in \{1, \ldots, m\} \setminus I$, and $z_i = 1 - y_i$ for all i. Lifting P to a suitable $r \times c \times 3$ line-sum polytope T with the coordinate y_0 embedded in the entry $x_{1,1,1}$ using Theorem 3, we find that T has a table with $x_{1,1,1} = 0$, and this value is unique among the tables in T if and only if there is *no* solution to the subset sum problem with $a_0, a_1, \ldots, a_m$. □

The next example demonstrates the construction of the proof of Theorem 1.

Example 2. **Encoding subset-sums in 2-margined 3-tables:** Given instance $m = 2$, $a_0 = 2$, $a_1 = a_2 = 1$ for subset-sum, the construction of Theorem 1 (incorporating the universality theorem) yields the line-sums below for tables of size $10 \times 8 \times 3$, where variables y_0, y_1, y_2 are embedded in entries $x_{1,1,1}, x_{3,7,1}, x_{4,8,1}$ respectively. The margins admit one table with $x_{1,1,1} = x_{3,7,1} = x_{4,8,1} = 0$ and one table with $x_{1,1,1} = x_{3,7,1} = x_{4,8,1} = 1$ corresponding to the subset-sum $a_0 = a_1 + a_2$ of $I = \{1, 2\}$.

$$\begin{pmatrix} 1&0&0&0&1&0&1&0&0&0 \\ 0&1&0&0&1&0&0&1&0&0 \\ 0&0&1&0&1&0&0&0&1&0 \\ 0&0&0&1&1&0&0&0&0&1 \\ 0&1&0&0&0&1&1&0&0&0 \\ 1&0&0&0&0&1&0&1&0&0 \\ 0&0&1&0&0&1&0&0&1&0 \\ 0&0&0&1&0&1&0&0&0&1 \end{pmatrix}, \quad \begin{pmatrix} 1&1&1&1&2&2&0&0&0&0 \\ 1&1&1&1&0&0&1&1&1&1 \\ 0&0&0&0&2&2&1&1&1&1 \end{pmatrix}, \quad \begin{pmatrix} 1&1&1 \\ 1&1&1 \\ 1&1&1 \\ 1&1&1 \\ 1&1&1 \\ 1&1&1 \\ 1&1&1 \\ 1&1&1 \end{pmatrix} .$$

3 Solvability for Long d-Way Tables: n-Fold Integer Programming

It is well known that integer programming problems are generally intractable. However, very recently, in [8], we were able to show that an important broad class of integer programming problems in variable dimension is polynomial time solvable. This result may seem a bit technical at a first glance, but is really very natural and has many applications in operations research and statistics, including clustering, partition problems and more. To state it, we need the following definition: given an $(r + s) \times t$ matrix A, let A_1 be its $r \times t$ sub-matrix consisting of the first r rows and let A_2 be its $s \times t$ sub-matrix consisting of the last s rows. Define the *n-fold matrix* of A to be the following $(r + ns) \times nt$ matrix,

$$A^{(n)} \quad := \quad (\mathbf{1}_n \otimes A_1) \oplus (I_n \otimes A_2) \quad = \quad \begin{pmatrix} A_1 & A_1 & A_1 & \cdots & A_1 \\ A_2 & 0 & 0 & \cdots & 0 \\ 0 & A_2 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & A_2 \end{pmatrix} .$$

Note that $A^{(n)}$ depends on r and s: these will be indicated by referring to A as an "$(r + s) \times t$ matrix".

In [8] we show the following theorem on integer programming over n-fold matrices.

Theorem 5. *For any fixed $(r + s) \times t$ integer matrix A there is a polynomial time algorithm that, given any n and any vectors $b \in \mathbf{Z}^{r+ns}$ and $c \in \mathbf{Z}^{nt}$, solves the n-fold integer programming problem*

$$\max\{cx : \ A^{(n)}x = b, \ x \in \mathbf{N}^{nt}\} .$$

As usual, *solving* the integer program means that the algorithm either returns an optimal solution $x \in \mathbf{N}^{nt}$, or asserts that the problem is infeasible, or asserts that the objective function is unbounded.

The equations defined by the n-fold matrix have the following, perhaps more illuminating, interpretation: splitting the variable vector and the right-hand-side vector into components of suitable sizes, $x = (x^1, \ldots, x^n)$ and $b = (b^0, b^1, \ldots, b^n)$, where $b^0 \in \mathbf{Z}^r$ and $x^k \in \mathbf{N}^t$ and $b^k \in \mathbf{Z}^s$ for $k = 1, \ldots, n$, the equations become $A_1(\sum_{k=1}^n x^k) = b^0$ and $A_2 x^k = b^k$ for $k = 1, \ldots, n$. Thus, each component x^k satisfies a system of constraints defined by A_2 with its own right-hand-side b^k, and the sum $\sum_{k=1}^n x^k$ obeys constraints determined by A_1 and b^0 restricting the "common resources shared by all components".

We proceed to prove Theorem 2 using Theorem 5. In fact, we show that any linear objective function can be maximized (and minimized) over long d-way tables with fixed margins. In particular, the *exact* smallest and largest values of any entry can be computed (as opposed to *approximative* lower and upper bounds obtainable from the linear programming relaxation of the multiway polytope).

Theorem 2. *For every fixed $d, m_1, \ldots, m_{d-1}$, there is a polynomial time algorithm that, given any n, any hierarchical collection of margins for $m_1 \times \cdots \times m_{d-1} \times n$ tables, and any entry index $(i_1, \ldots, i_d)$, determines whether or not the value of the entry $x_{i_1,\ldots,i_d}$ is the same in all tables with these margins.*

Proof. Let $\mathcal{F}$ be the family of subsets of $\{1, \ldots, d\}$ on which the given hierarchical collection of margins is supported. Re-index arrays $x = (x_{i_1,\ldots,i_d})$ as $x = (x^1, \ldots, x^n)$ where, for $k = 1, \ldots, n$,

$$x^k \quad := \quad (x_{i_1,\ldots,i_{d-1},k}) \quad := \quad (x_{1,\ldots,1,k}, \ldots, x_{m_1,\ldots,m_{d-1},k})$$

is a suitably indexed vector of length $t := \prod_{i=1}^{d-1} m_i$ representing the k-th layer of x. Then the margin equations $x_{i_1,\ldots,i_d} = u_{i_1,\ldots,i_d}$ for all tuples satisfying $supp(i_1, \ldots, i_d) \in \mathcal{F}$ can be written as $A^{(n)}x = b$ with $A^{(n)}$ the n-fold matrix of a suitable $(r+s) \times t$ matrix A, with r, s, A_1 and A_2 suitably determined from $\mathcal{F}$ and with the right-hand-side $b \in \mathbf{N}^{r+ns}$ determined from the given margins, in such a way that the equations $A_1(\sum_{k=1}^n x^k) = b^0$ represent the equations of all margins $x_{i_1,\ldots,i_d}$ with $i_d = +$ (where summation over layers occurs), whereas the equations $A_2 x^k = b^k$ for $k = 1, \ldots, n$ represent the equations of all margins $x_{i_1,\ldots,i_d}$ with $i_d \neq +$ (where summation is within a single layer k at a time).

Thus, by Theorem 5, for any integer vector $c \in \mathbf{Z}^{nt}$, we can solve in polynomial time the following integer programming problem over the multiway polytope,

$$\max\{cx : x \in \mathbf{N}^{m_1 \times \cdots \times m_{d-1}, n},\ x_{i_1,\ldots,i_d} = u_{i_1,\ldots,i_d},\ \mathrm{supp}(i_1, \ldots, i_d) \in \mathcal{F}\} = \max\{cx : A^{(n)}x = b,\ x \in \mathbf{N}^{nt}\}.$$

In particular, we can compute in polynomial time the smallest value l and largest value u of the entry $x_{i_1,\ldots,i_d}$ in all tables with these margins, by solving the following two n-fold integer programs,

$$l \ := \ \min\{x_{i_1,\ldots,i_d} \ : \ A^{(n)}x = b,\ x \in \mathbf{N}^{nt}\}\ ,$$

$$u \ := \ \max\{x_{i_1,\ldots,i_d} \ : \ A^{(n)}x = b,\ x \in \mathbf{N}^{nt}\}\ .$$

Clearly, entry $x_{i_1,\ldots,i_d}$ attains a unique value in all tables with the given (disclosed) hierarchical collection of margins if and only if $l = u$, completing the description of the algorithm and the proof. □

Example 3. Consider long 3-way tables of size $3 \times 3 \times n$ with all line-sums fixed, that is, with $d = 3$, $m_1 = m_2 = 3$, and the hierarchical collection of all 2-margins, supported on $\mathcal{F} = \{\{1,2\},\{1,3\},\{2,3\}\}$. Then $r = t = 9$, $s = 6$, and writing $x^k = (x_{1,1,k}, x_{1,2,k}, x_{1,3,k}, x_{2,1,k}, x_{2,2,k}, x_{2,3,k}, x_{3,1,k}, x_{3,2,k}, x_{3,3,k})$ for $k = 1, \ldots, n$, the $(9 + 6) \times 9$ matrix A whose n-fold product $A^{(n)}$ defines the $3 \times 3 \times n$ multiway polytope as in the proof of Theorem 2 consists of $A_1 = I_9$ the 9×9 identity matrix and

$$A_2 \quad = \quad \begin{pmatrix} 1\,1\,1\,0\,0\,0\,0\,0\,0 \\ 0\,0\,0\,1\,1\,1\,0\,0\,0 \\ 0\,0\,0\,0\,0\,0\,1\,1\,1 \\ 1\,0\,0\,1\,0\,0\,1\,0\,0 \\ 0\,1\,0\,0\,1\,0\,0\,1\,0 \\ 0\,0\,1\,0\,0\,1\,0\,0\,1 \end{pmatrix} \quad .$$

Already for this case, of $3 \times 3 \times n$ tables, the only polynomial time algorithm we are aware of for the corresponding entry uniqueness and integer programming problems is the one of Theorem 2 above.

4 Discussion

Since integer programming problems are generally intractable, a common practice by disclosing agencies is to compute a lower bound $\hat{l}$ and an upper bound $\hat{u}$ on the entry $x_{i_1,\ldots,i_d}$ in all tables with these margins, by solving the *linear programming relaxations* of the corresponding multiway programs,

$$\begin{aligned} \hat{l} \;:=\; \min\{x_{i_1,\ldots,i_d} : x \in \mathbf{R}_+^{m_1\times\cdots\times m_{d-1},n} \;, \\ x_{i_1,\ldots,i_d} = u_{i_1,\ldots,i_d},\ \mathrm{supp}(i_1,\ldots,i_d) \in \mathcal{F}\} \;, \end{aligned}$$

$$\begin{aligned} \hat{u} \;:=\; \max\{x_{i_1,\ldots,i_d} : x \in \mathbf{R}_+^{m_1\times\cdots\times m_{d-1},n} \;, \\ x_{i_1,\ldots,i_d} = u_{i_1,\ldots,i_d},\ \mathrm{supp}(i_1,\ldots,i_d) \in \mathcal{F}\} \;, \end{aligned}$$

that is, where the variables are nonnegative real numbers without integrality constraints. While this can be done efficiently for tables of any size, it is only an approximation on the true smallest value l and largest value u of that entry in (integer) tables, and can be far from the truth; it is easy to design examples (using again the universality Theorem 3) of line-sums for $r \times c \times 3$ tables where there is a unique integer entry $x_{1,1,1}$, while the linear programming bounds are arbitrarily far apart, that is,

$$\hat{l} \quad << \quad l \quad = \quad x_{1,1,1} \quad = \quad u \quad << \quad \hat{u} \quad ,$$

which may lead to erroneously declaring insecure margin disclosure as secure. Indeed, let u be any large positive integer. Consider the triangle $P_u := \{y \in \mathbf{R}_+^2 : 2y_1 + (2u+1)y_2 = 4u + 1\}$. It has just one integer point $y = (u, 1)$, with $y_1 = u$, while $\hat{l} := \min\{y_1 : y \in P_u\} = 0$ and $\hat{u} := \max\{y_1 : y \in P_u\} = 2u + \frac{1}{2}$. Lifting P_u to a suitable $r \times c \times 3$ line-sum polytope T_u with the coordinate y_1 embedded in the entry $x_{1,1,1}$ using Theorem 3, we find that T_u has just one integer table, where the entry $x_{1,1,1}$ attains the unique value $l = x_{1,1,1} = u$, while the linear programming bounds are $\hat{l} = 0 << u << 2u + \frac{1}{2} = \hat{u}$.

Our Theorem 2 provides, for the first time, a polynomial time algorithm allowing to compute the true smallest value l and largest value u (and moreover optimizing any linear functional) over long d-way tables, enabling exact solution of the entry uniqueness problem and taking accurate decisions.

References

1. Aoki, S., Takemura, A.: Minimal basis for connected Markov chain over $3 \times 3 \times K$ contingency tables with fixed two-dimensional marginals. Austr. New Zeal. J. Stat. **45** (2003) 229–249
2. Cox L.H.: Bounds on entries in 3-dimensional contingency tables. Inference Control in Statistical Databases - From Theory to Practice. Lec. Not. Comp. Sci., Springer, New York, **2316** (2002) 21–33
3. Cox L.H.: On properties of multi-dimensional statistical tables. J. Stat. Plan. Infer. **117** (2003) 251–273
4. De Loera, J., Onn, S.: The complexity of three-way statistical tables. SIAM J. Comp. **33** (2004) 819–836
5. De Loera, J., Onn, S.: All rational polytopes are transportation polytopes and all polytopal integer sets are contingency tables. Proc. 10th Ann. Math. Prog. Soc. Symp. Integ. Prog. Combin. Optim., (Columbia University, New York), Lec. Not. Comp. Sci., Springer, New York, **3064** (2004) 338–351
6. De Loera, J., Onn, S.: Markov bases of three-way tables are arbitrarily complicated. J. Symb. Comp. **41** (2006) 173–181
7. De Loera, J., Onn, S.: All linear and integer programs are slim 3-way transportation programs. SIAM J. Optim. (to appear)
8. De Loera, J., Hemmecke, R., Onn, S., Weismantel, R.: N-fold integer programming (submitted)
9. Duncan, G.T, Fienberg, S.E., Krishnan, R., Padman, R., Roehrig, S.F.: Disclosure limitation methods and information loss for tabular data. In: Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies, North-Holland (2001)
10. Irving, R., Jerrum, M.R.: Three-dimensional statistical data security problems. SIAM J. Comp. **23** (1994) 170–184
11. Vlach, M.: Conditions for the existence of solutions of the three-dimensional planar transportation problem. Disc. App. Math. **13** (1986) 61–78
12. Yemelichev, V.A., Kovalev, M.M., Kravtsov, M.K.: Polytopes, Graphs and Optimisation. Cambridge University Press, Cambridge (1984)

Combinations of SDC Methods for Microdata Protection

Anna Oganian and Alan F. Karr

National Institute of Statistical Sciences, NC, USA
{aoganyan, karr}@niss.org

Abstract. A number of methods have been proposed in the literature for masking (protecting) microdata. Nearly all of these methods may be implemented with different degrees of intensity, by setting the value of an appropriate parameter. However, even parameter variation may not be sufficient to realize appropriate levels of disclosure risk and data utility. In this paper we propose a new approach to protection of numerical microdata: applying multiple stages of masking to the data in a way that increases utility but controls disclosure risk.

Keywords: Statistical disclosure control (SDC), microdata, disclosure control methods, data utility, disclosure risk, combinations of SDC methods.

1 Introduction

Often, statistical agencies disseminate information only in form of tables. But, microdata — records which contain information about individuals or establishments — offers far greater flexibility for statistical research, especially of an exploratory nature, than tables. As a result, there has been an increasing demand from users for such data, and agencies would like to be able to this demand, provided that confidentiality is not compromised. In particular, there is well-recognized need to prevent both identity and attribute disclosure.

Before releasing microdata, agencies delete from the data direct identifiers, such as names and addresses. However, risk of identification still exists, for example, by means of linkage of the released data to external databases. So in addition, released microdata are typically perturbed, in order to make disclosure more difficult. Methods of statistical disclosure control (SDC),[1] for doing so, are described in §2.

All these methods may be characterized by a quantified *Data Utility* ($\mathbf{DU}$) representing the utility of the data for legitimate users and *Disclosure Risk* ($\mathbf{DR}$)—the risk to respondent confidentiality after releasing such data. Almost always, only a single method, chosen in advance, is applied. This method is sometimes chosen by solving an optimization problem: finding the method with maximal utility for the particular data set subject to the disclosure risk being below some threshold. A more flexible approach consists in defining *risk-utility frontiers* using the partial order $\preceq_{\mathrm{RU}}$ defined by $R_1 \preceq_{\mathrm{RU}} R_2$ if and only if $\mathbf{DR}(R_2) \leq \mathbf{DR}(R_1)$ and $\mathbf{DU}(R_2) \geq \mathbf{DU}(R_1)$. When $R_1 \preceq_{\mathrm{RU}} R_2$, the R_2 is preferred to R_1 because it has both lower disclosure risk and

[1] Or statistical disclosure limitation (SDL).

J. Domingo-Ferrer and L. Franconi (Eds.): PSD 2006, LNCS 4302, pp. 102–113, 2006.

higher utility. Only candidate releases on the risk-utility frontier of maximal elements of all possible releases with respect to the partial order need be considered further: for any other candidate, some element of the frontier has lower risk *and* higher utility. Calculation of the frontier can be done using existing algorithms for finding the maxima in a set of vectors [KLP75].

In [KKORS06] it was shown that the methods described in §2 differ rather dramatically in terms of disclosure risk and data utility. Some are good for one but not the other, some seem not especially good for either risk or utility, and none is uniformly good for both. So, there is no "best method," because for each particular situation the maximal tolerable risk and minimal utility may be different. Instead, there is a set of undominated methods that are on the risk-utility frontier. These methods are feasible options for the data owner.

This situation raises the intriguing possibility of *combining SDC methods* in such a way, ideally, that the combined method is superior in terms of both disclosure risk and data utility than either alone. In practice, this ideal is not always attained, but often combined methods yield dramatic improvements in utility at the price of only modest increases in disclosure risk.

Focusing for simplicity in combining two methods, the intuition is as follows. The first method should be one that is good from the perspective of disclosure risk *and* not necessarily good in terms of data utility, but whose utility consequences can be characterized. Then, the second method should be chosen to "reverse" the utility consequences of the first without harming disclosure risk.

In §5, we report the results of a simulation study in which the first method is a particular form of microaggregation and various second methods are employed. §2 is a brief introduction to relevant SDC methods. In §3 we describe the disclosure risk and data utility methods used in the paper. A general framework for combining SDC methods is articulated in §4.

2 Summary of SDC Methods

Here we present briefly several perturbation methods for SDC.

2.1 Additive Noise

Additive noise [B02, K86, L93, SF89, TM94] consists of adding random noise to the original data. Generally, the noise distribution has mean zero, to preserve means. The variance of the noise distribution commonly reflects either complete independence or the correlation structure of the original, unmasked data. Often, the noise distribution is Gaussian. Specifically, let $\mathbf{X}$ be original multivariate data set with covariance matrix $\mathbf{\Sigma}_{\text{orig}}$. The corresponding masked data $\mathbf{X}'$ are

$$\mathbf{X}' = \mathbf{X} + \mathbf{E}, \tag{1}$$

where $\mathbf{E} \sim N(\mathbf{0}, c\mathbf{\Sigma}_{\text{orig}})$, with the constant c is selected by the data releaser. This procedure is parameterized by c. In the experiments in §5, we used a linear transformation of noise $\mathbf{E}$ such that the sample mean is exactly $\mathbf{0}$-vector and sample covariance matrix of $\mathbf{E}$ is exactly $c\mathbf{\Sigma}_{\text{orig}}$.

2.2 Rank Swapping

Rank swapping is a form of data swapping [DR82]. While originally designed for ordinal variables [M96], it can also be used for numerical variables. To implement rank swapping, we first rank the values of each variable X_i in ascending order. Each ranked value then is swapped with another ranked value randomly chosen within a restricted range. This process is repeated for each variable.

Rank swapping leaves (univariate) means and variances unchanged but may seriously affect the correlation structure of the data. It is parameterized by the parameter p that specifies that ranks of two swapped values cannot differ by more then p percent of the total number of records. Large values of p lead to greater distortions in the data whereas the smaller ones to higher disclosure risk.

2.3 Microaggregation

Microaggregation involves clustering records into groups of size at least k, where k is a parameter of the method [DN93]. Rather than release the original value of X for a given record, the componentwise average of the X's over all records in the cluster containing the given record is released. Classical microaggregation requires that all clusters, except perhaps one, be of size k.

Different variants of microaggregation exist, defined by the clustering method. These include: 1) Individual ranking, in which each variable is grouped independently of other variables; 2) Multivariate ranking, in which all the variables (or subsets of variables) are grouped by similarity of values; and 3) z-scores projection and principal components projection [A93, DA95, DN93], in which the multivariate data first are ranked by projecting them onto a single axis, using either the sum of z-scores or the first principal component, and then are aggregated into groups of size k, except possibly for one group of larger size (from k to $2k-1$ elements).

In this paper, we consider both multivariate microaggregation and projection microaggregation methods. Microaggregation with individual ranking will not be considered because recent empirical and analytical results ([DFMSOT02], [O04]) prove the lack of security for this method.

Microaggregation leaves means unchanged but decreases variances: the variance decrease by microaggregation is equal to $\frac{1}{N}\sum_{j=1}^{p}\sum_{j=1}^{k_j}\delta_{ij}^2$, where p is the number of clusters, k_j is number of records in the cluster j, δ_{ij} is distance between the centroid of cluster j and records i belonging to this cluster and N is the total number of records in the data file.

Regarding higher moments microaggregated sets are usually more leptokurtic than original data sets, especially microaggregation using sum of z-scores or principal component projection, because microaggregation substitutes the values of the records in a cluster by its means and produce a shrinkage of the data towards the center of mass of the distribution, thus increasing kurtosis.

3 Disclosure Risk and Data Utility

The SDC methods described in §2 affect data in different ways. The common feature for all SDC methods is that all of them are designed with two goals in a mind. First is to minimize disclosure risk, that is the risk to respondent confidentiality that the data releaser would experience as a consequence of releasing the data. And the second one is to maximize data utility, that is the value of the released data to a legitimate data user.

3.1 Disclosure Risk Measures

Two varieties of disclosure risk are usually considered. *Identification disclosure* occurs when an intruder[2] can associate a released record with the individual or establishment to which it pertains. Typically, identification disclosure is effected by record linkage (see, e.g., [O04]) to an external database containing identifiers. One measure of disclosure risk, then, is the percentage of masked records that are linked "correctly" to their parent records in the original data. This measure is used, for instance, in [KKORS06], as well as in §5.[3]

An *attribute disclosure* occurs when the intruder's target is an original value of a particular attribute. Attribute disclosure risk can be measured by the tightness of bounds for attribute values in the original data given the masked data, as in [DFKS02] in the context of tabular data.

3.2 Transparency Risk and Utility

An important, and largely unaddressed issue in SDC is how much a statistical agency can (from the risk perspective) or should (from the utility perspective) reveal regarding the methods it employs to protect released microdata, a practice we term *transparency*. To date, the issue has been considered solely from a risk standpoint. For example, agencies are fiercely protective of swap rates when rate swapping is employed, and will sometimes not even reveal which attributes have been swapped, or under what constraints [GKS06].

Less attention has been the data utility consequences of transparency. A compelling analogy exists in cryptography, where it is almost universal not to depend on hiding knowledge of cryptographic methods (as opposed to values of keys). In the setting of the example, this reasoning would argue for the agency's stating that noise had been added, and that it had covariance that is a multiple of $\mathbf{\Sigma}_{\text{orig}}$, but not revealing the value of c.

To illustrate, consider the masked data $\mathbf{X}'$ defined by (1). An agency would not release $\mathbf{X}'$, whose covariance $(1+c)\mathbf{\Sigma}_{\text{orig}}$ is not that of the original data. One strategy would be to withhold the value of c and release $\mathbf{X}'' = \frac{1}{\sqrt{1+c}}\mathbf{X}'$, which does have the same covariance as the original data $\mathbf{X}$. A more transparent strategy would be to release $\mathbf{X}'$ and c, in which case users could calculate $\mathbf{X}''$, and could also tailor analyses to the value of c. For example, ordinary regressions could be replaced by errors-in-variables models. But this strategy *is* risky. Given knowledge of $\mathbf{X}'$ and c, an intruder can construct confidence ellipsoids around masked records for corresponding original records. While these ellipsoids may not yield precise attribute disclosures, they may suffice for identity disclosure, especially for original data points in sparse regions.

[2] The generic term for an illegitimate use of the data.

[3] Care must be taken because "parent" is method-specific. For microaggregation with cluster size k, each masked record has, in effect, k parents.

As a second illustration, consider rank swapping. Suppose the intruder is interested in the record $x_i = (x_{i,1}, \dots, x_{i,t})$. When the parameter p of rank swapping is released together with the data, upper and lower bounds for the original value are:

$$\begin{aligned} x_u &= x_{i_1+\lfloor N*p \rfloor,1} \cdots, x_{i_t+\lfloor N*p \rfloor,t} \\ x_l &= x_{i_1-\lfloor N*p \rfloor,1} \cdots, x_{i_t-\lfloor N*p \rfloor,t}. \end{aligned} \tag{2}$$

Here x_u and x_l are upper and lower bounds; the first index of x denotes the rank of $x_{i,j}$ and the second is the index of the variable; N is the number of records in the data file. If the algorithm used to perform rank swapping were known in detail, then intervals narrower than those given by (2) could be obtained.

As a specific example, consider a data set of 1000 records in which variable j has a lognormal distribution. Suppose that the range of variable j is $[0.04, 25.57]$. Let rank swapping with $p = 0.05$ be applied to the data. Then upper and lower bounds for every masked (swapped) value are given by (2). An intruder using this knowledge can, by repeated simulation of the procedure, construct empirical distributions for different ranks. For most ranks the distribution of the intervals is very close to the uniform, but for the lowest and the largest ranks the shape of the distribution is triangular skewed to the left or right, becoming more and more uniform-like with the growth of the rank. The narrowest intervals are in the densest areas and could be as narrow as $[0, 42, 0.58]$ for this particular example.

3.3 Data Utility Measure

Proposed utility measures can be found in the recent literature regarding; see for example, [DFMST99], [O04] and [YWW02]. Many of these are based on differences between point estimates of the first and second moments of released data and corresponding estimates for the original data. In [KKORS06], measures are proposed based on differences between inferences based on original and released data. However, these measures are tailored to normally distributed data and one specific linear regression analysis.

In this paper, we adopt a broader utility measure, called *propensity score utility*, recently proposed in [WROK06]. This measure is both suitable for any distribution of the data and not tied to a particular data analysis.

Propensity score utility measures the distance between distributions of original and masked data by the means of classification of the pooled data in two groups: one corresponding to the original and other group corresponding to the masked data. We call an assignment of a particular record to the original data as treatment 1 and an assignment of the masked data, treatment 0. In symbols, let $r_i = 1$ if record x_i is assigned to the treatment 1 and $r_i = 0$ if it is assigned to the treatment 0. The propensity score $e(x)$ is the probability of being assigned to treatment 1 given the observed record $\mathbf{x}$: $e(\mathbf{x}) = P(r = 1|\mathbf{x})$. In [WROK06] it was shown that in order to test that the distribution of $\mathbf{x}$ is the same for treated and control records, all the records must have approximately the same propensity scores. That is, testing the differences of two sample distributions is equivalent to testing $e(\mathbf{x}) = c$ for all $\mathbf{x}$, where c is some constant. As in practice the propensity scores are unknown, they should be estimated by modeling propensity scores using logistic model or tree model.

Then, the propensity score utility measure is $\sum_{i=1}^{N}(\hat{e}(\mathbf{x_i}) - c)^2$. We suppose that the number of records in the original and masked files is the same, so that $c = 1/2$. Methods with higher utility have smaller values of this measure.

4 Combination of SDC Methods

In §3.2, we touched upon data utility and risks of attribute disclosure and identity disclosure when information about masking method and its parameters is released together with the data. While the utility benefits are clear, so is the risk.

To improve the security of the methods, and in particular to make it harder for an intruder to estimate bounds on original values, we propose to use several stages of masking, when different methods are used at different stages. Ideally, such combination of SDC methods can also improve utility.

4.1 An Example

In §2 we showed that different SDC methods have different properties. For example, data masked with noise has larger variance than the original, whereas data masked with microaggregation has smaller variance than the original. So, we might first mask the original data using microaggregation and then add noise to restore the "lost" variability in the data. An added benefit of doing so is that masked data values would be unique, whereas microaggregation produces masked data with k-tuples of identical values, thereby revealing the value of k.

At least two issues ensue. First, how much noise should be added to restore the variability in the data? Second, what distribution should be used to generate the noise?

To help make these issues concrete, suppose that microaggregation using z-scores projection was applied to the original data. Figure 1 contains scatterplot of a normally distributed two-dimensional data set (green circles) and corresponding microaggregated data (red circles). As it can be seen there, the microaggregated data set is a shrunk version of the original, and the degree of shrinkage is different in different directions. This happens because the only criterion to form a cluster for microaggregation with z-scores projection is the closeness of the sum of z-scores of the records. Even if the points are far away in the Euclidean sense, but have the same sum of z-scores, they can be chosen by the algorithm to form a cluster.In case of data with two variables, points with the same sum of z-scores are located along the line $x_2 = z - x_1$, where x_2 and x_1 are the scaled variables, and z is any constant. Therefore, clusters tend to be stretched in the direction of negative correlation.

To reverse the effect of microaggregation, we therefore need to "add more variation" in the direction of negative correlation than in the direction of positive correlation. One approach would be to add noise whose distribution if $N(O, \text{Cov}(\text{original}) - \text{Cov}(\text{micz}))$. We tried this approach, and it worked very well for certain types of analyses, such as linear regression, because in this case the released data have the same covariance matrix as the original data.

4.2 A General Formulation for Two-Stage Masking

For data that are not normally distributed, adding normal noise obviously would distort other moments. Ideally, the noise distribution should be chosen to reproduce the

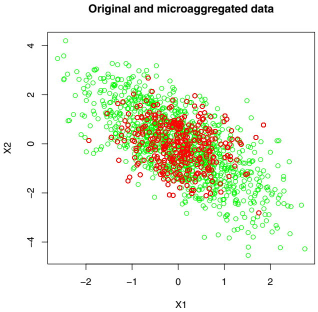


Fig. 1. Original and microaggregated data, using microaggregation with z-scores projection. Green circles correspond to the original data and red circles to the masked data.

distribution of the original data $\mathbf{X}$. Denoting by $\mathbf{X}'$ the first-stage masked data (in §4.1), then one would want the distribution of the noise $\mathbf{E}$ to be chosen so that

$$f_{\mathbf{X}'} * f_{\mathbf{E}} = f_{\mathbf{X}}, \tag{3}$$

where the f's are associated (estimated) density functions.

Computationally, solution of (3) is not feasible, especially for high-dimensional data. First of all, densities in (3) would need to be estimated, which is not possible in even moderately high dimensions. Second, (3) would need to be discretized and solved numerically, in which case zeroes of estimated density functions become problematic. Finally, of course, there is no guarantee that (3) has a positive solution integrating to one.

In low (for example, two) dimensions, solution of (3) is possible. Discretization of 2-space, say into squares, in order to solve (3) in effect then simply allows the noise distribution to be local; see §5 for further discussion.

One may ask, of course, why if we "knew" the density of the original data, we would not simply use it to generate synthetic data having the same distribution as the original data. In the context of (3), this amounts to choosing the first stage SDC method in such a way that $\mathbf{X}'$ is a constant. One way to do this is with microaggregation in which the cluster size equals the data set size. This is a relevant criticism, to which there are at least two rejoinders. First, reproducing the distribution of the original data is not the only measure of utility, and it may be that two-stage masking preserves other aspects of the original data not present in an independent replication.

Second, and more important, if one has only a poor estimate $\hat{f}_{\mathbf{X}}$, then any version of (3), including the synthetic data version, is not useful. Seen in this way, two-stage masking in which the second stage consists of adding noise may be the "best of both worlds:" in (1) readily modeled characteristics of the noise are captured in $\mathbf{E}$, while hard-to-model characteristics are retained in $f_{\mathbf{X}'}$. There remains, of course, the central question: is this reality or just wishful thinking? The simulation study in §5 addresses this question.

4.3 Multi-stage Masking

The paradigm of "capture the easy-to-model" in the noise and "leave the hard-to-model" in the masked data generalizes to multi-stage masking, in which more and more subtle features of the original data may be captured.

Let $\mathbf{X}_0$ denote the original data, $\mathbf{X}_1$ the result of the first-stage masking, and $\mathbf{X}_2$ the result of adding noise to $\mathbf{X}_1$. Then we can rewrite (1) as

$$\mathbf{X}_2 = \mathbf{X}_1 + \mathbf{E}. \tag{4}$$

The key point is that in (4), $\mathbf{E}$ is an approximation to $\mathbf{X}_0 - \mathbf{X}_1$, so that an agency could instead replace $\mathbf{E}$ by *a masked version of* $\mathbf{X}_0 - \mathbf{X}_1$.

Changing notation, let $\mathcal{M}_j(\cdot)$ denote the stage-j masking method, so that for example, $\mathbf{X}_1 = \mathcal{M}_1(\mathbf{X}_0)$. Then, (4) generalizes to

$$\mathbf{X}_2 = \mathbf{X}_1 + \mathcal{M}_2(\mathbf{X}_1 - \mathbf{X}_0) = \mathcal{M}_1(\mathbf{X}_0) + \mathcal{M}_2(\mathbf{X}_1 - \mathbf{X}_0). \tag{5}$$

But then, of course, there is no need to employ only two stages, leading to a more general version of (5):

$$\mathbf{X}_k = \sum_{j=1}^{k} \mathcal{M}_j(\mathbf{X}_{j-1} - \ldots - \mathbf{X}_0). \tag{6}$$

Harking back to the intuition articulate in §1, if the early stages of masking are good from the standpoint of keeping disclosure risk, then the later stages may be chosen to (incrementally) improve data utility. Of course, no stage should "undo" the utility improvements resulting from earlier ones.

Full investigation of (6) remains a topic for future research. In §5, we explore some of the possibilities, considering in particular how the selection of masking method should reflect the statistical characteristics of X_0.

5 Simulation Study

To understand the potential usefulness of the paradigms in §4, we generated a series of non-normally distributed data sets. We emphasize non-normal data, because this is a more complicated case in SDC in a sense that if the multivariate normal data sets, we could simply generate synthetic data with the same mean and covariance matrix as in the original data and the resulting data would have virtually maximal utility and minimal disclosure risk.

So, replicates of eight of two-dimensional data set were created by crossing the following three cases: (1) Whether the (non-normal) data distribution is symmetric (or not); (2) Whether the two variables in the data are highly correlated (or not); and (3) Whether the correlation between the two variables is positive or negative.The samples of size $n = 10,000$ were drawn from various data structures.

The SDC methods we combine are those usually considered as "quite distorting," for example, microaggregation using projections, multivariate microaggregation, rank swapping and noise generated approximately as $\mathbf{X}_0 - \mathcal{M}_1(\mathbf{X}_0)$. Not surprisingly, these methods also tend to have good disclosure risk properties, so in effect we are investigating whether data utility can be improved while maintaining the good disclosure risk behavior. We did not consider SDC methods with high risk of identity disclosure, such as microaggregation with individual ranking, or resampling.

Specifically, as first stage method we used microaggregation with z-scores projection with $k = 3$ records per cluster. The reason for this choice is that this method has the lowest disclosure risk among the methods with relatively good utility for non-normal data sets (see, for example, [KKORS06]). Moreover, microaggregation is a good potential candidate for combining with noise, because noise restores the variance in the data that is diminished by microaggregation. For two-dimensional negatively correlated data sets, microaggregation using z-scores projection is especially distorting, so improving utility improvement in this case would be especially valuable.

Regarding second-stage methods, in principle, any method can be used that can be described as "distortive." In our experiments we used:

micz03-micz03: Microaggregation with z-scores projection and $k = 3$ the number of records per cluster.

micz03-micpcp03: Microaggregation with principal component projection and $k = 3$ the number of records per cluster.

micz03-micmul10: Multivariate microaggregation with $k = 10$ the number of records per cluster.

micz03-rank1: Rank swapping with $p = 1$ the maximal percentage difference between ranks of swapped values.

micz03-noise100 "Coarse noise," with $p = 100$ the number of partitions for density estimation in (3).

Choices of the parameters described above were based primarily on the magnitude of $\mathbf{X}_0 - \mathscr{M}_1(\mathbf{X}_0)$. For example, projection microaggregation usually perturbs data more than multivariate microaggregation. Rank swapping even with small parameter may also perturb the data considerably ([DFMST99], [O04], [KKORS06]). So the values of k or p are based on these considerations and on values reported as the best in [DFMST99] and [O04].

The results of our experiment are presented in Tables 1 and 2, which contain the disclosure risk and data utility and values for all eight data sets and all five combined methods, as well as the first-stage masking—micz03—alone.

Methods of the risk-utility frontier are indicated using **boldface**. It is clear from these tables, that the combined methods significantly outperform Micz03 alone. Regarding identification disclosure risk, we can see that in four data sets out of eight, Micz03 alone is not on the frontier, which means that there is a combination that outperforms it in both utility and disclosure risk! For other data sets, disclosure risk is smaller for Micz03 alone than for the combinations, however the gain in utility is in general much more significant. Note also, that the risk remains very low for the combinations: less than one percent of records are correctly identified. The exception is Micz03-Micpcp03, which having high utility, also has quite high values of risk—up to 29% for some data sets. As noted in §3, high utility and low disclosure risk are conflicting goals, in the sense that methods leading substantially increase data utility also carry increased disclosure risk. So, these empirical results showed that the combinations could be very promising as in half of the cases both goals were attained! In general, what most combinations accomplish is dramatic increases in utility accompanied by only modest increases in risk.

Table 1. Disclosure risk values for the simulation study. Lower values represent lower risk.

	Symmetric				Non-symmetric			
	High Correlation		low Correlation		High Correlation		Low Correlation	
	Neg	Pos	Neg	Pos	Neg	Pos	Neg	Pos
micz03	0.0025	0.0036	0.0019	0.0024	0.0011	0.0043	0.0011	0.0012
micz03-noise100	0.0044	0.0077	0.0044	0.0076	0.0079	0.0144	0.0098	0.0039
micz03-micmul10	0.0046	0.0077	0.0025	0.0203	0.0947	0.1122	0.1265	0.0071
micz03-micpcp03	0.2516	0.0198	0.0029	0.0133	0.2926	0.0806	0.18	0.0477
micz03-micz03	0.0015	0.0275	0.0023	0.0035	0.0004	0.0033	0.0009	0.0033
micz03-rank1	0.0119	0.0092	0.0067	0.0087	0.0079	0.034	0.0091	0.0096

Table 2. Propensity score utility values for the simulation study. Smaller values represent higher utility.

	Symmetric				Non-symmetric			
	High Correlation		low Correlation		High Correlation		Low Correlation	
	Neg	Pos	Neg	Pos	Neg	Pos	Neg	Pos
micz03	281.51	**128.14**	**233.40**	**132.12**	639.42	592.07	639.04	**463.78**
micz03-noise100	**26.49**	**9.00**	16.97	**9.27**	**15.5**	**5.69**	**7.53**	**28.75**
micz03-micmul10	**16.53**	18.37	**12.97**	14.84	**9.15**	5.48	8.68	**11.15**
micz03-micpcp03	**9.31**	12.83	**9.33**	**7.86**	**3.39**	**4.99**	**5.76**	**8.61**
micz03-micz03	**28.30**	23.48	**33.92**	**37.27**	**180.10**	**45.09**	**94.10**	**40.04**
micz03-rank1	34.81	28.58	29.42	27.26	42.04	14.82	**26.45**	39.48

6 Discussion and Conclusion

In this paper we studied the possibility of combining SDC methods designed for the protection of numerical microdata. An additive, two-stage scheme for SDC was proposed. In the first stage, an SDC method is applied to the data and then some synthetic data ("noise"), which is a function of the difference between original and first-stage masked data is added.

Based on the simulation study described in §5, this approach seems very promising, mainly due to its iterative increase in date utility, whereas the identity disclosure risk remains low. Regarding attribute disclosure and transparency risk, we think that publishing the details of the masking algorithm and their parameters could be safer for the combinations of methods than for the single methods, since the problem of finding reliable bounds for original data values becomes more difficult when several stages of masking are applied. Thorough assessment of attribute disclosure of the combinations of SDC methods and single methods are topics of current research.

Acknowledgements

This research was supported by NSF grant EIA–0131884 to the National Institute of Statistical Sciences (NISS). Any opinions, findings, and conclusions or recommendations

expressed in this publication are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

[A93] N. Anwar, *Micro-aggregation—the small aggregates method*, Research Report, 1993. Luxembourg: Eurostat.

[B02] R. Brand, "Microdata protection through noise," in *Inference Control in Statistical Databases*, J. Domingo-Ferrer, ed., *Lecture Notes in Computer Science*, **2316**, 97–116, 2002. Berlin: Springer.

[DR82] T. Dalenius and S. P. Reiss, "Data-swapping: A technique for disclosure control," *Journal of Statistical Planning and Inference*, **6** 73–85, 1982.

[DA95] D. Defays and N. Anwar, "Micro-aggregation: a generic method," in *Proceedings of the 2nd International Symposium on Statistical Confidentiality*, 69–78, 1995. Luxembourg: Office for Official Publications of the European Communities.

[DN93] D. Defays and P. Nanopoulos, "Panels of enterprises and confidentiality: the small aggregates method", in *Proceedings of the 92 Symposium on Design and Analysis of Longitudinal Surveys*, 195–204, 1993. Ottawa: Statistics Canada.

[DFKS02] A. Dobra, S. E. Fienberg, A. F. Karr, and A. P Sanil, "Software systems for tabular data releases," *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems*, **10(5)**, 529–544, 2002.

[DFMS99] J. Domingo-Ferrer and J. M. Mateo-Sanz, "On resampling for statistical confidentiality in contingency tables," *Computers and Mathematics with Applications*, **38**, 13–32, 1999.

[DFMSOT02] J. Domingo-Ferrer, J. M. Mateo-Sanz, A. Oganian and A. Torres, "On the security of microaggregation with individual ranking: analytical attacks", *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems*, **10(5)**, 477–492, 2002.

[DFMST99] J. Domingo-Ferrer, J. M. Mateo-Sanz and V. Torra, "Comparing SDC methods for microdata on the basis of information loss and disclosure risk of disclosure control methods." In *Proc. of ETK–NTTS '2001*, 807–825. Luxembourg: Eurostat.

[GKS06] S. Gomatam, A. F. Karr and A. P. Sanil, "Data swapping as a decision problem," *Journal of Official Statistics*, to appear, 2006. Available on-line at www.niss.org/dgii/technicalreports.html.

[J89] M. A. Jaro, "Advances in record-linkage methodology as applied to matching the 1985 Census of Tampa, Florida," *Journal of the American Statistical Association*, **84**, 414–420, 1989.

[H93] G. R. Heer, "A Bootstrap procedure to preserve statistical confidentiality in contingency tables." In *Proceedings of the International Seminar on Statistical Confidentiality*, (D. Lievesley, ed.), 261–271, 1993. Luxembourg: Office for Official Publications of the European Communities.

[KKORS06] A. F. Karr, C. N. Kohnen, A. Oganian, J. P. Reiter and A. P. Sanil "A framework for evaluating the utility of data altered to protect confidentiality", The American Statistician, **60(3)**, 224–232, 2006.

[K86] J. J. Kim, "A method for limiting disclosure in microdata based on random noise and transformation." In *Proceedings of the ASA Section on Survey Research Methodology*, 303–308, 1986. Alexandria VA: American Statistical Association.

[KLP75] H. T. Kung, F. Luccio, F.P. Preparata, "On finding the maxima of a set of vectors", *J. ACM*, 22:469–476, 1975.

[L93] R. J. A. Little, “Statistical analysis of masked data,” *Journal of Official Statistics*, **9**, 407–426, 1993.

[MDF99] J. M. Mateo-Sanz and J. Domingo-Ferrer, “A method for data-oriented multivariate microaggregation.” In *Proceedings of Statistical Data Protection’98*. Luxembourg: Office for Official Publications of the European Communities, 89–99, 1999.

[M96] R. Moor, “Controlled data swapping techniques for masking public use microdata sets.” U.S. Census Bureau, 1996.

[O04] A. Oganian, *Security and Information Loss in Statistical Database Protection.* Ph. D. thesis, Universitat Politecnica de Catalunya, 2004.

[ODF04] A. Oganian and J. Domingo-Ferrer, “On the complexity of optimal microaggregation for statistical disclosure control,” *Statistical Journal of the United Nations Economic Commission for Europe*, **18(4)**, 345–353, 2001.

[PS98] D. Pagliuca and G. Seri, “Some results of individual ranking method on the system of enterprise accounts annual survey,” Esprit SDC Project, Deliverable MI-3/D2, 1999.

[SF89] G. Sullivan and W. A. Fuller, “The use of measurement error to avoid disclosure.” In *Proceedings of the ASA Section on Survey Research Methodology*, 802–807, 1989. Alexandria VA: American Statistical Association.

[TM94] P. Tendik and N. Matloff, “A modified random perturbation method for database security,” *ACM Transactions on Database Systems*, **19(1)**, 47–63, 1994.

[WROK06] M.-J. Woo, J. P. Reiter, A. Oganian and A. F. Karr,“Global measures of data usefulness for microdata altered for disclosure limitation.” Technical Report, National Institute of Statistical Sciences, 2006.

[YWW02] W. E. Yancey, W. E. Winkler and R. H. Creecy, “Disclosure risk assessment in perturbative microdata protection.” In Domingo-Ferrer, editor, *Inference Control in Statistical Databases*, 135–152. Berlin: Springer-Verlag.

A Fixed Structure Learning Automaton Micro-aggregation Technique for Secure Statistical Databases

Ebaa Fayyoumi and B. John Oommen*

School of Computer Science, Carleton University, Ottawa, Canada : K1S 5B6
efayyoum@scs.carleton.ca,
oommen@scs.carleton.ca

Abstract. We consider the problem of securing statistical databases and, more specifically, the micro-aggregation technique (MAT), which coalesces the individual records in the micro-data file into groups or classes, and on being queried, reports, for the all individual values, the aggregated means of the corresponding group. This problem is known to be NP-hard and has been tackled using many heuristic solutions. In this paper we present the first reported Learning Automaton (LA) based solution to the MAT. The LA modifies a fixed-structure solution to the *Equi-Partitioning Problem* (EPP) to solve the micro-aggregation problem. The scheme has been implemented, rigorously tested and evaluated for different real and simulated data sets. The results[1] clearly demonstrate the applicability of LA to the micro-aggregation problem, and to yield a solution that obtains a lower information loss when compared to the best available heuristic methods for micro-aggregation.

1 Introduction

A lot of attention has recently been dedicated to the problem of maintaining the confidentiality of statistical databases through the application of statistical tools, so as to limit the identification of information on individuals and enterprises. The objective in statistical databases is to guarantee the confidentiality of the information provided, and to simultaneously provide useful (unbiased) statistical summaries of the data to the user [1].

One of the most recent techniques proposed involves the strategy called "Micro-aggregation". The latter comprises of a family of statistical disclosure limitation techniques used to protect micro-data files containing records on individual data subjects. These belong to the family of substitution/perturbation approaches [2,3,4], where individual values are replaced by values computed on small aggregates *prior* to publication.

* Chancellor's Professor and Fellow of the IEEE.

[1] The authors would like to thank Dr. Domingo-Ferrer for all his support, and advice, and for providing us with the real-life data sets. We are also thankful to him, Dr. Mateo-Sanz, and Dr. Francesc Sebe for answering our queries. Without their help, we would not have been able to produce these results.

J. Domingo-Ferrer and L. Franconi (Eds.): PSD 2006, LNCS 4302, pp. 114–128, 2006.

The Micro-Aggregation Problem (MAP) as formulated in [3,4,5,6], can be stated as follows: A micro-data set $\mathcal{U} = \{U_1, U_2, \ldots, U_n\}$ is specified in terms of the n "individuals", namely the $U_i's$, each representing a data vector whose components are p continuous variables. Micro-aggregation involves partitioning the n data vectors into m groups so as to obtain a k-partition $\mathbb{P}_k = \{G_i \mid 1 \leq i \leq m\}$, such that each group, G_i, of size n_i, contains between k and $2k-1$ data vectors. The optimal k-partition $\mathbb{P}_k^*$ is defined to be the one that maximizes the within-group homogeneity, which is defined as the *Sum of Squares Error* (SSE) computed on the basis of the Euclidean distances of each individual data vector X_{ij} to the centroid $\bar{X}_i$ of the group to which it belongs, and is given by: $SSE = \sum_{i=1}^{m} \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_i)^T (X_{ij} - \bar{X}_i)$. Analogously, the between-groups similarity is defined as the *Sum of Squares Among* the groups (SSA), and is given as: $SSA = \sum_{i=1}^{m} n_i (\bar{X}_i - \bar{X})^T (\bar{X}_i - \bar{X})$. The *Total Sum of Squares* is denoted by $SST = SSA + SSE$. The *Information Loss* is quantified as: $L = \frac{SSE}{SST}$.

Contribution of the Paper

The main contribution of this paper is to demonstrate that the information loss, which can be obtained from a MAT, can be reduced by using two criteria, namely that of maintaining the total SSE to be as minimum as possible, and simultaneously by studying the relations between the individual records in the micro-data file. This, in turn, is achieved by invoking the newly proposed Object Migrating Micro-aggregated Automaton ($OMMA$). The paper thus demonstrates the power of LA in minimizing the information loss, leading to results comparable to those obtained from the best available heuristic methods for micro-aggregation such as the *Maximum Distance to Average Vector* ($MDAV$) [7], and the *Minimum Spanning Tree Partitioning Algorithm* (MST) [6]. The observed reduction in the information loss that can be noticed when micro-aggregating *multivariate* data (which sometimes exceeds 13% compared to the $MDAV$ and MST) renders the contribution of this paper significant. But apart from this we argue that the applicability of LA to the MAT provides a promising strategy to effectively protect sensitive data in the micro-data file.

It should be mentioned that as argued in [8], the $OMMA$ can actually be applied to many types of attributes - continuous, ordinal or nominal.

2 State-of-Art $MATs$

Initial research in the field proposed "fixed " $MATs$ which required that the size of each group was a fixed constant, k. These, in turn, led to the *Fixed-Size Micro-aggregation* algorithm [9,10]. Recent developments [3,4,11,12] have concentrated on further reducing the information loss by using variable-sized data-dependent groups, leading to families of *Data-Oriented Micro-aggregation* algorithms. The philosophy that is utilized is the fact that groups need to consist of *at least* k data vectors. They also preserve the natural data aggregate by allowing the group size to be between k and $2k-1$, depending on the structure of the data, so as to lead to more homogenous groups.

Two alternative heuristic approaches which incorporate variable-size micro-aggregation have been presented in [3,11,13]. The authors of [3] presented a genetic algorithm that appears as an alternative linear heuristic. It presents the k-partitions as a binary string, and combines directed and random search strategies to attain a global optimum. A hierarchical classification method can be used to obtain building blocks for the heuristic MAT such as the k-Ward's algorithm [4,11], which was extended to a *Secure-k-Ward* scheme in order to enhance the individual's privacy [14]. In order to enhance the micro-aggregation speed, optimized versions of the latter were proposed in [15].

An efficient polynomial algorithm to solve the univariate MAP was presented in [5]. Here, optimal partitions were shown to correspond to the shortest path on a graph. It is necessary to highlight that all the above described univariate $MATs$ can easily be extended to multivariate $MATs$ using any projection method[2]. Our aim is to develop a multivariate MAT without any projection.

The first algorithm to accomplish this was proposed in 2002 by Domingo-Ferrer *et al.* [3], called the *Maximum Distance to Average Vector* ($MDAV$). It micro-aggregates the multivariate micro-data file based on the concept of the diameter distance of the data set. In 2005, an enhanced version of the $MDAV$ appeared in [7], and was implemented as a built-in technique in the μ-ARGUS Software tool version 4.0 [16]. The modification is based on computing the centroid of the data set instead of computing its diameter, in order to increase the micro-aggregation speed, and to reduce the information loss.

The computation of the Minimum Spanning Tree leads to another multivariate MAT [6]. This yields a new clustering algorithm obtained by splitting the minimum spanning tree using a constraint on the minimum group size. The MST can be considered to be a potential strategy for any practical application.

3 Learning Automata (*LA*)

The functionality of the LA can be described as a sequence of repetitive feedback cycles. The feedback cycle involves two entities, the *Random Environment* and the LA. During each cycle the automaton chooses an action, which triggers a response from the Environment, and uses the received response - that can be either a reward or a penalty- with the knowledge gained from the previous cycles to determine which is the next action to be chosen. By the process of learning, the automaton adapts itself to the Environment and determines the optimal action, i.e., the action which has the minimum penalty probability.

Incorporating LA in any application domain is an evidence of the power of the philosophy. Basically, LA learn from the random environment. The actual technique involved in applying the LA philosophy in the different applications involves modeling the actions, simulating the transforming functions, and representing the system's output, in order to have reward or penalty responses. This is where the creativity of the researcher becomes apparent.

[2] The projection of multivariate data vectors onto a single-axis, can be done by using either the First Principal Component, the Sum of Z-scores, or a particular variable.

Of all classes of LA, the pioneering ones are those which belong to the Fixed Structure Stochastic Automata ($FSSA$) families. These $FSSA$ have the property that their transition and output functions do not change with time. These LA seem to possess powerful properties useful for solving different NP-hard problems, as we will show in this paper.

The basic idea used to solve the MAP is based on a sub-class of LA solutions that have been used to solve the object partitioning problem [17,18]. As documented in the literature, the object partitioning problem involves partitioning a set of $|\mathbb{P}|$ objects into $|\mathbb{N}|$ groups, where the main aim is to partition the objects into groups that mimic an unknown grouping. In other words, the objects which are accessed together must reside in the same group [18]. In the special case when all the groups are required to contain the same number of objects, the problem is referred to as the *Equi-partitioning Problem* (EPP). Many solutions involving LA have been proposed to solve the EPP, but the most efficient algorithm is the *Object Migrating Automaton* (OMA), which was proposed by Oommen and Ma [18], and some modifications were added by Gale *et al.* [17].

The OMA assumes that it has a sequence of queries, where each query is represented in this form $< O_1, Q_2 >$ indicating that these two objects are accessed together. Because these two objects are more often accessed together, they are then migrated between the different groups based on the current group of the object set $\mathbb{P}$, and the current query $< O_1, O_2 >$. The OMA tries to group the two objects of a query on the same group. It is important for the convergence of the OMA that the random sequence of the queries reflects an optimal partition. We refer the interested reader to [8,18] for more information regarding the fundamentals of LA, and the OMA.

3.1 Restrictions of the OMA to the MAP

The reported instances of the OMA are not directly applicable for the MAP. To develop our solution, we highlight the necessary enhancements which must be added to the OMA in order for it to be useful here.

- In case of the MAP, the user does not have access to the stream of random queries. Rather, the only available data is the micro-data file. It is apparent that we have to artificially "generate" a sequence of "queries" which can be used to operate on the OMA. Also, for the EPP, the placement of the objects in the automaton and the stream of random queries, together, serve to either reward or penalize the automaton. However, in the case of the MAP, the question of obtaining a reward/penalty response is not provided by the user, but it has to be inferred.
- Unlike the EPP, which has no way of penalizing "non accessed elements", a solution to the MAP must develop a strategy for penalizing such records by considering how similar the records within the same groups are. In the present problem, it is crucial that an automaton can quantify how fitting a record is for any given group.

- The optimal partition for the EPP yields crucial information in the stream of random queries. As opposed to this, in the context of the MAP, the system has no notion of how to characterize the optimal partition.
- The definition of the optimal partition for the EPP is quite different from that of the corresponding solution for the MAP. In the case of the EPP, all objects which are accessed together more frequently should be in the same partition, while in the MAP all similar records should be in the same group.
- As explained in [8], the criteria which are used to reward and penalize the automaton in the EPP is quite unlike the one used for the MAP.
- Although the EPP and MAP will utilize analogous rules for a reward phenomenon, as we shall see, they differ in performing the penalty rules. In case of the EPP, the automaton enforces the rule that the pertinent object migrates, if and only if at least one of the accessed objects is at the boundary state of the different partitions. As opposed to this, in the MAP, the automaton enforces the rule that the records are migrated to another group whenever migrating the object reduces the overall SSE.
- The automaton used to solve the EPP is said to have converged, when all the objects are found in the most (or the last two) internal states of each partition. However, we propose that the convergence in the MAP occur when the *measure* of the information loss is unchanged.

4 Object Migrating Micro-aggregated Automaton (*OMMA*)

In this Section, we define the Object Migrating Micro-aggregated Automaton ($OMMA$) as an 8-tuple as below: ($\mathbb{U}$, $\underline{\Phi}$, $\underline{\alpha}$, $\mathbb{B}$, $\mathbb{Q}$, $\mathbb{G}$, $\mathbb{D}$, $\mathbb{L}$), where

- $\mathbb{U} = \{U_1, U_2, \ldots, U_n\}$ is the micro-data file.
- $\underline{\Phi} = \{\phi_1, \phi_2, \ldots, \phi_{hM}\}$ is the set of states.
- $\underline{\alpha} = \{\alpha_1, \alpha_2, \ldots, \alpha_h\}$ is the set of h actions, each representing a group into which the records of $\mathbb{U}$ must fall.
- $\mathbb{Q}$ is the transition function, explained presently, which specifies how the records should move between the various states.
- The function $\mathbb{G}$ partitions the set of states for the groups. For each action group α_j, there is a set of states $\{\phi_{(j-1)M+1,\ldots,\phi_{jM}}\}$, where M is the depth of memory. Thus, $G(\phi_i) = \alpha_j$, where $(j-1)M+1 \leq i \leq jM$. This means that the record in the automaton chooses α_1, if it is in any of the first M states, etc. We assume that $\phi_{(j-1)M+1}$ is the most internal state of action α_j, and that ϕ_{jM} is the boundary state. These are called the state of *"Maximum Certainty"* and *"Minimum Certainty"*, respectively.
- $\mathbb{D}$ is the similarity distance matrix.
- $\mathbb{L}$ is the similarity list specifying the records deemed to be collectively similar. It is stored as a list of triples of term $< R_i, R_{i+1}, 1 >$ (where i is odd) are stored, and where the records included are those for which the similarity index is greater than or equal to a predefined threshold, θ.

The $OMMA$ process (where n is a multiple of k) consists of 8 phases involved to generate the micro-aggregated data file as follows: (1) Loading the micro-data file, (2) Standardizing the data, (3) Building the distance matrix, (4) Constructing the similarity list, (5) Constructing the group structure, (6) Distributing the records among the automaton groups (actions) randomly, (7) Performing the learning cycle, which is where the MAP is solved, and (8) Generating the micro-aggregated file. The overall schematic for this application is given pictorially in [8], and omitted here in the interest of brevity.

We now specify how the variables can be standardized after loading the data file. The standarization process replaces the individual values x_i for a specific variable V_j, by $(x_i - \bar{X}_j)/\sigma_j$, where $\bar{X}_j$ and σ_j are, respectively, the average and the standard deviation of the values taken by the variable V_j. The importance of this phase is evident when we consider multivariate micro-data [8].

The next issue to be considered is that of specifying how the records themselves are to be compared. This depends on the type of the relevant attribute, and whether it is continuous, nominal, or ordinal. Here, we quantify the similarity between continuous attributes in terms of their "Euclidean" distance.

The next phase involves creating the symmetric $N \times N$ matrix (with a zero along the main diagonal). This poses no major handicap, especially since Solanas *et al.* have recently demonstrated that when the number of records is very large, the distance matrix can be stored by applying the blocking technique [19]. Each entry in the matrix, $D(R_i, R_j)$, is computed using the Euclidean metric between the two records. The similarity matrix is used in constructing the similarity list.

The first step involved in building the similarity list, $\mathbb{L}$, requires comparing the elements of the distance matrix to a certain threshold, θ. The question of determining θ is open. As it stands now, we have used a single heuristic: θ is quite simply set as: $\theta = \frac{1}{2}[Max_{i,j}\ D(i,j) - Min_{i,j}\ D(i,j)]$, and is used to make the scheme arrive at an efficient, and hopefully optimal, grouping. The next phase involves the migration. We migrate pairs of records between the groups based on the following similarity metric: If $D(R_i, R_j)$ is greater than or equal to θ, we add an element $< R_i, R_j, 1 >$ to the similarity list, where "1" represents the activity of this tuple. The rationale of the activity status will be explained presently.

The learning phase is the core of the clustering, and it utilizes the philosophy of the OMA without requiring any background knowledge of the true optimal partition. It is important to highlight that the different states within a given group quantify the measure of certainty that the scheme has for a given record belonging to that group. The model is initialized by placing all the records in the boundary state of their initially randomly-chosen groups. This indicates that the scheme is initially uncertain of the placement of the records. As the learning cycle proceeds, similar records will be rewarded for their being together in the same group, and they will be penalized by either moving toward their boundary state, or to another group.

Before starting the learning cycle, and after distributing the records randomly among the groups, we have to crystallize the group structure by computing the value of the SSE, and this quantity involves the summation of the values for the records. Subsequently, we have to assign the index of each record to the group it belong to. This will minimize the time required in the learning phase.

The $OMMA$ moves into its main learning loop by setting the old value of the information loss to be equal to ∞ (a large positive number), and by then processing the constructed similarity list, $\mathbb{L}$, one tuple at a time as follows. We first check the activity attribute in the tuple $< R_i, R_j, 1 >$. If this quantity is equal to "1", the two records will either be rewarded or penalized depending on their states in the automaton. Otherwise, the automaton will not process this tuple, and it will ignore it by processing the next tuple in the similarity list. This is because, the distance between the records is, by itself, not adequate to measure the similarity between them. Indeed, we have to also maintain the total SSE to be as small as possible. Based on these two criteria, at the beginning, all the tuples in the similarity list are rendered active in the first cycle, but as the cycles proceed, some tuples are retained to be active while others are rendered inactive. The activity of the tuple is set to "0" after being sure that, although these two records are close to each other, their being in the same group will lead to *increasing* the total SSE. Thus, if we are reasonably sure that these two records will not be in the same group, reprocessing this tuple in the coming cycles will merely aid to increasing the processing time.

The list $\mathbb{L}$ is now traversed repeatedly, and similar records in the active tuple are processed. If they are both assigned to the same group, the automaton is rewarded. However, if they are assigned to distinct groups, the automaton is penalized. After $\mathbb{L}$ has been processed, we compare the newly computed value of the information loss, L_{new}, with the old value, L_{old}, produced in the previous cycle. If $L_{new} < L_{old}$, the learning phase continues by entering the next learning cycle. But, if $L_{new} = L_{old}$, the learning phase terminates and the micro-aggregated file is generated. It should be mentioned that L_{new} cannot be greater than L_{old}, because the records will move, if and only if, the SSE remains the same or is reduced. Clearly, this leads to the same value of the information loss or to a smaller value, relative to the one obtained in the previous cycle.

We now describe the actual transitions represented by $\mathbb{Q}$ for each of these operations.

1. Transitions for Rewards
   On being rewarded, since the elements R_i and R_j are in the same group, say, α_u, both of them are moved toward the most internal state of that group, one step at a time. See Figure 1.a in Appendix A.
2. Transitions for Penalties
   This case is encountered, when two similar records,R_i and R_j, are allocated in distinct groups. say, α_u and α_v respectively (*i.e.,* R_i is in state ζ_r, where $\zeta_r \epsilon \{\Phi_{(u-1)M+1}, \cdots, \phi_{uM}\}$, and R_j is in state ζ_s, where $\zeta_s \epsilon \{\Phi_{(v-1)M+1}, \cdots, \phi_{vM}\}$). In this case, they are moved as follows:
   - Case 1: If both ζ_r and ζ_s are not the boundary states ϕ_{uM} and ϕ_{vM}, respectively, R_i and R_j are moved one state toward the corresponding boundary state. See Figure 1.b in Appendix A .
   - Case 2: This occurs when at least one of R_i or R_j is in the boundary state of *Minimum Certainty*, say, $\zeta_s = \phi_{uM}$. Studying the effect of migrating any record R_x in α_u with R_j, under the condition that $R_x \neq R_i$, requires

investigating the effect of the potential moves on the summation of the SSE_u and SSE_v, which we shall refer to as SSE_{uv}. Here we need to consider the $k-1$ different swapping possibilities. If the new value SSE_{uv} is less than or equal to the previous value of SSE_{uv}, we have to physically swap the chosen record R_x, which leads to the minimum SSE_{uv} value, with R_j and proceed to update the group structure. Subsequently, we assign both R_j and R_x to the boundary state of α_u and α_v, respectively. Otherwise, the tuple $< R_i, R_j, 1 >$ will be deactivated by setting the active attribute to "0". This is clarified in Figure 2 in Appendix A.

- Case 3: If both R_i and R_j are at the boundary states of their different groups, say, $\zeta_r = \phi_{uM}$ and $\zeta_s = \phi_{vM}$, we have to study all the different possibilities of swapping any record R_x, under the condition that $R_x \neq R_i$ and $R_x \neq R_j$, in α_u or α_v with R_j or R_i, respectively, on the SSE_{uv}. We then choose the option that leads to reducing the value of SSE_{uv}. In such a case, we physically swap the records and update the group structure, besides, assigning both records to the boundary state of the group which they belong to. This scenario is described in Figure 3 in Appendix A.

For sake of completeness, the discussed scheme is algorithmically described in [8], and omitted here in the interest of brevity.

5 Experimental Result

5.1 Data Sets

The $OMMA$ has been rigorously tested and the results obtained seem to be very promising. We have tested them on two benchmark real data sets, which have been used as benchmarks in previous studies [3,20]: (i) The **Tarragona Data Set** contains 834 records with 13 variables [3], and (ii) The **Census Data Set** contains $1,080$ records with 13 variables [20].

To further investigate the scalability of the $OMMA$ with respect to data size, dimensionality, and the group size, we have also tested it with two simulated multivariate data sets generated using *Matlab's* built-in-functions with the following parameters: (i) **Uniform distribution** (min=0; max=1000). (ii) **Normal distribution** (μ=0; σ=0.05). The results for the Normal distribution are included in Appendix A, and the related discussion can be found in [8].

5.2 Results

For a given value of the minimum group size k, we compared the percentage value of the information loss $L = SSE/SST$ (as defined in Section 1) resulting from the $OMMA$ and the $MDAV$[3] strategies.

[3] The $MDAV$ was implemented based on the centroid concept and not a diameter concept. We did not program the $MDAV$ scheme. We are extremely thankful to Dr.Francesc Sebe for giving us his source code.

Table 1. Comparison of the percentage of the information loss between the *MDAV* and the *OMMA* on the Tarragona and Census data sets (Univariate Methods). In this case the value of k was set to be $k = 3$.

Tarragona Data Set				**Census Data Set**			
Variable	*MDAV*	*OMMA*		Variable	*MDAV*	*OMMA*	
	I.L.	I.L.	Converge		I.L.	I.L.	Converge
Var1	7.15200	7.15199	2	Var1	0.13155	0.13155	4
Var2	0.63586	0.63586	4	Var2	0.00138	0.00138	4
Var3	0.51702	0.51702	5	Var3	0.00828	0.00828	5
Var4	1.48854	1.48854	4	Var4	0.00489	0.00489	4
Var5	1.69394	1.69394	5	Var5	0.02449	0.02449	4
Var6	0.47503	0.47503	4	Var6	0.03262	0.03262	4
Var7	1.96623	1.96623	5	Var7	0.00171	0.00172	4
Var8	0.42182	0.42182	4	Var8	0.43418	0.43418	5
Var9	1.28625	1.28627	4	Var9	0.72176	0.72176	5
Var10	1.74929	1.74929	4	Var10	0.00611	0.00611	5
Var11	2.58368	2.58367	3	Var11	0.01353	0.01353	4
Var12	4.14703	4.14703	5	Var12	0.00689	0.00689	5
Var13	5.00563	5.00563	4	Var13	0.00808	0.00808	8

Table 1 shows a comparison between the *MDAV* and the *OMMA*. In this table, both strategies have been applied on univariate data sets, for each of the "Tarragona" and "Census" real data sets containing 13 continuous variables, and the value of k was set to 3 in all the experiments. The results clearly show that the *OMMA* obtains values of the information loss exactly same as those obtained by the *MDAV* scheme. The reason behind this agreement is the existence of only one optimal solution. It is worth mentioning, that the *OMMA* reaches the minimum value of the information loss with an average of 4 successive learning cycles in the Tarragona data set, while 5 cycles are required, on the average, to reach the minimum value of the information loss in the Census data set. The computation time required to micro-aggregate each variable independently was as low as 1.93 seconds (on the average) for the Tarragona data set, while for the Census data set it was, on the average, about 2.93 seconds. Thus, we can unequivocally conclude that for uni-variate micro-aggregation the time required to micro-aggregate all the individual records to reach the minimum information loss increases with respect to the data size. Besides, the number of learning cycles, required to lead to the minimum value of the information loss, is proportional to the computation time.

But the power of the *OMMA* is clearly shown, when it is used on multivariate data sets, since the *OMMA* has the ability to measure the similarity between the individual records based on two different criteria. The first criterion, C_1, involves studying the relation between the records quantified in terms of the distance between each record and the other records. As opposed to this, the second criterion, C_2, attempts to maintain the *SSE* as low as possible. The

Table 2. Comparison of the percentage of the information loss and the computation time between the *MDAV* and the *OMMA* on the Tarragona and Census data sets for multivariate methods

Data Set	k value	*MDAV*		*OMMA*			Improv. (%)
		I.L.	Time	I.L.	Time	Converge	
Tarragona	2	9.27500	0.281	9.24351	5.438	7	0.34%
	3	16.96611	0.203	15.12901	11.687	5	10.83%
	6	26.40474	0.109	24.26087	36.078	5	8.12%
Census	2	3.16518	0.50	3.25152	8.70	6	-2.73%
	3	5.65353	0.33	5.22899	20.16	5	7.51%
	4	7.44143	0.25	6.76231	30.00	4	9.13%
	5	8.88401	0.27	8.09004	48.84	4	8.94%
	6	10.19413	0.24	9.14287	63.97	4	10.31%

power of the *OMMA* is that it is used to prioritize between these two criteria, and to reflect them in the penalization and reward responses. In the reward case, C_1 is given a higher priority than C_2. Similarly in the penalization rule, in Case 1, processing the active tuple and moving both records one state towards the boundary state lends C_1 a higher priority than C_2. But, in Case 2 and Case 3 in the penalization rules, the *OMMA* forces the corresponding records to migrate to groups that preserve the number of records in each group, and to move towards the absolute minimum value of the *SSE*. This obviously awards C_2 a higher priority than C_1.

Table 2 shows the results of using the *OMMA* on multivariate real data sets. This time we micro-aggregate *all* the individual records with all the variables simultaneously. In this case, we have tested the *OMMA* with different values of k in order to determine the effect of increasing the number of records per group. It seems to be apparent that, increasing the number of records per group tends to increase the value of the information loss beside increasing the required computational time. But, in this case, in spite of increasing the required time, the number of learning cycles required to reach the minimum value of the information loss is inversely proportional to the computation time. However, further investigations are required to improve our understanding of the increment in the computational time as well as the reduction in the information loss.

Table 2 shows that the values of the information loss measured by the *OMMA* are less than the corresponding values measured by the *MDAV*. The percentage of improvement in the information loss is as high as 10.83% in the Tarragona data set when $k = 3$, requiring only 11.68 seconds. Similarly, in the Census data set the improvement is as high as 10.31% when $k = 6$, requiring 63.97 seconds. In term of comparison, we believe that, minimizing the information loss is more important than minimizing the computational time.

Because we are using the same recommended data sets are used in [6], we compare our information loss values with the corresponding loss values yielded by the *MST*. In the Tarragona data set, the percentage improvement in the information loss is up to 3.02% when $k = 3$. But, in the Census data set, the percentage improvement in the information loss reaches up to 3.52% when $k = 3$, 6.34% when $k = 4$, and 8.17% when $k = 5$. This clearly demonstrates how the $OMMA$ competes in a superior manner against all the state-of the art strategies.

Finally, Table 3 studies the performance of the $OMMA$ and $MDAV$ with respect to other three issues.

Issue 1: To investigate the scalability with respect to the data set size, we have tested both the schemes on the Uniform and the Normal distributions with data set cardinalities of 1200, 2400, 300, 4500, and 5400 with 16 variables and $k = 3$. The impact of increasing the size of the data set leads to minimizing the information loss value, increasing the computational time and, in the case of the $OMMA$, increasing the number of learning cycles required to reach to the minimum value of the information loss. The percentage of the improvement in the information loss that the $OMMA$ obtains, ranges from 6.69% to 3.03% when the data size equals $1,200$ and $4,500$, respectively for the uniform distribution.

Issue 2: To Investigate the scalability with respect to dimensionality, we have also tested both strategies on the Uniform and the Normal distributions for different numbers of variables, including 12, 14, 16, 18, 20, 22 and 24, when the data size was set to $3,000$ and the value of k was set to 3. Again, we observe that, the value of the information loss is proportional to the number of the variables used in the micro-aggregation process. The interesting point here is that the computational time required to micro-aggregate all the individual records seems to be inversely proportional to the dimensionality in the $OMMA$ scheme, but proportional for the $MDAV$ scheme. The highest percentage of the improvement in the information loss is 4.04% when the number of variables is 16 for the Uniform distribution.

Issue 3: The scalability of the $OMMA$ and the $MDAV$ regarding the group size has also been studied for both the Uniform and the Normal distributions, where the group sizes were 3, 4, 5, 6, 8, 10, and 12, and for the data set cardinality of $2,400$, with the dimensionality of 16. Here, the value of the information loss increases with the number of records per group. In the $OMMA$ the computational time required to micro-aggregate the records increases with the number of records per group, as can be justified [8]. Thus, increasing the group size leads to increase the efficiency of the $OMMA$ by leading to a minimum value of the information loss, as opposed to the $MDAV$. In the Uniform distribution the percentage of improvement for the information loss is as high as 13.27% which is obtained when the group size equals 12.

6 Conclusions

In this paper we have presented, to our knowledge, the first reported Learning Automaton (LA)-based solution to the Micro-Aggregation Problem (MAP). We have shown that our newly devised scheme, the $OMMA$ can successfully be used to micro-aggregate a multivariate micro-data file. The $OMMA$ competes in a superior manner to the state-of the art $MDAV$ and MST methods in minimizing the loss in the information. The percentage of improvement reaches up to 13% and 8% when compared to the $MDAV$ and MST schemes, respectively, on real-life data sets. The proposed strategy also scales well with respect to the size of the data set, the dimensionality, and the group size. Preliminary tests show that the $OMMA$ can thus be highly recommended for advantageous micro-aggregation.

In conclusion, our work has demonstrated the intractability of the micro-aggregation problem and presented a promising tool for solving it. We foresee three venues for future work. First of all, we propose to extend the $OMMA$ for the Data-Oriented micro-aggregation, where the group size, n_i, satisfies $k \leq n_i < 2k$. A second avenue for future work is to enhance the $OMMA$ scheme for other families of LA, especially those of a variable structure. Finally, we have to investigate how much actual run time is required to carry out $OMMA$ compared to $MDAV$ on data sets that are very large, typically over 50,000 records.

References

1. Adam, N., Wortmann, J.: Security-Control Methods for Statistical Databases: A comparative Study. ACM Computing Surveys **21**(4) (1989) 515–556
2. Baeyens, Y., Defays, D.: Estimation of Variance Loss Following Microaggregation by the Individual Ranking Method. In: Proceedings of Statistical Data Protection'98, Luxembourg: Office for Official Publications of the European Communities (1999) 101–108
3. Domingo-Ferrer, J., Mateo-Sanz, J.: Practical Data-Oriented Microaggregation for Statistical Disclosure Control. IEEE Trans. on Know. and Data Eng. **14**(1) (2002) 189–201
4. Mateo-Sanz, J., Domingo-Ferrer, J.: A Method for Data-Oriented Multivariate Microaggregation. In: Proceedings of Statistical Data Protection'98, Luxembourg: Office for Official Publications of the European Communities (1999) 89–99
5. Hansen, S., Mukherjee, S.: A Polynomial Algorithm for Univariate Optimal Microaggregation. IEEE Trans. on Know. and Data Eng. **15**(4) (2003) 1043–1044
6. Laszlo, M., Mukherjee, S.: Minimum Spanning Tree Partitioning Algorithm for Microaggregation. IEEE Trans. on Know. and Data Eng. **17**(7) (2005) 902–911
7. Domingo-Ferrer, J., Torra, V.: Ordinal, Continuous and Heterogeneous k-Anonymity Through Microaggregation. Data Mining and Knowledge Discovery **11**(2) (2005) 195–212
8. Fayyoumi, E., Oommen, B.: (Using Learning Automaton to Micro-Aggregate the Continuous Micro-data File) Unabridged Version of This Paper.
9. Defays, D., Anwar, N.: Micro-Aggregation: A Generic Method. In: Proceedings of the 2nd International Symposium on Statistical Confidentiality, Luxembourg: Office for Official Publications of the European Communities (1995) 69–78
10. Defays, D., Nanopoulos, P.: Panels of Enterprises and Confidentiality: the Small Aggregates Method. In: Proceedings of 92 Symposium on Design and Analysis of Longitudinal Surveys, Ottawa: Statistics Canada (1993) 195–204
11. Mateo-Sanz, J., Domingo-Ferrer, J.: A Comparative Study of Microaggregation Methods. Questiio **22**(3) (1998) 511–526
12. Solanas, A., Martínez-Ballesté, A.: V-MDAV: A Multivariate Microaggregation With Variable Group Size. In: 17th COMPSTAT Symposium of the IASC, (2006)

13. Domingo-Ferrer, J., Mateo-Sanz, J.: On Resampling for Statistical Confidentiality in Contingency Tables. Computers and Mathematics with Applications **38** (1999) 13–32
14. Li, Y., Zhu, S., Wang, L., Jajodia, S.: A privacy-enhanced microaggregation method. In: FoIKS '02: Proceedings of the Second International Symposium on Foundations of Information and Knowledge Systems, London, UK, Springer-Verlag (2002) 148–159
15. Fayyoumi, E., Oommen, B.: On Optimizing the k-Ward Micro-Aggregation Technique for Secure Statistical Databases. (In: 11th Australasian Conference on Information Security and Privacy Proceeding)
16. Hundepool, A., Wetering, A., Ramaswamy, R., Franconi, L., Capobianchi, A., Wolf, P., Domingo-Ferrer, J., Torra, V., Brand, R., Giessing, S.: M-ARGUS Version 4.0 Software and User's Manual. (2004)
17. Gale, W., Das, S., Yu, C.: Improvements to an Algorithm for Equipartitioning. IEEE Trans. Comput. **39**(5) (1990) 706–710
18. Oommen, B., Ma, D.: Deterministic Learning Automata Solutions to the Equipartitioning Problem. IEEE Transction Computer **37**(1) (1988) 2–13
19. Solanas, A., Martínez-Ballesté, A., Domingo-Ferrer, J., Mateo-Sanz, J.: A 2d-Tree-Based Blocking Method for Microaggregating Very Large Data Sets. In: The First International Conference on Availability,Reliability and Security. The International Dependability Conference Bridging Theory and Practice. (2006)
20. Domingo-Ferrer, J., Torra, V.: A Quantitative Comparison of Disclosure Control Methods for Microdata. In Doyle, P., Lane, J., Theeuwes, J., Zayatz, L., eds.: Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies, Amesterdam: North-Holland, Berlin: Springer-Verlag (2002) 113–134

Appendix A

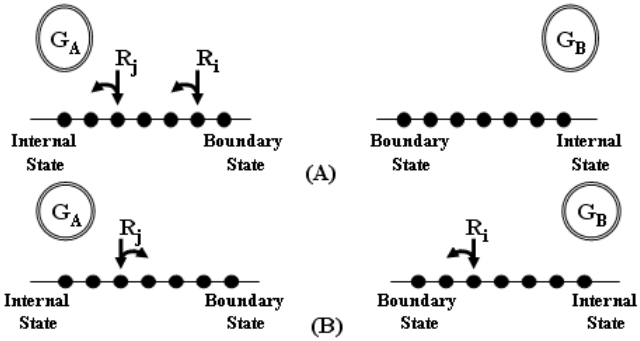


Fig. 1. (A) The automaton is rewarded, since both R_i and R_j are similar and located in the same group. (B) The automaton is penalized, since R_i and R_j are similar but located in distinct groups. None of them at the boundary state.

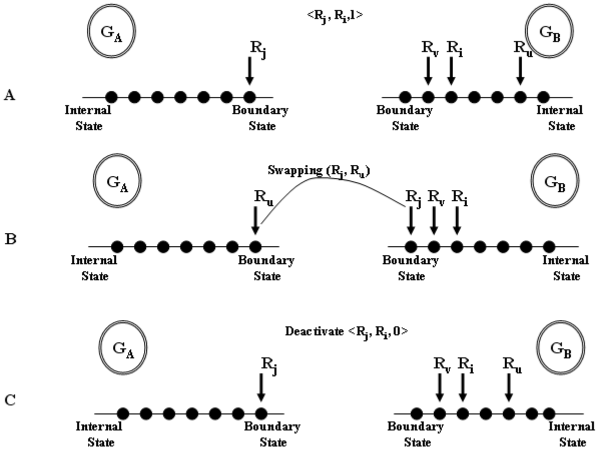


Fig. 2. (A) The automaton is penalized, since R_i and R_j are similar but located in distinct groups. R_j is at the boundary state, while R_i is not at the boundary state. After searching for the most suitable record which leads to the minimum amount of the SSE (*i.e., $k = 3$*, we have two choices), a physical swapping between (R_u, R_j) is done in (B), while in (C) there is no record which leads to a smaller value of SSE, the tuple $< R_i, R_j, 1 >$ is deactivated.

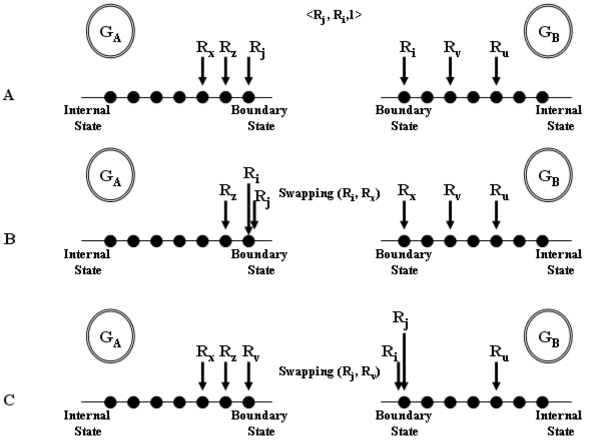


Fig. 3. (A) The automata is penalized, since R_i and R_j are similar but located in distinct groups. However none of them is in a boundary state. After searching for the most suitable record which leads to the minimum amount of the SSE (*i.e., $k = 3$*, we have four choices), a physical swapping between (R_i, R_x) is done in (B) while (R_j, R_v) in (C).

Table 3. Comparison of the percentage of the information loss and the computation time between the *MDAV* and the *OMMA* on simulated Uniform and Normally distributed data sets for multivariate methods

Investigating scalability with respect to data size

Data Set	No. of Records	No. of Groups	Dim.	k value	*MDAV*		*OMMA*			Improv. (%)
					I.L.	Time	I.L.	Time	Converge	
Uniform Distribution	1,200	400	16	3	27.67671	0.42	25.82418	0.66	6	6.69
	2,400	800			24.64686	1.28	23.71665	5.13	6	3.77
	3,000	1000			23.66435	2.00	22.69465	9.88	8	4.10
	4,500	1500			22.38678	4.45	21.60147	21.64	7	3.51
	5,400	1800			21.71397	6.42	21.05564	33.05	7	3.03
	6,000	2000			21.38408	7.95	20.63458	40.58	8	3.50
Normal Distribution	1,200	400	16	3	27.74187	0.36	26.22697	0.65	7	5.46
	2,400	800			24.80187	1.28	24.00719	3.76	6	3.20
	3,000	1000			23.82211	1.99	23.06686	6.45	7	3.17
	4,500	1500			22.29773	4.45	21.57716	16.23	6	3.23
	5,400	1800			21.62495	6.46	20.96986	29.97	9	3.03

Investigating scalability with respect to data dimensionality

Data Set	No. of Records	No. of Groups	Dim.	k value	*MDAV*		*OMMA*			Improv. (%)
					I.L.	Time	I.L.	Time	Converge	
Uniform Distribution	3,000	1000	12	3	17.55012	1.61	16.92778	20.75	7	3.55
			14		20.80110	1.81	20.05538	12.38	6	3.59
			16		23.64962	2.00	22.69465	9.88	8	4.04
			18		26.07906	2.23	25.23600	6.25	7	3.23
			20		28.59153	2.45	27.45749	4.17	7	3.97
			22		30.51047	2.67	33.41902	1.97	6	-9.53
			24		32.14988	2.84	39.14110	1.94	6	-21.75
Normal Distribution	3,000	1000	12	3	17.65475	1.59	16.85623	29.66	6	4.52
			14		20.93491	1.81	20.30601	18.66	9	3.00
			16		23.82211	1.99	23.06686	6.47	7	3.17
			18		26.29432	2.24	25.36438	5.03	7	3.54
			20		28.62054	2.42	27.56436	3.67	7	3.69
			22		30.57049	2.64	29.41102	2.97	6	3.79
			24		32.11612	2.98	38.03268	1.96	6	-18.42

Investigating scalability with respect to the number of records per group

Data Set	No. of Records	No. of Groups	Dim.	k value	*MDAV*		*OMMA*			Improv. (%)
					I.L.	Time	I.L.	Time	Converge	
Uniform Distribution	2,400	800	16	3	24.64686	1.27	23.71665	5.13	6	3.77
		600		4	31.39147	1.28	29.17778	7.28	7	7.05
		480		5	36.55690	1.30	33.12187	10.42	7	9.40
		400		6	40.37307	1.30	36.19842	15.89	9	10.34
		300		8	46.01553	1.30	40.65105	19.64	7	11.66
		240		10	49.91325	1.30	43.61896	28.50	7	12.61
		200		12	53.22947	1.30	46.16533	37.58	7	13.27
Normal Distribution	2,400	800	16	3	24.80187	1.28	24.00719	3.71	6	3.20
		600		4	31.68612	1.31	29.27261	8.19	10	7.62
		480		5	36.54596	1.33	33.24953	8.23	8	9.02
		400		6	40.23421	1.31	36.30028	13.55	8	9.78
		300		8	46.32321	1.30	41.14944	21.86	9	11.17
		240		10	50.22509	1.33	43.99177	27.86	7	12.41
		200		12	52.78059	1.32	46.64860	37.38	7	11.62

Optimal Multivariate 2-Microaggregation for Microdata Protection: A 2-Approximation

Josep Domingo-Ferrer and Francesc Sebé

Rovira i Virgili University of Tarragona
Department of Computer Engineering and Maths
Av. Països Catalans 26
E-43007 Tarragona, Catalonia
{josep.domingo, francesc.sebe}@urv.cat

Abstract. Microaggregation is a special clustering problem where the goal is to cluster a set of points into groups of at least k points in such a way that groups are as homogeneous as possible. Microaggregation arises in connection with anonymization of statistical databases for privacy protection (k-anonymity), where points are assimilated to database records. A usual group homogeneity criterion is within-groups sum of squares minimization SSE. For multivariate points, optimal microaggregation, *i.e.* with minimum SSE, has been shown to be NP-hard. Recently, a polynomial-time $O(k^3)$-approximation heuristic has been proposed (previous heuristics in the literature offered no approximation bounds). The special case $k = 2$ (2-microaggregation) is interesting in privacy protection scenarios with neither internal intruders nor outliers, because information loss is lower: smaller groups imply smaller information loss. For 2-microaggregation the existing general approximation can only guarantee a 54-approximation. We give here a new polynomial-time heuristic whose SSE is at most twice the minimum SSE (2-approximation).

Keywords: Clustering, Statistical databases, Statistical disclosure control, Privacy-preserving data mining, Microaggregation.

1 Introduction

Microaggregation [7,8] is a technique for privacy in statistical databases, a discipline also known as statistical disclosure control (SDC). It is used to mask individual records in view of protecting them against re-identification. More generally, microaggregation can be mathematically modeled as a special kind of clustering problem where the goal is to cluster a set of p-dimensional points (the records in the SDC application) into groups of at least k points in such a way that groups are as homogeneous as possible. For the sake of concreteness, we talk about records rather than points in what follows.

Let $\mathbf{X}$ be a p-dimensional dataset formed by n records, that is, the result of observing p attributes on n individuals. Attributes will be assumed numerical (continuous) in this paper. Microaggregation is operationally defined in terms of two steps. Given a parameter k, the first step partitions records of $\mathbf{X}$ into groups

J. Domingo-Ferrer and L. Franconi (Eds.): PSD 2006, LNCS 4302, pp. 129–138, 2006.

of at least k records each. The second step replaces each record by the centroid of its group to obtain the masked dataset $\mathbf{X}'$. In a microaggregated dataset, no re-identification within a group is possible, because all k records in a group are identical: the best that an intruder can hope is to track what is the group where a target individual has been masked into.

Microaggregating with minimum information loss has been known to be an important —and difficult— issue ever since microaggregation was invented as an SDC masking method for microdata. However, it was often argued that optimality in SDC is not just about minimum information loss but about the best tradeoff between low information loss and low disclosure risk. The recent application [11] of microaggregation to achieve k-anonymity [21,20,24,25] for numerical microdata leaves no excuse to circumvent the problem of trying to reduce information loss as much as possible: once a value k is selected that keeps the re-identification risk low enough, the only job left is to k-anonymize (that is, to microaggregate) with as little information loss as possible.

A partition P such that all of its groups have size at least k is called a k-partition [8] and microaggregation with parameter k is sometimes denoted as k-microaggregation.

In [8], optimal microaggregation is defined as the one yielding a k-partition maximizing the within-groups homogeneity. The rationale is that, the more homogeneous the records in a group, the less variability reduction when replacing those records by their centroid (average record) and thus the less information loss. The within-groups sum of squares SSE is a usual measure of within-groups homogeneity in clustering [27,12,15,16], so a reasonable optimality criterion for a k-partition $P = \{G_1, \ldots, G_g\}$ is to minimize SSE, *i.e.* to minimize

$$SSE(P) = \sum_{i=1}^{g} \sum_{j=1}^{|G_i|} (x_{ij} - c(G_i))'(x_{ij} - c(G_i))$$

where $|G_i|$ is the number of records in the i-th group, $c(G_i)$ is the mean record (centroid) over the i-th group and x_{ij} is the j-th record in the i-th group. It was shown in [8] that groups in the optimal k-partition have sizes between k and $2k-1$.

The optimal microaggregation problem has been shown to be NP-hard in the multivariate case, that is, when $p > 1$ ([19]). Therefore, algorithms for multivariate microaggregation are heuristic [6,8,22,17,18].

1.1 Contribution and Plan of This Paper

In [10], the first approximation algorithm in the literature to optimal multivariate microaggregation was described. For any integer $k \geq 2$, the SSE of the k-partition P provided by the heuristic given in [10] is shown to verify

$$SSE(P) \leq 2(2k-1)[\max(2k-1, 3k-5)]^2 SSE(P^{opt})$$

where P^{opt} is the optimal k-partition.

When using microaggregation to protect a dataset, the lower k, the lower SSE and the less information loss caused. Define an internal intruder as an intruder who has contributed one or more records to the dataset. In the presence of internal intruders or outliers, $k > 2$ should be chosen, so that an internal intruder cannot exactly guess the contribution of the other individual in her/his group. However, if internal intruders are unlikely and there are no outliers, a value as low as $k = 2$ would do for anonymity (2-anonymity): groups of records of size between $k = 2$ and $2k - 1 = 3$ are formed and each record in a group is replaced by the group average record (2-microaggregation).

Thus 2-microaggregation is a relevant case deserving specific attention. For $k = 2$, the heuristic in [10] guarantees a bound $SSE(P) \leq 54 \cdot SSE(P^{opt})$, even though empirical results show that $SSE(P)$ is usually well below that bound. We propose in this paper a new heuristic for 2-microaggregation yielding a 2-partition P for which we can prove that $SSE(P) \leq 2 \cdot SSE(P^{opt})$.

Section 2 gives some background on the minimum-weight $[1, 2]$-factor problem and its algorithmic solution. Section 3 presents the 2-approximation heuristic for 2-microaggregation. The 2-approximation bound is proven in Section 4. Section 5 gives empirical results on the actual performance of the 2-approximation heuristic. Section 6 is a conclusion.

2 Background: The Minimum-Weight $[1, 2]$-Factor Problem

Given a graph $G = (V, E)$ and a function $w : E \to \mathbb{R}$ that assigns a weight to each edge, the minimum-weight $[1, 2]$-factor problem consists of finding the spanning subgraph F_{min} of G that satisfies:

- Each node in F_{min} has degree 1 or 2
- The sum of weights of edges in F_{min} is minimum.

This problem can be solved in strongly polynomial time [23], *i.e.* in running time bounded polynomially by a function only of the inherent dimensions of the problem (number of edges and nodes) and independent of the sizes of the numerical data. The graph library GOBLIN [13] solves the minimum weight factor problem over a weighted graph $G = (V, E)$ (with $|V| = n$ and $|E| = m$) by transforming the graph into a balanced flow network N_G [14] consisting of $n' = 2n + 4$ nodes and $m' = 2m + 4n + 6$ edges and solving a minimum weight balanced flow over N_G. This solution determines a minimum weight factor over G. The Enhanced Primal Dual Algorithm [14] solves this problem in $O(n'^2 m')$ time.

In this way, a [1,2]-factor problem over a complete graph G having n nodes (and $n(n - 1)/2$ edges) is solved polynomially in $O(n^4)$ time.

3 A 2-Approximation Algorithm for 2-Microaggregation

Next, we present a 2-approximation for the multivariate 2-microaggregation problem. Our solution adapts for 2-microaggregation a corrected version of the [1] and [2] approach to 2-anonymizing categorical data through partial suppression. The algorithm in [1] and [2] could not be used as published, as it relies on [4] to solve a minimum weight $[1, 2]$-factor, a problem not dealt with by [4], but by the substantially more recent literature mentioned in Section 2.

Algorithm 1 (2-μ-Approx)

1. *Given a dataset* $\mathbf{X}$*, build a weighted complete graph* $\mathbf{G} = (\mathbf{V}, \mathbf{E})$ *as follows:*
   (a) *Each record of* $x \in \mathbf{X}$ *is mapped to a different vertex* $v \in \mathbf{V}$.
   (b) *Given two vertices* $v, v' \in \mathbf{G}$ *corresponding to records* $x, x' \in \mathbf{X}$*, the edge* $vv' \in E$ *(the one linking nodes* v *and* v'*) is assigned weight* $w(vv') = d(x, x')^2$*, where* $d(x, x')$ *is the Euclidean distance between* x *and* x'.
2. *Compute the minimum weight [1,2]-factor* $\mathbf{F}_{min}$ *of graph* $\mathbf{G}$ *(see Section 2). By optimality,* $\mathbf{F}_{min}$ *consists only of connected components with a single edge (two vertices) or two adjacent edges (three vertices).*
3. *The 2-partition* P *is obtained by mapping each connected component in* $\mathbf{F}_{min}$ *to the group in* P *containing the records corresponding to the vertices in the component.*
4. *Microaggregate* $\mathbf{X}$ *based on* P.

4 The 2-Approximation Bound

We exploit in this section the properties of Algorithm 2-μ-Approx to prove that it yields a 2-approximation to optimal 2-microaggregation. We first give some notation, then a preliminary lemma and finally the theorem with the approximation bound.

Given a 2-partition $P = \{G_1, \ldots, G_g\}$ of $\mathbf{X}$, such that all groups have size 2 or 3, we denote by $SSE(G_i)$ the within-group sum of squares of group G_i, that is, $SSE(G_i) = \sum_{j=1}^{|G_i|}(d(x_{ij}, c(G_i)))^2$, where $c(G_i)$ is the centroid of group G_i.

Consider the complete graph $\mathbf{G}$ built in Step 1 of Algorithm 2-μ-Approx. Define $T(G_i)$ as the minimum-weight component of a [1,2]-factor of $\mathbf{G}$ containing the vertices corresponding to records in G_i. If G_i consists of two records, $T(G_i)$ contains a single edge connecting the two corresponding vertices. If G_i consists of three records, $T(G_i)$ contains the minimum-weighted two edges connecting the three corresponding vertices.

Lemma 1. *For any group* $G_i \in P$ *consisting of two or three records, it holds that*

$$\frac{1}{2} \leq \frac{SSE(G_i)}{w(T(G_i))} \leq 1$$

where $w(T(G_i))$ *is the sum of weights of edges in* $T(G_i)$.

Proof

i) Consider a two-record group $G_i = \{x_{i1}, x_{i2}\}$ and let $d(x_{i1}, x_{i2})$ be the Euclidean distance between both records. When microaggregating G_i, both records will be replaced by their mean vector $c(G_i)$ (*i.e.*, the centroid of G_i). It holds that

$$d(x_{i1}, c(G_i)) = d(x_{i2}, c(G_i)) = \frac{d(x_{i1}, x_{i2})}{2}$$

Thus,

$$SSE(G_i) = 2 \cdot \left(\frac{d(x_{i1}, x_{i2})}{2}\right)^2 = \frac{(d(x_{i1}, x_{i2}))^2}{2}$$

On the other hand, by construction of the graph **G** in Algorithm 2-μ-Approx, we have $w(T(G_i)) = (d(x_{i1}, x_{i2}))^2$. Thus, for any group G_i with two records it holds that

$$\frac{SSE(G_i)}{w(T(G_i))} = 1/2.$$

ii) Let us now take a three-record group $G_i = \{x_{i1}, x_{i2}, x_{i3}\}$. Its corresponding minimum weight tree $T(G_i)$ consists of three vertices v_{i1}, v_{i2}, v_{i3} and the two minimum-weight edges connecting them. Let us denote $d(x_{i1}, x_{i2}) = d_1$, $d(x_{i1}, x_{i3}) = d_2$ and $d(x_{i2}, x_{i3}) = d_3$. It is well known that, in a triangle, the sum of the squared lengths of the sides is three times the sum of the squared vertex-centroid distances. In our notation, this equality can be written as

$$d_1^2 + d_2^2 + d_3^2 = 3 \cdot SSE(G_i) \tag{1}$$

Without loss of generality, we consider that the edges of $T(G_i)$ are $e_1 = v_{i1}v_{i2}$ and $e_2 = v_{i1}v_{i3}$. By the minimality of $T(G_i)$ and using Equation (1), we get

$$w(T(G_i)) = d_1^2 + d_2^2 \leq (2/3)(d_1^2 + d_2^2 + d_3^2) = 2 \cdot SSE(G_i)$$

Thus,

$$\frac{SSE(G_i)}{w(T(G_i))} \geq 1/2$$

Another fact of elementary geometry is that, for any triangle, the sum of the squared lengths of any two sides is at least one third of the sum of the squared lengths of three sides. Using this, we can write

$$w(T(G_i)) = d_1^2 + d_2^2 \geq (1/3)(d_1^2 + d_2^2 + d_3^2) = SSE(G_i)$$

Thus,

$$\frac{SSE(G_i)}{w(T(G_i))} \leq 1$$

□

Theorem 1 (2-Approximation bound). *If P is a 2-partition found by Algorithm 2-μ-Approx and P^{opt} is the optimal 2-partition, then $SSE(P) \leq 2 \cdot SSE(P^{opt})$.*

Proof: Consider the minimum weight [1,2]-factor $\mathbf{F}_{min}$ of graph $\mathbf{G}$ computed at Step 2 of Algorithm 2-μ-Approx. Let us denote its cost, that is the sum of its edge weights, as $w(\mathbf{F}_{min})$. By Lemma 1, for any group $G_i \in P$ it holds that

$$SSE(G_i) \leq w(T(G_i)) \tag{2}$$

Extending Inequality (2) for all $G_i \in P$ and taking into account that $T(G_i)$ are the components of $\mathbf{F}_{min}$, we get

$$SSE(P) \leq w(\mathbf{F}_{min}) \tag{3}$$

Let us now take the optimal k-partition P^{opt} for the dataset $\mathbf{X}$. For each group $G_i^{opt} \in P^{opt}$, we take its corresponding vertices in $\mathbf{G}$ and connect them with one edge (if G_i^{opt} consists of two records) or the two minimum-weighted adjacent edges (if G_i^{opt} consists of three records); call the resulting graph component $T(G_i^{opt})$. The union of all $T(G_i^{opt})$ is a non-minimum weight [1,2]-factor $\mathbf{F}$ for $\mathbf{G}$. By Lemma 1 we know that $w(T(G_i^{opt})) \leq 2 \cdot SSE(G_i^{opt})$. Applying this inequality to all clusters, we get

$$w(\mathbf{F}) \leq 2 \cdot SSE(P^{opt}) \tag{4}$$

On the other hand, by definition of minimum weight [1,2]-factor

$$w(\mathbf{F}_{min}) \leq w(\mathbf{F}) \tag{5}$$

If we combine Inequalities (3),(4) and (5), the 2-approximation bound of the theorem follows. □

5 Empirical Results

We will show in this section that the new 2-approximation heuristic can perform even better than the best microaggregation heuristics in the literature in terms of low within-groups sum of squares SSE. We have used two reference datasets from the European project "CASC" [3]:

- The "Tarragona" dataset contains 834 records with 13 numerical attributes corresponding to financial information on 834 companies located in the area of Tarragona, Catalonia. The "Tarragona" dataset was used in the "CASC" project and in [8,18,10].
- The "EIA" dataset contains 4092 records with 11 numerical attributes (plus two additional categorical attributes not used here). This dataset was used in the "CASC" project and in [5,10] and partially in [18] (an undocumented subset of 1080 records from "EIA", called "Creta" dataset, was used in the latter paper). For the sake of speed, we have used in our experiments reported below a block with only the first 600 records of the "EIA" dataset; we call "EIA-600" the resulting dataset.

Table 1 gives the information loss under various methods for different values of k. For each case, SSE and $L_{SSE} = 100 \times SSE/SST$ are given, where SST

is the total sum of squares (sum of squared Euclidean distances from all records to the dataset centroid). The advantage of L_{SSE} is that it is bounded within the interval $[0, 100]$. The methods considered in the comparison include the best heuristics in the literature, according to the comparison in [10], namely:

- An improved version of the heuristic in [8] called MDAV (Maximum Distance to Average Vector, [11]). MDAV is the microaggregation method implemented in the μ-Argus package [17] resulting from the "CASC" project.
- The μ-Approx general approximation heuristic described in [10].
- The 2-μ-Approx heuristic proposed in this paper.

It can be seen that 2-μ-Approx yields the lowest SSE for the "Tarragona" dataset. For the "EIA-600" dataset, 2-μ-Approx ranks second after MDAV. Anyway, the differences in terms of L_{SSE} are not really meaningful. Furthermore, note that even if MDAV can slightly outperform the approximation heuristics for particular datasets, the latter have the advantage of always guaranteeing an SSE within a known multiple of the minimum SSE; this is especially valuable when that multiple is as small as twice the minimum SSE, as is the case for 2-μ-Approx.

The price paid to get the 2-approximation is that, since 2-μ-Approx basically requires to solve a minimum-weight $[1, 2]$-factor, it runs in time $O(n^4)$ (see Section 2), whereas MDAV and the general approximation μ-Approx run in $O(n^2)$. For example, the time needed to run 2-μ-Approx on the EIA-600 dataset is 81 minutes and 17 seconds, whereas MDAV and mu-Approx take a few seconds. Nonetheless, this is less painful than it would appear at first sight: all heuristics being at least quadratic-time, blocking attributes must always be used to microaggregate large datasets, so the only adaptation needed to run an $O(n^4)$ heuristic is to take smaller blocks.

Table 1. Information loss measures for the "Tarragona" and "EIA-600" datasets under various microaggregation heuristics ($k = 2$)

	Method	SSE	L_{SSE}
"Tarragona"	MDAV	1005.59	9.27499
	μ-Approx	1148.32	10.5914
	2-μ-Approx	958.496	8.84058
"EIA-600"	MDAV	59.2535	1.06927
	μ-Approx	66.8219	1.20585
	2-μ-Approx	65.8851	1.18895

Finally, we give some experimental results on how close to optimality are the partitions obtained using 2-μ-Approx. In order to be able to find the optimal 2-partition by exhaustive search, we are constrained to using very small datasets. We have taken the third reference dataset in [3], called "Census", which contains 1080 records with 13 numerical attributes and was used in the CASC project and [9,5,28,18,11,10]. From the "Census" dataset, we have drawn 10 random samples of $n = 15$ records with $p = 13$ attributes each. Those samples have been 2-microaggregated optimally by exhaustive search and also heuristically using 2-μ-Approx. For each sample, Table 2 shows the optimal SSE, the SSE obtained with 2-μ-Approx and the ratio between the former and the latter. It can be seen that such a ratio is 1 or close to 1 in all cases. Thus, even if the approximation bound only guarantees that the SSE obtained with 2-μ-Approx is no more than twice the optimum, it actually tends to be very close to the optimum.

Table 2. Optimal SSE vs SSE obtained with 2-μ-Approx ($k = 2$) for 10 random samples drawn from the "Census" dataset ($n = 15$ and $p = 13$)

Sample	Optimal SSE	SSE 2-μ-Approx	Ratio
1	12.042	12.042	1
2	12.2066	12.8186	0.9522
3	14.8156	14.8156	1
4	12.5545	12.5545	1
5	51.6481	52.0665	0.9920
6	74.1998	74.1998	1
7	15.6783	16.5705	0.9462
8	9.9702	9.9702	1
9	21.4293	22.7064	0.9438
10	33.3882	33.3882	1

6 Conclusion

The polynomial-time 2-approximation presented here improves for $k = 2$ on the general $O(k^3)$-approximation for multivariate microaggregation. Even though 2-microaggregation is not usable if internal intruders are likely or outliers are present, it can be an interesting option to implement 2-anonymity in other cases, because it results in low information loss and thus in high data utility. Thus, the availability of a 2-approximation for 2-microaggregation is relevant. Suggested directions for future research include: i) to devise heuristics that, for specific values of k other than 2, provide better approximations than the general $O(k^3)$-approximation; ii) to adapt Algorithm 2-μ-Approx to come up with an approximation to 2-microaggregation of non-numerical (categorical) microdata (categorical microaggregation was defined in [26]).

Acknowledgments

Thanks go to David Hartvigsen and Gérard Cornuéjols for their advice on the [1, 2]-factor literature. We also acknowledge partial support by the Government of Catalonia under grant 2005 SGR 00446, by the Spanish Ministry of Science and Education under project SEG2004-04352-C04-01 "PROPRIETAS" and by Eurostat-European Comission under contract "CENEX-SDC".

References

1. G. Aggarwal, T. Feder, K. Kenthapadi, R. Motwani, R. Panigrahy, D. Thomas, and A. Zhu. Anonymizing tables. In T. Eiter and L. Libkin, editors, *Proceedings of ICDT'2005*, volume 3363 of *Lecture Notes in Computer Science*, pages 246–258, Berlin Heidelberg, 2005.
2. G. Aggarwal, T. Feder, K. Kenthapadi, R. Motwani, R. Panigrahy, D. Thomas, and A. Zhu. Approximation algorithms for k-anonymity. *Journal of Privacy Technology*, 2005. Paper no. 20051120001.
3. R. Brand, J. Domingo-Ferrer, and J. M. Mateo-Sanz. Reference data sets to test and compare sdc methods for protection of numerical microdata, 2002. European Project IST-2000-25069 CASC, http://neon.vb.cbs.nl/casc.
4. G. P. Cornuéjols. General factors of graphs. *Journal of Combinatorial Theory*, B45:185–198, 1988.
5. R. Dandekar, J. Domingo-Ferrer, and F. Sebé. Lhs-based hybrid microdata vs rank swapping and microaggregation for numeric microdata protection. In J. Domingo-Ferrer, editor, *Inference Control in Statistical Databases*, volume 2316 of *Lecture Notes in Computer Science*, pages 153–162, Berlin Heidelberg, 2002. Springer.
6. D. Defays and N. Anwar. Micro-aggregation: a generic method. In *Proceedings of the 2nd International Symposium on Statistical Confidentiality*, pages 69–78, Luxemburg, 1995. Eurostat.
7. D. Defays and P. Nanopoulos. Panels of enterprises and confidentiality: the small aggregates method. In *Proc. of 92 Symposium on Design and Analysis of Longitudinal Surveys*, pages 195–204, Ottawa, 1993. Statistics Canada.
8. J. Domingo-Ferrer and J. M. Mateo-Sanz. Practical data-oriented microaggregation for statistical disclosure control. *IEEE Transactions on Knowledge and Data Engineering*, 14(1):189–201, 2002.
9. J. Domingo-Ferrer, J. M. Mateo-Sanz, and V. Torra. Comparing sdc methods for microdata on the basis of information loss and disclosure risk. In *Pre-proceedings of ETK-NTTS'2001 (vol. 2)*, pages 807–826, Luxemburg, 2001. Eurostat.
10. J. Domingo-Ferrer, F. Sebé, and A. Solanas. A polynomial-time approximation to optimal multivariate microaggregation. *Manuscript*, 2005.
11. J. Domingo-Ferrer and V. Torra. Ordinal, continuous and heterogenerous k-anonymity through microaggregation. *Data Mining and Knowledge Discovery*, 11(2):195–212, 2005.
12. A. W. F. Edwards and L. L. Cavalli-Sforza. A method for cluster analysis. *Biometrics*, 21:362–375, 1965.
13. C. Fremuth-Paeger. Goblin: A library for graph matching and network programming problems. release 2.7, 2005.
http://www.math.uni-augsburg.de/opt/goblin.html.
14. C. Fremuth-Paeger and D. Jungnickel. Balanced network flows. vii: Primal-dual algorithms. *Networks*, 39(1):35–42, 2002.
15. A. D. Gordon and J. T. Henderson. An algorithm for euclidean sum of squares classification. *Biometrics*, 33:355–362, 1977.
16. P. Hansen, B. Jaumard, and N. Mladenovic. Minimum sum of squares clustering in a low dimensional space. *Journal of Classification*, 15:37–55, 1998.
17. A. Hundepool, A. Van de Wetering, R. Ramaswamy, L. Franconi, A. Capobianchi, P.-P. DeWolf, J. Domingo-Ferrer, V. Torra, R. Brand, and S. Giessing. *μ-ARGUS version 4.0 Software and User's Manual.* Statistics Netherlands, Voorburg NL, may 2005. http://neon.vb.cbs.nl/casc.

18. M. Laszlo and S. Mukherjee. Minimum spanning tree partitioning algorithm for microaggregation. *IEEE Transactions on Knowledge and Data Engineering*, 17(7):902–911, 2005.
19. A. Oganian and J. Domingo-Ferrer. On the complexity of optimal microaggregation for statistical disclosure control. *Statistical Journal of the United Nations Economic Comission for Europe*, 18(4):345–354, 2001.
20. P. Samarati. Protecting respondents' identities in microdata release. *IEEE Transactions on Knowledge and Data Engineering*, 13(6):1010–1027, 2001.
21. P. Samarati and L. Sweeney. Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression. Technical report, SRI International, 1998.
22. G. Sande. Exact and approximate methods for data directed microaggregation in one or more dimensions. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(5):459–476, 2002.
23. A. Schrijver, editor. *Combinatorial Optimization: Polyhedra and Efficiency. Volume A*. Springer Verlag, Berlin, 2003.
24. L. Sweeney. Achieving k-anonymity privacy protection using generalization and suppression. *International Journal of Uncertainty, Fuzziness and Knowledge Based Systems*, 10(5):571–588, 2002.
25. L. Sweeney. k-anonimity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge Based Systems*, 10(5):557–570, 2002.
26. V. Torra. Microaggregation for categorical variables: a median based approach. In J. Domingo-Ferrer and V. Torra, editors, *Privacy in Statistical Databases*, volume 3050 of *Lecture Notes in Computer Science*, pages 162–174, Berlin Heidelberg, 2004. Springer.
27. J. H. Ward. Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58:236–244, 1963.
28. W. E. Yancey, W. E. Winkler, and R. H. Creecy. Disclosure risk assessment in perturbative microdata protection. In J. Domingo-Ferrer, editor, *Inference Control in Statistical Databases*, volume 2316 of *Lecture Notes in Computer Science*, pages 135–152, Berlin Heidelberg, 2002. Springer.

Using the Jackknife Method to Produce Safe Plots of Microdata

Jobst Heitzig

Federal Statistical Office Germany, Gustav-Stresemann-Ring 11, 65189 Wiesbaden, Germany
jobst.heitzig@destatis.de
http://www.destatis.de

Abstract. We discuss several methods for producing plots of uni- and bivariate distributions of confidential numeric microdata so that no single value is disclosed even in the presence of detailed additional knowledge, using the jackknife method of confidentiality protection. For histograms (as for frequency tables) this is similar to adding white noise of constant amplitude to all frequencies. Decreasing the bin size and smoothing, leading to kernel density estimation in the limit, gives more informative plots which need less noise for protection. Detail can be increased by choosing the bandwidth locally. Smoothing also the noise (i.e. using correlated noise) gives more visual improvement. Additional protection comes from robustifying the kernel density estimator or plotting only classified densities as in contour plots.

Keywords: jackknife method, histogram, kernel density estimation, robustness, contour plot, noise, remote access.

1 Introduction

Although plots of microdata are doubtlessly a very useful tool both in presentations and statistical analysis, they pose special problems when the data is confidential. This paper proposes to follow the jackknife methodology introduced in [3] to find ways to produce high-quality plots of data which can guarantee confidentiality. After recalling the jackknife method briefly, we first examine histograms before moving on to density plots utilizing kernel density estimation.

2 The Jackknife Method of Confidentiality Protection

Assume that we want to publish the result $f(M)$ of some statistical analysis (e.g., a mean, a model parameter estimate, the p-value of some test statistic, etc.) computed from confidential microdata which are contained in a matrix $M = (x_{ij})_{ij}$, and that we want to make sure that not even a person who knows all of M except a single individual value x_{ij} can infer anything useful about the "target value" x_{ij}. The idea underlying the jackknife method is to achieve

J. Domingo-Ferrer and L. Franconi (Eds.): PSD 2006, LNCS 4302, pp. 139–151, 2006.

this by reducing the precision of the published result just as much as needed, for example by publishing a small interval instead of the precise value, or by applying some kind of noise to the result.

In [3], it was shown that the necessary amount of imprecision is proportional to the maximal influence which the removal of a single value x_{ij} in M or the replacement of one such value with a randomly chosen value can have on $f(M)$. This analogy to the jackknife method of variance estimation motivates the naming of the protection method. The maximal influence can either be determined by actually performing the removals and replacements or by analysing the *influence function* of f known from the theory of robust estimation. It was also shown in [3] that, in general, the resulting relative imprecision is of order $O(1/N)$ and is thus asymptotically smaller than the relative standard error of order $O(1/\sqrt{N})$, where N is the number of observations in M. For an example of highly skewed, real-world data, Scheffler [5] showed that the quality of various kinds of numerical results protected with this method can well compete with that of results produced from traditionally anonymized microdata files, which however provide a lower level of protection.

For single numeric results f such as sample means, correlation coefficients, model parameter estimates or p-values, the preferred form of the published result is an interval containing the precise value with some high probability or even with certainty or, equivalently, an approximate result together with a certain or confidence bound for its deviation from the precise value. Gopal et. al. [1] discuss a slightly more complex way to produce safe results of this kind.

Graphical representations of distributions (the main interest of this paper) can be interpreted as vectors or matrices of real numbers in the obvious way: a histogram as a vector of frequencies, a density line plot as a vector of density values, a greyscale-coded 2-dimensional density plot as a matrix of brightness values. To produce a safe plot, the jackknife method can be applied to these individual numbers, giving a vector or matrix of approximate frequencies, densities or brightnesses.

Since the usefulness of the resulting plot does not so much depend on any individual deviation between precise and approximate values but more on their distribution, it will in general suffice to publish an aggregate measure of precision together with the approximate plot instead of a vector or matrix of intervals. This is even more so when the aim of the plot is not a quantitative but a qualitative one, e.g. to determine modality, skewness, groups of outliers or shape of correlation, or to find useful transformations. Still, some kinds of plots allow us to incorporate more detailed information about the precision of the plot into the plot itself, e.g. in the form of bands in a density line plot.

3 Histograms

For a *frequency table* of a numeric variable X with fixed class boundaries $b_0 < b_1 < \cdots < b_k$, the jackknife method first recognizes that the frequencies f_c of any cell c can change by at most one when an individual value x_i is removed, added or changed. Consequently, no individual value can be disclosed when from the published result one cannot tell with sufficient certainty whether the actual cell frequency is f_c, $f_c - 1$ or $f_c + 1$. This can be achieved by adding a, say, normally distributed "protection error" to all frequencies. If it is done independently and with equal amplitude for all cells, we have added "white" (that is, uncorrelated) noise. Note that this is quite different from such anonymization methods as perturbation of the underlying microdata, since we add noise only to the analysis results (i.e., the frequencies), thereby achieving much better quality asymptotically.

In a *histogram,* the class boundaries b_c and their number k are usually not fixed but are determined from the data, hence we must additionally make sure that nothing can be disclosed from that information either.

3.1 Choice of Bin Width and Boundaries

Popular choices of a constant bin width $w = b_c - b_{c-1}$ are $w_\mathrm{s} := 3.49sN^{-1/3}$ or $w_\mathrm{q} := 2(q_3 - q_1)N^{-1/3}$, where s, q_1 and q_3 are the sample standard deviation and first and third quartile estimators, respectively (see [4]). The outer boundaries b_0 and b_k are typically chosen just outside the minimum and maximum values $x_{(1)}$ and $x_{(N)}$. It is obvious that the latter cannot be published unmodified and that q_1 and q_3 often refer to individual values, too. Similarly, using w_s is more risky than using w_q, even when no additional knowledge is assumed. From the knowledge of s, N and the sample mean m alone one can infer that all values x_i are between $m \pm \sqrt{N}s$, and extreme values in skewed distributions may be bounded too narrowly by this. For example, a simple simulation shows that in a sample of ten values from a log-normal distribution, the upper bound $m + \sqrt{N}s$ is below $\frac{11}{10}x_{(N)}$ in more than half of all cases.

For these reasons, the outer boundaries and the bin width must also be chosen in a "safe" way, and this can be done by first applying the jackknife method to s or $q_3 - q_1$, and to $x_{(1)}$ and $x_{(N)}$, as described in [3], before using them in the histogram. This kind of "plug-in" mechanism is a typical way to apply the jackknife method to all kinds of complex tasks such as automatic variable selection in regression models or reweighting observations in robust estimation. For such tasks one often uses fairly robust auxiliary estimators which are not much affected by single values, hence the imprecision introduced by the jackknife method at the plug-in stage is quite small. This motivates the use of the robust bin width w_q instead of the considerably non-robust w_s. Even more so, the outer boundaries $x_{(1)}$ and $x_{(N)}$ are extremely non-robust functions of the sample, hence we should instead use either a fixed or user-specified range or a function of some more robust statistics such as $x_{(3)}$ and $x_{(N-2)}$.

Figure 1 shows a histogram of a Normal(0,1)-distributed sample with $N = 100$, $q_1 = -0.71$, $q_3 = 0.61$, $x_{(1)} = -2.43$ and $x_{(N)} = 2.05$ which was protected by adding white noise of amplitude $a = 1$ (i.e., uncorrelated Normal$(0, 1)$-distributed errors) and using fixed outer boundaries of $\approx \pm 3$ and the jackknife-protected bin width $w'_\mathrm{q} = 0.64$ instead of the true $w_\mathrm{q} = 0.57$. In addition, a confidence interval of size $\pm 2a$ and the normal density function are shown.

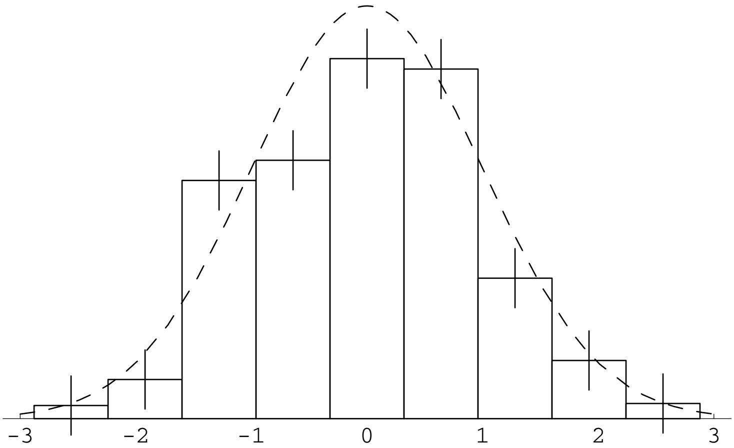


Fig. 1. Histogram of a standard normal sample of size 100, protected by using a jackknife-protected bin-size of 0.64 and adding white noise of amplitude 1, with 95%-confidence intervals and underlying density function

3.2 Choice of Protection Amplitude

One way to assess the level of protection attained with noise of some amplitude a would be to compute something like the worst-case ratio between the likelihood of f_c and the larger of the likelihoods of $f_c - 1$ and $f_c + 1$, given the published $f'_c = f_c + \varepsilon_c$ with $\varepsilon_c \sim \text{Normal}(0, a^2)$ and known a. Such cell-based measures can in fact be determined for all kinds of frequency tables. This worst-case ratio is $e^{1/2a^2}$ and it is attained for $\varepsilon_c = 0$. In other words, even when some individual random error value is actually zero, the frequencies $f_c - 1$ and $f_c + 1$ still have $e^{-1/2a^2}$ times the likelihood of the true frequency f_c. For $a = 1$, this ratio is 0.61.

In our case of histograms, we can alternatively study how accurately a recipient could estimate from $(f'_c)_c$ a single individual value x_i if she knew all other values x_j $(j \neq i)$. The maximum likelihood estimator of x_i given $(f'_c)_c, (x_j)_{j \neq i}$ is the centre of the bin c for which $f'_c - g_c$ is maximal, where the g_c are the bin frequencies of the known values x_j. The solid line in Figure 2 shows the standard error of this estimator for samples of the kind of Figure 1, with $x_i = 0$ and different choices of a.

So far, both considerations show that an amplitude of $a = 1$ can be considered a safe choice. However, Figure 1 indicates that we should also be concerned about the accuracy of the resulting histogram.

3.3 Accuracy of the Resulting Distribution Function

As our published histogram might be used to assess quite different aspects of the underlying distribution, let us evaluate its accuracy by comparing the following distribution functions: the underlying normal distribution function F, the empirical distribution function F_e of the sample, the distribution function

F_{u} that corresponds to the unmodified histogram, and finally that of the published histogram F_{p}, the latter two evaluated at the bin borders and interpolated linearly within. In particular, we examine the ratio between the average squared L_2-distances $||F_{\mathrm{u}} - F||_2^2$ and $||F_{\mathrm{p}} - F||_2^2$, and the ratio between the averages of $||F_{\mathrm{u}} - F_{\mathrm{e}}||_2^2$ and $||F_{\mathrm{p}} - F_{\mathrm{e}}||_2^2$.

The dotted lines in Figure 2 show these quotients for 10,000 simulated samples of the kind of Figure 1, keeping the bin width at $w = 0.64$ but varying a. With $a = 1$, for instance, these "efficiency"-type measures are about 0.86 and 0.69, respectively. That is, seen as an estimate of the underlying distribution, the jackknife-protected histogram is 0.86 times as efficient as the original, unprotected, histogram, while as an estimate of the empirical distribution of the sample, it is 0.69 times as efficient. For $a = \frac{1}{2}$, which still ensures that a single value x_i cannot be estimated more accurately than up to ± 2, these efficiencies are much better (0.96 and 0.90).

4 Plots Using Kernel Density Estimation

A usual way to improve the information given by a histogram is to smooth it by averaging neighboured bin counts while simultaneously increasing the bin number. In the limit, this gives a function f which estimates the density at x as the average of the values $K(\frac{x-x_i}{h_i})/h_i$ over all sample values x_i, where K is some standardized symmetric density function, called the *kernel,* and h_i is some positive *bandwidth* which specifies the amount of smoothing. In the typical case where K is the Gaussian kernel (i.e., the standard normal density function), we thus have

$$f(x) = \frac{1}{N\sqrt{2\pi}} \sum_{i=1}^{N} \frac{1}{h_i} \exp\left(- \left(\frac{x - x_i}{h_i} \right)^2 /2 \right).$$

Most simply, h_i can be chosen equal for all i and determined somewhat similarly to the bin width of a histogram, but shrinking more slowly as N increases, e.g., $h_i = h_{\mathrm{s}} := 1.06 s N^{-1/5}$ (the "simple normal reference rule") or $h_i = h_{\mathrm{t}} := 0.9 \min\{s, \frac{q_3 - q_1}{1.34}\} N^{-1/5}$ (Silverman's rule of thumb, see [6]). The bottom three lines in Figure 4 show such kernel density estimates for the sample $\{-21, -15, -11, -9, -7, -3, 3, 7, 9, 11, 15, 21, 51\}$, with either $h_i = h_{\mathrm{t}} = 8.32$, $h_i = 2h_{\mathrm{t}} = 16.64$ or $h_i = \frac{1}{2} h_{\mathrm{t}} = 4.16$. Only with the smallest choice of constant bandwidth, both the seeming bimodality of the distribution and the outlier are clearly visible. With Silverman's rule of thumb (which is said to "oversmooth" often), the bimodality gets obscured. The outlier, on the other hand, is smoothed out only for the largest bandwidth. Before considering non-constant bandwidths, let us study how safe this is.

4.1 Still, Confidentiality Is an Issue

At first glance, one may think that, if only the bandwidth is large enough, such a density line plot discloses no individual value x_i since no individual peaks are seen. But, unfortunately, at least when the bandwidth is constant and the

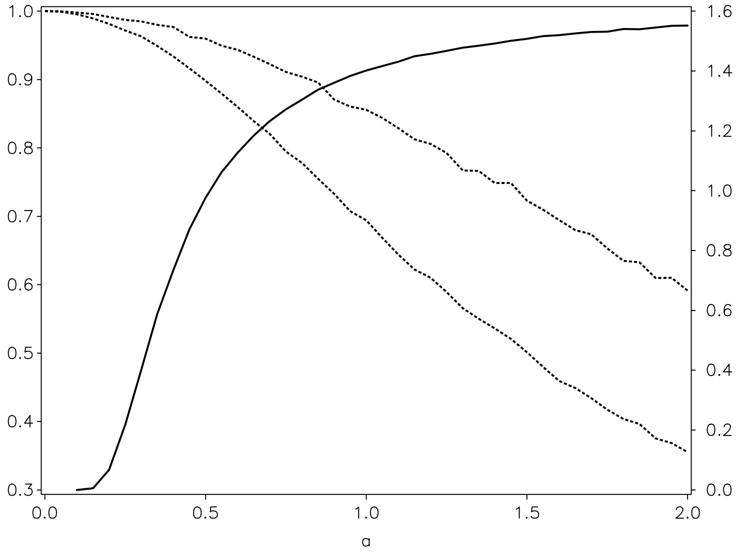


Fig. 2. Standard error (solid line, right scale) and corresponding efficiencies (dotted lines, left scale, higher is better) of the ML-estimator of $x_i = 0$ given a protected *histogram* of the kind of Figure 1, for varying noise amplitude a

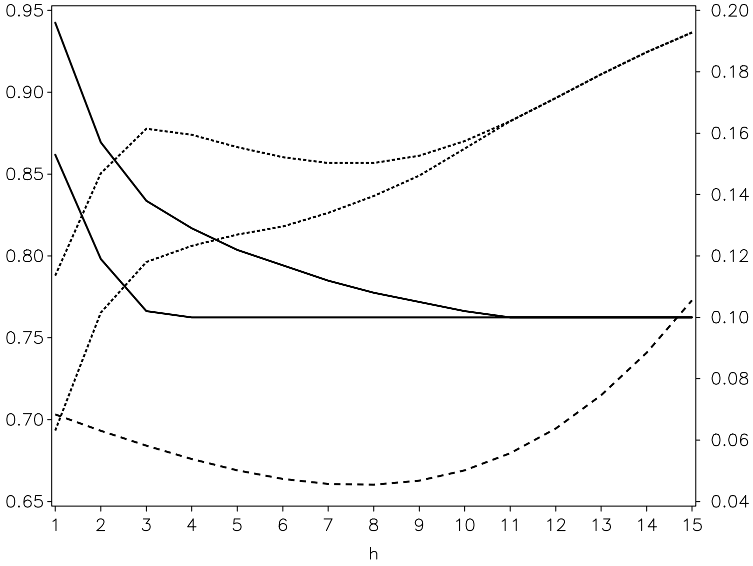


Fig. 3. Required noise amplitude a (solid lines, right scale) and resulting efficiencies (dotted lines, left scale, higher is better) for *kernel density estimation* of 100 normally-distributed values, for varying constant bandwidth h. In addition, the squared L_2-distance of the estimated from the true distribution is shown (dashed line, right scale).

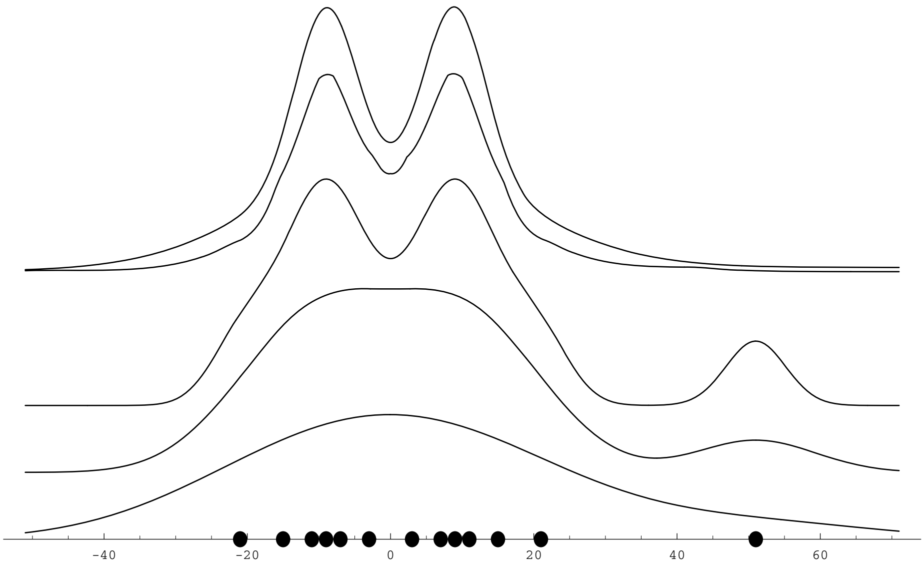


Fig. 4. Different kernel density estimates for the sample shown at the bottom. The bottom three lines have constant bandwidths of $2h_\mathrm{t}$, h_t and $\frac{1}{2}h_\mathrm{t}$, respectively. The top two have local bandwidths proportional to the distance to the nearest neighbour, the lower one being robustified additionally by subtracting the highest influence kernel contributions at each point.

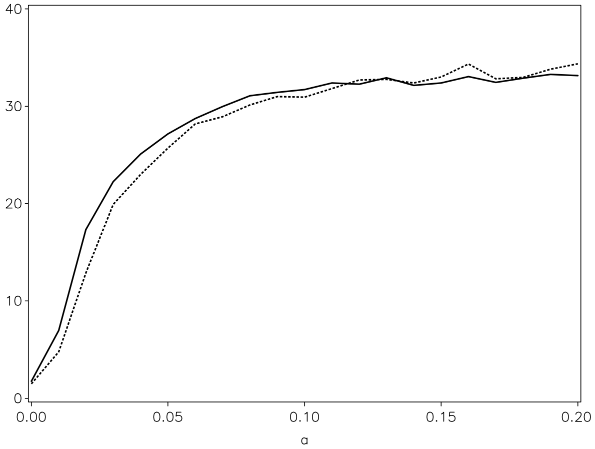


Fig. 5. Comparison of the standard errors of the ML-estimators for the rightmost value $x_{13} = 51$ of the sample from Figure 4, given the non-robustified (dotted line) or robustified (solid line) local-bandwidth density line diagram shown in the top two lines in the other figure, after protection with noise of amplitude a, assuming knowledge of the other values x_j and of the bandwidth $h = 2$

kernel is known, the theory of integral transforms tells us that the transformation $(x_i)_i \mapsto f$ can be inverted easily by using an "inverse" kernel. Applying a Fourier-transform, dividing by the Fourier-transform of the kernel and Fourier-transforming again gives the original values. This means that we have to add noise here, too. However, because of the smoothing, we can hope to need less noise than in the case of a histogram.

In practice, the density plot to be published will only use the values $f(y_c)$ for a number of k equidistant grid points y_c, so that we may treat it in analogy to a histogram with k bins. Doing so, the published density line diagram corresponds to the values

$$f'_c := \frac{1}{N\sqrt{2\pi}} \sum_{i=1}^{N} \frac{1}{h_i} \exp\left(-\left(\frac{y_c - x_i}{h_i}\right)^2 /2\right) + \frac{\varepsilon_c}{N}, \quad \varepsilon_c \sim \text{Normal}(0, a^2)$$

or to their non-negative and normalized version $\max(f'_c, 0)/\sum_c \max(f'_c, 0)$.

As above, let us look at how accurately a recipient could estimate from $(f'_c)_c$ a single individual value x_i if she knew the (possibly constant) bandwidths h_j and all other values x_j $(j \neq i)$. Because of the smoothing, the maximum likelihood estimator of x_i given $(f'_c)_c, (h_j)_j, (x_j)_{j\neq i}$ is now approximately the grid point y_d for which the log-likelihood

$$L(y_d) \propto -\sum_c \left(g'_c - \frac{1}{N\sqrt{2\pi}h_i} \exp\left(-\left(\frac{y_c - y_d}{h_i}\right)^2 /2\right)\right)^2$$

is maximal, where

$$g'_c := f'_c - \frac{1}{N\sqrt{2\pi}} \sum_{j\neq i} \frac{1}{h_j} \exp\left(-\left(\frac{y_c - x_j}{h_j}\right)^2 /2\right).$$

For the same samples as in Figure 2, the solid lines in Figure 3 show what noise amplitude is needed in order that the standard error of this estimator is 1 (lower line) or 1.5 (upper line), depending on the chosen constant bandwidth h. For these amounts of noise, again the accuracy of the resulting density line plots was assessed by computing the ratio (dotted lines) between the averages of $||F_\text{u} - F||_2^2$ and $||F_\text{p} - F||_2^2$, where F_u and F_p are now the distribution functions of the unmodified and protected density estimates f and f'. Here the upper line corresponds to a target standard error of 1, the lower one to a target of 1.5.

As we see, the necessary amplitudes are not zero but generally much smaller (about $\frac{1}{10}$) than those needed with histograms. This is in accordance with the jackknife method's result since when one x_i is changed, added or removed, $f(x)$ can change by at most $1/(N-1)\sqrt{2\pi}h$, so the necessary amplitude should be of that order. At the same time, the efficiency of the resulting distribution functions are comparable to those for histograms. For the choice of h which minimizes $||F_\text{u} - F||_2^2$ (dashed line), $h \approx 8$, the protected plot has again an efficiency of about 0.85.

Instead of adding noise one could draw a synthetic sample of size N from the estimated density and publish a scatterplot of that sample instead of the density line, as is done in some resampling approaches to anonymization (e.g. [2]). That, however, would reduce the quality of the information not only more than adding noise to the results but even more than perturbing all microdata points individually.

4.2 Better Trade-Off with Local Bandwidths

As the discussion of Figure 4 exemplified, it is generally not optimal to use constant bandwidths since either outliers are too clearly visible, or detail in more populated areas of the plot is smoothed out. Obviously and in accordance with the rationale behind the simple normal reference rule and Silverman's rule of thumb, one should increase bandwidth in less populated areas and decrease it in dense regions. This is even more essential when we want to produce bivariate density plots. Moreover, the factor $N^{-1/5}$ leads to a too slow decreasing bandwidth in our case where protection is the main purpose and not density estimation in itself, since it smoothes out too much detail that could safely be shown for larger values of N.

Although the literature (e.g. [8]) contains quite sophisticated adaptive and iterative methods to determine local bandwidths h_i which are in some sense optimal for estimation, let us here, for simplicity's sake, only look at a simple heuristic way to choose bandwidths locally: we put

$$h_i := \max\{\beta \min_{j \neq i} |x_i - x_j|, h_0\},$$

that is, we use a constant multiple of the distance to the nearest value, but bounded below by some minimal bandwidth h_0. This is somewhat similar to the *generalized nearest neighbour estimates* discussed in [6]. The topmost line in Figure 4 shows this for $\beta = 2$ and $h_0 = 2$. The bimodality is clearly visible but the outlier is smoothed out. Because of the latter effect, using local bandwidths increases both quality and safety of the plot. In addition, one should expect that even less noise needs to be added since the ML-estimators for single values (and also the algebraic back-transformation) loose their simple form so that disclosure would be much harder even without noise. However, the simulations in 5.1 will show that still some noise is needed for protection although no visible peak shows the outlier directly.

Analysing this with the jackknife method is somewhat more difficult than for constant bandwidths. How much $f(x)$ can change when adding, removing or changing a single x_i now also depends on how much all the h_j can change in this. For one- or two-dimensional density plots, removing x_i can only increase at most two or five of the remaining h_j, respectively (namely those having x_i as their nearest neighbour). Analogously, adding some new x_i can only decrease at most that many bandwidths. Consequently, an upper bound for the maximal change of $f(x)$ is $(\gamma+1)/(N-1)\sqrt{2\pi}h_0$, where γ is 2 or 5 according to dimension. Hence it is important to choose h_0 not *too* small.

4.3 Visual Improvement by Posterior Smoothing

The lower line in Figure 6 shows a local bandwidth density line plot for real-world data from a typical skewed sample of $N = 439$ magnitudes between 0 and 1 million, using $k = 201$ grid points at multiples of 5000, with a bandwidth factor of $\beta = 10$ and a minimal bandwidth of $h_0 = 50{,}000$, protected by adding white noise of amplitude $a = 3/\sqrt{2\pi}h_0 \approx 0.000024$. Although the shape of the distribution is clearly visible, the upper line in that figure is much more appealing because of its calmer appearance. That plot was produced by simply smoothing the noise-protected graph a second time with a small constant bandwidth. This does not remove the uncorrelated noise that we just added but only converts it into correlated, less "visible" noise. The calmer plot contains exactly the same information as the restive one since the smoothing is a bijective transformation. In particular, the same level of protection is attained by both! In Figure 6, the posterior smoothing was done by simply averaging the values at three neighboured grid-points and repeating this five times.

5 Improving the Protection

5.1 Robust Estimation

Obviously, the less a result depends on single values the smaller the protection amplitude needs to be. In other words, the more robust the estimator, the better the precision we can publish it with. Many point estimators can be "robustified" by either trimming (e.g. for the mean), switching from squared to absolute values (e.g. using Gini's mean absolute difference instead of standard deviation), replacing values by their ranks (e.g. using Spearman instead of Pearson correlation), or by plugging in robust ingredients for non-robust ingredients (e.g. using the median absolute deviation from the median instead of the mean squared deviation from the mean).

Similar things can be done with the kernel density estimator $f(x)$ since it is just the mean of individual kernel contributions. A radical idea would be to replace the mean by the median, but that would reduce the efficiency of the estimator dramatically since the distribution of the individual kernel distributions is highly skewed at each position. A more suitable approach is trimming, that is, we just leave out the largest contribution for instance:

$$f_{\mathrm{r}}(x) := \frac{1}{(N-1)\sqrt{2\pi}} \left(\sum_{i=1}^{N} z_i - \max_{i=1}^{N} z_i \right), \quad z_i = \frac{1}{h_i} \exp\left(- \left(\frac{x - x_i}{h_i} \right)^2 /2 \right).$$

The second line from the top in Figure 4 shows such a robustified plot, while Figure 5 shows how this robustification affects the standard error of the ML-estimator for the rightmost value $x_{13} = 51$ of that same sample, given knowledge of the other values x_j ($j \neq 13$) and of $h = 2$, for varying noise amplitude a. While for small values of a, there is a significant improvement, for larger values of a there does not seem to be one.

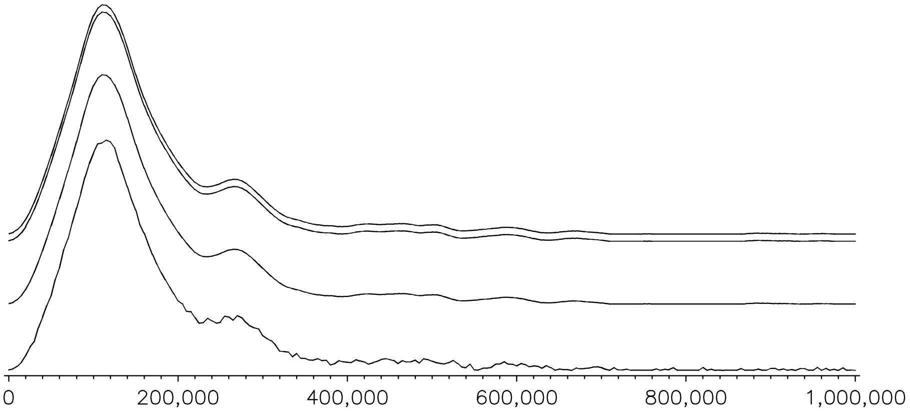


Fig. 6. Example of a local bandwidth density line plot for real-world data with $N = 439$, with a bandwidth factor of $\beta = 10$ and a minimal bandwidth of $h_0 = 50{,}000$, protected by adding white noise of constant amplitude (lower line). The second line is derived from this by posterior smoothing, the upper graph shows the corresponding 95%-confidentiality bands.

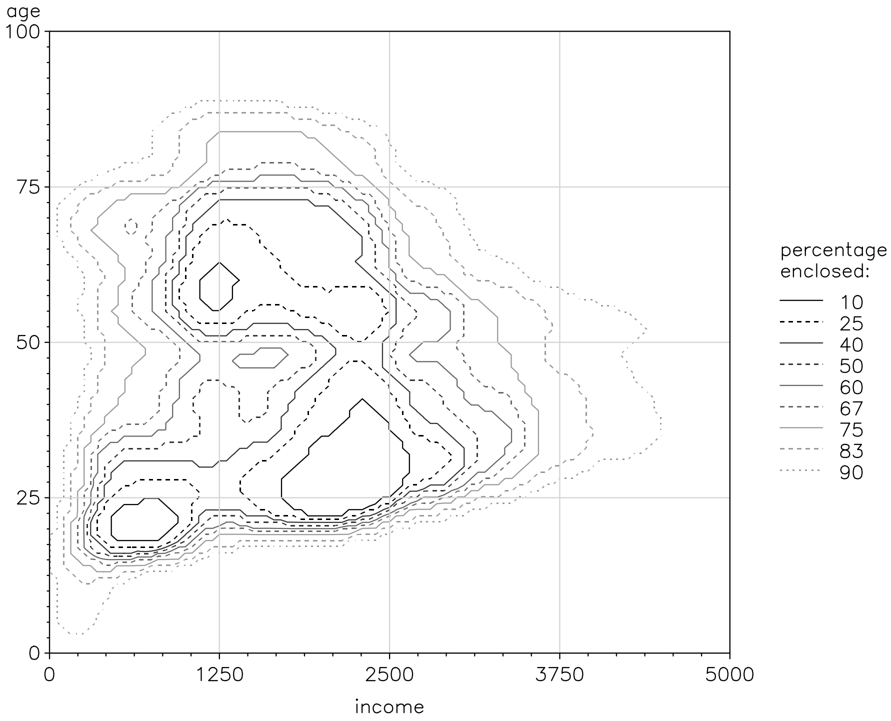


Fig. 7. Example of a 101×101-grid contour plot of a robustified, local-bandwidth, bivariate kernel density estimate based on a large real-world sample, protected by adding white noise of constant amplitude with posterior smoothing, and densities classified into 10 groups. Bandwidth factor was 8, minimal bandwidths were 50 (income) and 1 (age).

For larger samples, one could even afford to trim the largest $\gamma+1$ (see above) contributions z_j, so that the upper bound on how much $f_\mathrm{r}(x)$ can change by adding, removing or changing a single x_i becomes smaller in less populated areas: since at most $\gamma+1$ of the z_j are affected by such a modification, $f_\mathrm{r}(x)$ can change by at most the sum of the largest $\gamma+1$ values among the z_j, which is noticeably less than $(\gamma+1)/(N-1)\sqrt{2\pi}h_0$ in sparsely populated areas. This would imply that one also needs less noise there. However, one has to be careful when using a noise amplitude depending too much on local data since the amplitude can quite accurately be estimated from the published plot by studying the variance in a small region, and then the trimmed values could also be estimated too accurately from this. In a separate study, one should therefore examine the possibility of using for such higher-trimmed kernel density estimates a local noise amplitude determined by *smoothing* the sum of the trimmed values over a sufficiently large region, so that the resulting amplitude cannot reveal local detail.

5.2 Classification of Densities: Contour Plots

Often, specifically in the bivariate case, one does not plot the actual estimated density but rather draws only a number of meaningful contour lines, thereby classifying the density values into groups. The choice of contour levels can for instance be such that the enclosed density mass represents specific proportions of the whole population, as in the real-world example in Figure 7. In order that the contours be sufficiently continuous and scattered noise-induced "islands" be avoided, posterior smoothing is again essential, even more than for non-classified densities.

Since classification reduces the information profoundly while preserving most of the interesting features of the distribution, it will obviously also decrease the risk of disclosure. From a contour plot of the type of Figure 7 one cannot as easily proceed with a maximum-likelihood estimation of a single value as we did in the case of histograms and univariate density line plots. This should help reducing the level of noise further.

6 Conclusion

As we have seen, addition of noise not to the underlying microdata but to results (frequencies, densities), as suggested by the jackknife method's approach to confidentiality protection, allows us to produce various kinds of high-quality but safe plots of microdata. Because in principle this requires no manual intervention, such plots could not only be valuable in publications but could also be used in statistical databases or other kinds of remote access facilities.

Further research should address in more detail the exact amount of necessary noise and the usefulness of the resulting plots for different tasks such as assessment of modality or skewness, identification of clusters or groups of outliers, classification of shapes of correlation, finding useful transformations of the data, deciding on homoscedasticity, et cetera. It will also be interesting to evaluate how these are affected by the described possibilities to robusitify or classify the densities, or by the use of adaptive ways to choose bandwidths locally.

References

1. Gopal, R., Garfinkel, R., Goes, P.: Confidentiality Via Camouflage: The CVC Approach to Disclosure Limitation When Answering Queries to Databases. Operations Research **50** (2002) 501–516
2. Gottschalk, S.: Microdata disclosure by resampling – empirical findings for business survey data. J. German Statist. Soc. (3) **88** (2004) 279–302
3. Heitzig, J.: The "Jackknife" Method: Confidentiality Protection for Complex Statistical Analyses. Invited paper, Joint UNECE/Eurostat work session on statistical data confidentiality (Geneva, Switzerland, 9–11 November 2005). URL: `http://www.unece.org/stats/documents/ece/ces/ge.46/2005/wp.39.e.pdf`
4. Izenman, A.: Recent Developments in Nonparametric Density Estimation. J. Amer. Statist. Ass. **86** (1991) 205
5. Scheffler, M.: Jackknife-Geheimhaltung im Vergleich zur Nutzung von Scientific-Use-Files am Beispiel der Kostenstrukturerhebung. Methoden–Verfahren–Entwicklungen, Vol. I/2006, Federal Statistical Office Germany, Wiesbaden (2006)
6. Silverman, B.W.: Density Estimation for Statistics and Data Analysis. Monographs on Statistics and Applied Probability, Chapman and Hall, London (1986).
7. Steel, Ph., Reznek, A.: Issues in Designing a Confidentiality Preserving Model Server. Invited paper, Joint UNECE/Eurostat work session on statistical data confidentiality (Geneva, Switzerland, 9–11 November 2005). URL: `http://www.unece.org/stats/documents/ece/ces/ge.46/2005/wp.4.e.pdf`
8. Wand, M.P., Jones, M.C.: Comparison of Smoothing Parameterizations in Bivariate Kernel Density Estimation. J. Amer. Statist. Ass. **88** (1993) 520–528

Combining Blanking and Noise Addition as a Data Disclosure Limitation Method

Anton Flossmann and Sandra Lechner

Department of Economics, Box D-124, University of Konstanz, Center
for Quantitative Methods and Survey Research, 78457 Konstanz, Germany
Phone: ++49-7531-88-3214; Fax: -4450
sandra.lechner@uni-konstanz.de

Abstract. Statistical disclosure limitation is widely used by data collecting institutions to provide safe individual data. In this paper, we propose to combine two separate disclosure limitation techniques blanking and addition of independent noise in order to protect the original data. The proposed approach yields a decrease in the probability of reidentifying/disclosing the individual information, and can be applied to linear as well as nonlinear regression models.

We show how to combine the blanking method and the measurement error method, and how to estimate the model by the combination of the Simulation-Extrapolation (SIMEX) approach proposed by [4] and the Inverse Probability Weighting (IPW) approach going back to [8]. We produce Monte-Carlo evidence on how the reduction of data quality can be minimized by this masking procedure.

Keywords: disclosure limitation technique, error-in-variables, blanking, SIMEX, IPW.

1 Introduction

During the last twenty years, we are witnessing an increasing demand for microdata carried away by the advances in computer technology and the development of econometric softwares. In this time, data collecting institutions have faced a double difficulty because they have not only to find a way to provide a maximum amount of information to the data user, but also to guaranty confidentiality and privacy to the respondents in this case firms or households. Therefore, data collecting institutions become interested in the provision of scientific-use-files that optimally combine both these interests.

Statistical offices can apply various disclosure limitation techniques,[1] such as noise addition, blanking, local suppression, imputation, data swapping, microaggregation, ... in order to protect the confidentiality of the data.[2] However, each

[1] We relate the reader back to [12] [13] and [1] who provide a description of several statistical disclosure limitation techniques.

[2] See, for example [5] and [10] for the effects of some disclosure limitation techniques on the estimation properties.

J. Domingo-Ferrer and L. Franconi (Eds.): PSD 2006, LNCS 4302, pp. 152–163, 2006.

anonymization technique has its protection limits, such that the probability of reidentifying/disclosing the individual information for some observations is not minimized. For example, additive measurement errors modify only slightly the original value, especially when the original value is high, and data blanking only protects specific observations.

This paper is concerned with the combination of two disclosure limitation techniques: blanking and addition of independent noise. To our knowledge, the combination of both approaches has not been analyzed before. The idea behind the proposed method is that observations with high original values, which are not optimally protected by noise addition, are finally protected by data blanking. Therefore, our proposed approach has the advantage of guaranteeing more disclosure protection than the application of both methods separately. However, from the perspective of the researcher, an appropriate estimation method is needed in order to get consistent parameter estimates for linear or nonlinear regression models. Therefore, the basis of combining both disclosure limitation methods lies in the combination of two estimation methods. On the one hand, we apply the Simulation-Extrapolation (SIMEX) method developed by [4] which is well suited for estimating and reducing the bias due to additive measurement error. On the other hand, the Inverse Probability Weighting (IPW) estimator due to [8] is applied to account for the blanking process.

The outline of the paper is as follows. Section 2 introduces both disclosure limitation techniques, blanking and addition of independent noise, separately, and shows how to combine them in order to get consistent estimates of the parameters of interest. Section 3 presents evidence on the power of this approach. Based on the results of a Monte-Carlo experiment, we show that the SIMEX method combined with the IPW approach nicely corrects for the estimation bias introduced by data masking through noise addition and blanking. Finally, Section 4 summarizes the main results and addresses further research questions.

2 The Model

In the following we propose an approach that can be applied to data that are masked by blanking and additive measurement error. First we present a blanking method. Second we briefly explain masking by additive noise in order to finally present a proposed combination of both methods.

2.1 Blanking as a Data Disclosure Limitation Method

Identification Problems Caused by Blanking. Blanking constitutes a data disclosure limitation technique where observations with sensitive information are blanked out of the sample. This, however, creates severe identification problems for the researcher. For expositional purposes, and in order to show the identification problems in a very general form, we rely on the M-estimation setup. The interest lies in estimating parameters of the conditional expectation function, $\mathrm{E}[Y_i|X_i] = \mu(X_i, \theta_0)$, where θ_0 is the true $k \times 1$ parameter vector. In the case of the linear regression model $\mu(X_i, \theta_0) = X_i'\beta_0$ with $\theta_0 = \beta_0$. Let $Z_i = (Y_i, X_i')'$ and $q(Z_i, \theta)$ be an objective function, where $\theta \in \Theta$. For the linear least squares case $q(Z_i, \theta) = (Y_i - X_i'\beta)^2$, where $\theta = \beta$. It is now assumed that θ_0 uniquely solves the following minimization problem:

$$\min_{\theta \in \Theta} \mathrm{E}[q(Z_i, \theta)] \tag{1}$$

Based on this assumption, the M-estimator $\hat{\theta}$ of θ_0 is defined as the solution to the problem

$$\min_{\theta \in \Theta} n^{-1} \sum_{i=1}^{n} q(Z_i, \theta) \tag{2}$$

Under specific regularity conditions, the M-estimator can be shown to be consistent and asymptotically normally distributed (see for example [14], chapter 12). The sample average of $q(Z_i, \theta)$ is an estimator for the population objective function, $\mathrm{E}[q(Z_i, \theta)]$.

Now, if the data set contains blanked values for some variables, $\mathrm{E}[q(Z_i, \theta)]$ is not identified without additional assumptions. To formalize the blanking process, let D_i be a dummy variable taking the value 1 if the realization of Z_i is not blanked and 0 otherwise. The M-estimator based on a complete case analysis solves the problem

$$\min_{\theta \in \Theta} n^{-1} \sum_{i=1}^{n} D_i q(Z_i, \theta) \tag{3}$$

The corresponding estimator $\hat{\theta}$ is only consistent for θ_0 if θ_0 is the solution to

$$\min_{\theta \in \Theta} \mathrm{E}[D_i q(Z_i, \theta)]. \tag{4}$$

However, without further assumptions, the solutions of (4) and (1) are not the same. One approach to encounter this problem is to assume that the missing data mechanism is ignorable:

$$Z_i \perp D_i | W_i, \tag{5}$$

where $\perp$ stands for independence. W_i is a vector of observable random variables, which determine both the blanking process and Z_i. This assumption is also known as the Missing at Random (MAR) assumption, [11]. Based on the MAR assumption, [15] and [16] analyze the Inverse Probability Weighting (IPW) going back to [8]. This method weights the observed moment function by the inverse of the individual probability of being in the sample given the vector of covariates, $\mathrm{P}(D_i = 1|W_i)$. It can easily be shown that

$$\mathrm{E}\left[\frac{D_i q(Z_i, \theta_0)}{\mathrm{P}(D_i = 1|W_i)}\right] = \mathrm{E}[q(Z_i, \theta_0)]. \tag{6}$$

and the weighted M-estimator is the solution for

$$\min_{\theta \in \Theta} n^{-1} \sum_{i=1}^{n} \frac{D_i q(Z_i, \theta)}{\mathrm{P}(D_i = 1|W_i)} \tag{7}$$

[15] shows consistency and asymptotic normality for the weighted M-estimator if in a first step $\mathrm{P}(D_i = 1|W_i)$ is estimated by a parametric model.

The results show so far that in order to identify and consistently estimate the conditional expectation function under M-estimation, the researcher needs to observe the factors which determine both the blanking process and the outcome variable.

A Blanking Method. The identification issues caused by a blanked data set, discussed in the previous section, lead to the consideration that if the data collecting institution would blank the data on the basis of a stochastic process that fulfills these identifying assumptions, the researcher could apply the proposed methods for estimation of linear and nonlinear models. This means that a way has to be found to create blanked data consistent with the selection on observables assumption, and at the same time to minimize the risk of disclosing certain observations by the researcher. In the following we propose a combination of data masking by noise addition and weighting methods. The model setup follows closely [9].

Let Z_i be the vector of all L variables in the data set for observation i. An observation will not be masked if all variables Z_i lie between the quantiles θ_l and θ_u. An indicator D_i for non-blanking can therefore be defined in the following manner:

$$D_i = \begin{cases} 1, \text{ if } q_{\theta_l}(Z_{1j}, ..., Z_{nj}) < Z_{ij} < q_{\theta_u}(Z_{1j}, ..., Z_{nj}) & \forall j = 1, ..., L \\ 0, \text{ otherwise.} \end{cases}$$

where $q_\theta(.)$ is the θ-quantile of the variables Z_j with $\theta_l < \theta_u$. Since missing data on the dependent variable cause the selection bias in estimating the model, we distinguish between D_{yi} and D_{xi} where

$$D_{y_i} = \mathbb{1}\{\ q_{\theta_l}(Y_1, ..., Y_n) < Y_i < q_{\theta_u}(Y_1, ..., Y_n)\} \tag{8}$$

$$D_{x_i} = \mathbb{1}\{q_{\theta_l}(X_{1j}, ..., X_{nj}) < X_{ij} < q_{\theta_u}(X_{1j}, ..., X_{nj})\}\ \ \forall j = 1, ..., L_x \tag{9}$$

Let v_i and $\omega_{ij},\ j = 1, ..., L_x$, be zero mean iid variables. Then, define $Y_i^* = Y_i + v_i$ and $X_{ij}^* = X_{ij} + \omega_{ij},\ j = 1, ..., L_x$. Hence, the variables Y_i^* and X_{ij}^* are created by noise addition on the original values. Define the indicator function to be

$$D_{y_i}^* = \mathbb{1}\{q_{\theta_l}(Y_1, ..., Y_n) < Y_i^* < q_{\theta_u}(Y_1, ..., Y_n)\} \tag{10}$$

$$D_{x_i}^* = \mathbb{1}\{q_{\theta_l}(X_{1j}, ..., X_{nj}) < X_{ij}^* < q_{\theta_u}(X_{1j}, ..., X_{nj})\}\ \ \forall j = 1, ..., L_x \tag{11}$$

Therefore, the conditional probability of observing a non blanked observation, given Y_i is:

$$\mathrm{P}(D_{y_i}^* = 1|Y_i) = \mathrm{P}(q_{\theta_l}(Y_1, ..., Y_n) - Y_i < v_i < q_{\theta_u}(Y_1, ..., Y_n) - Y_i). \tag{12}$$

Let $P(Y_i) \equiv \mathrm{P}(D_{y_i}^* = 1|Y_i)$. This is a valid weighting probability for the M-estimation problem since

$$\mathrm{E}\left[\frac{D_{x_i}^* D_{y_i}^* q(Z_i, \theta)}{\mathrm{P}(Y_i)}\right] = \mathrm{E}\left[\mathrm{E}\left[D_{x_i}^* D_{y_i}^* \frac{q(Z_i, \theta)}{\mathrm{P}(Y_i)}\middle| Y_i\right]\right] = \mathrm{E}[D_{x_i}^* q(Z_i, \theta)] \tag{13}$$

Under the additional assumption that θ_0 solves

$$\min_{\theta \in \Theta} \mathrm{E}[q(Z_i, \theta)|X_i], \tag{14}$$

the solutions of (13) and (1) are the same. Hence, if the researcher would know or dispose of estimates of $P(Y_i)$, he could carry out M-estimation by minimizing the following empirical objective function:

$$n^{-1} \sum_{i=1}^{n} \frac{D_{xi}^{*} D_{yi}^{*} q(Z_i, \theta)}{\hat{\mathrm{P}}(Y_i)}, \tag{15}$$

where $\hat{P}(Y_i)$ is an estimator for $P(Y_i)$. This result gives rise to the following data protection method for the data releasing institution:

1. Create a blanked data set by removing the observations lying outside the critical quantile range following (10) and (11).
2. Compute the corresponding conditional probabilities.
3. Provide the researcher with the blanked data set and the conditional probabilities.

The drawback of this method is that since the error term v_i is introduced in the blanking process, observations lying outside but at the margin of the critical quantile range may be kept in the sample. To what extent this is the case depends on the variance of the simulated errors v_i. A high variance would increase the risk of disclosure since more sensitive observations are not blanked, but on the other hand, this would raise the quality of the estimation results since the amount of randomness in the selection process is larger. Another tradeoff between disclosure risk and estimation quality is faced in the choice of the θ-quantiles. The more observations are blanked, the lower the risk of disclosure, but the lower also the efficiency of the estimators due to a larger loss of observations.

2.2 Noise Addition as Disclosure Limitation Technique

Measurement Error in the Explanatory Variable. A very simple way to protect data is to add some independent noise to the covariates. This leads then to the well known error-in-variables problems, discussed by [7] for the linear regression model. Without loss of generality, suppose that the explanatory variable X_i contains sensitive information, which should be protected against disclosure. Rather than observing X_i, we observe a masked explanatory variable X_i^m defined as:

$$X_i^m = X_i + u_i, \tag{16}$$

where u_i is an independent random variable with $\mathrm{E}[u_i|X_i] = 0$ and $\mathrm{V}[u_i|X_i] = \sigma_u^2$, that is added to the original variable in order to mask it.

It is standard textbook wisdom that in the bivariate linear regression model, the ordinary least squares (OLS) estimate of β is inconsistent because the error term and the regressor are correlated, when measurement error occurs. Using the Weak Law of Large Numbers (WLLN), one can easily show that instead of obtaining an estimate of the parameter of interest, OLS estimates the product of the reliability ratio and the true parameter value consistently, i.e:

$$\text{plim } \hat{\beta}_{naive} = \frac{\sigma_x^2}{\sigma_x^2 + \sigma_u^2}\beta = \kappa_{xx}\beta, \tag{17}$$

where σ_x^2 is the variance of X_i and the *reliability ratio* $\kappa_{xx} \leq 1$. [3]

The advantage of this method is that the data generating process is known in the case of disclosure limitation, so that a correction of the asymptotic bias of the OLS estimator can easily be implemented for the linear regression model. Therefore, if the data collecting institutions provide the data user with the value of the reliability ratio, or at least with the variance of the measurement error, $\text{V}[u_i]$, it is possible to construct an unbiased OLS estimator of the parameter of interest. On the other hand, data collecting institutions complain that this disclosure limitation technique does not protect enough the confidentiality of the data, so that there exist a non negligible probability that a "data attacker" is able to re-identify one unit in the scientific use file. Indeed, additive measurement errors modify only slightly the original value, especially when the original value is high. This means that the probability of reidentifying/disclosing the individual information for those observations is not minimized.

In this paper, we apply the SIMEX approach in order to correct the estimates because this method does not depend on the functional form of the model and is quite general.

SIMEX for Additive Measurement Errors

The SIMEX method is a two step simulation based method of estimating and reducing the bias due to additive measurement error. In the simulation step of the SIMEX algorithm, additional measurement errors are added to the covariate measured with error. We generate B new covariates $X_{i,b}^m(\lambda_t)$ by the rule:

$$X_{i,b}^m(\lambda_t) = X_i^m + \sqrt{\lambda_t}u_{i,b}, \qquad b = 1, \ldots, B,\ t = 1, \ldots, T,\ i = 1, \ldots, n, \tag{18}$$

where $0 = \lambda_0 < \lambda_1 < \lambda_2 < \ldots < \lambda_T = 2$, are given parameters controlling for the variance of the measurement error[4], and $\{u_{i,b}\}_{b=1}^B$ are iid computer simulated normal random numbers with mean zero and variance σ_u^2. This means that for each λ_t the simulation step creates B additional datasets with the same dependent variable Y_i and the explanatory variable $X_{i,b}^m(\lambda_t)$ whose variance

$$\text{V}[X_{i,b}^m(\lambda_t)] = \sigma_x^2 + (1 + \lambda_t)\sigma_u^2. \tag{19}$$

Given the B estimates for each λ_t, we compute an average estimate $\hat{\beta}(\lambda_t) = \frac{1}{B}\sum_{b=1}^B \hat{\beta}_b(\lambda_t)$ of the vector of naive estimates, $\hat{\beta}_b(\lambda_t)$, obtained by regression of Y on $\{X_b^m(\lambda_t)\}$ for each λ_t.

[3] For more explanations about measurement error in linear models, we relate back to [7]. If we consider a multivariate regression model, the reliability ratio is a little bit more complex, see [3].

[4] The value $\lambda_T = 2$ is recommended by [3].

In the extrapolation step, each component of the vector $\hat{\beta}(\lambda_t)$ is modelled as a function of λ_t for $\lambda_t \geq 0$. The SIMEX estimator is finally defined as the extrapolation of $\hat{\beta}(\lambda_t)$ to $\hat{\beta}(\lambda_t = -1)$, which represents the bias free estimate of β.[5] The standard errors of the parameter of interest are estimated with the bootstrap method.[6] [2] derive the asymptotic distribution of the SIMEX estimator for parametric models.

2.3 Blanking and Noise Addition

The idea of combining these data disclosure methods described above is based on compensating the drawbacks of both methods. However, from the researcher's perspective an appropriate estimation method is required in order to analyze such an anonymized data set.

Consider a masked data set by noise addition. If in addition we would blank the data set by the method proposed in (10) and (11), estimation methods have to take account of both the blanking process and the addition of noise. A consistent estimate of the parameter of interest can be obtained by applying the SIMEX procedure to the IPW-estimator. The proposed estimation method would imply the following data disclosure limitation protocol for the data collecting institution:

1. Add independent noise to the sensitive variables.
2. Create a blanked data set from the masked data by removing the observations lying outside the critical quantile range following (10) and (11).
3. Compute the corresponding conditional probabilities.
4. Provide the researcher with the blanked data set, the conditional probabilities, and the variance of the measurement error term u_i.

Then the researcher can combine SIMEX with IPW. Since he disposes of the conditional probability estimates, $\hat{\mathrm{P}}(Y_i)$, and the variance of the measurement error u_i, he can apply the empirical objective function in (15) as objective function in the SIMEX-procedure. The main advantage of our method is an increase in data protection while keeping the flexibility of the methods described above in estimating a variety of model specifications.

3 Monte-Carlo Experiment

The Monte-Carlo experiment illustrates the quantitative effect of our proposed approach, i.e, blanking and measurement errors. In order to get a better understanding on the impact of our proposed method on the properties of the estimates, we focus on the linear regression model.[7] Without loss of generality,

[5] The estimates of all parameters in the model are obtained in the same way.

[6] See [6] for more details.

[7] One can argue that we don't need the SIMEX method as an estimation procedure because we consider a linear regression model and are able to correct directly the estimated parameters. However, our future work will be concerned with analyzing the finite sample properties of the proposed method when estimating nonlinear models. Therefore, a more generalized estimation procedure is needed.

let us consider a multivariate linear regression model given by:

$$Y_i = \alpha + \beta X_{1i} + \gamma X_{2i} + \varepsilon_i, \quad i = 1, \cdots, n, \tag{20}$$

where $\varepsilon_i \sim N(0,1)$. We suppose that the two explanatory variables X_1 and X_2 follow a multivariate normal distribution:

$$\begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0.25 \\ 0.25 & 1 \end{bmatrix} \right) .$$

For each simulation we suppose that the true values for the parameters of interest are $\alpha = 0.5$, $\beta = 1$, and $\gamma = -1$. As masked variable we consider X_{1i}^m, which is created by adding an iid noise u_i drawn from a $N(0, \sigma_u^2)$ distribution to X_{1i}. We base our blanking procedure on Y_i and X_{1i}^m by computing first the θ-quantiles $q_\theta(Y_1, ..., Y_n)$ and $q_\theta(X_{11}^m, ..., X_{1n}^m)$, respectively.[8] The indicator for non-data protection is generated as

$$D_i = \mathbb{1}\{X_{1i}^m + v_{1i} < q_{\theta_u}(X_{11}^m, ..., X_{1n}^m)\} \times \mathbb{1}\{Y_i + v_{2i} < q_{\theta_u}(Y_1, ..., Y_n)\}, \tag{21}$$

where v_{1i} and $v_{2i} \sim N(0, \sigma_v^2)$. Our Monte-Carlo results are based on two different samples of size $n = 100$ and $n = 1000$, which are replicated $R = 1000$ times. For the SIMEX approach, we suppose that the λ_t's are equidistant in the interval $[0, 2]$, so that $0 = \lambda_0 < \lambda_1 = 0.5 < \lambda_2 = 1 < \lambda_3 = 1.5 < \lambda_4 = 2$, and generate for each value of λ, B=50 samples.

We consider 4 Monte-Carlo designs for different specifications of the blanking and noise parameters.

Table 1. Different Monte-Carlo designs

Design	σ_v^2	σ_u^2	q_{θ_u}
1	1.0	0.01	0.95
2	1.0	0.01	0.90
3	1.0	0.5	0.95
4	1.0	0.5	0.90

Tables 2 to 5 contain the results of the Monte-Carlo simulations for all 4 different designs listed in Table 1. $\hat{\alpha}_{true}, \hat{\beta}_{true}$ and $\hat{\gamma}_{true}$ correspond to the estimates obtained with the original dataset. These estimators show how close our estimates come to the estimates of the original data. In order to have a benchmark how strongly our proposed approach bias the estimates, we also report the naive OLS-estimates $\hat{\alpha}_{naive}, \hat{\beta}_{naive}$ and $\hat{\gamma}_{naive}$ on the blanked and mismeasured data. Finally, $\hat{\alpha}_{BSIMEX}, \hat{\beta}_{BSIMEX}$ and $\hat{\gamma}_{BSIMEX}$ represent the estimates of the model where the combined approach is used in order to correct the estimates.

[8] Usually the risk of disclosure is particularly high only for large values, such that blanking of values below θ_l can be neglected.

Table 2. Estimation results for Design 1. This table contains the results of the Monte-Carlo simulations for sample size $n = 100$ in the left part and the results for sample size $n = 1000$ in the right part of the table.

	$n = 100$				$n = 1000$			
	Mean	Bias	RMSE	RELSE	Mean	Bias	RMSE	RELSE
$\hat{\alpha}_{true}$	0.500	0.000	0.101	1.001	0.501	0.001	0.033	1.002
$\hat{\beta}_{true}$	1.000	0.000	0.110	0.958	0.998	-0.000	0.033	0.997
$\hat{\gamma}_{true}$	-0.997	0.003	0.104	1.009	-0.999	0.001	0.032	1.005
$\hat{\alpha}_{naive}$	0.392	-0.108	0.162	0.927	0.404	-0.096	0.103	0.911
$\hat{\beta}_{naive}$	0.910	-0.090	0.159	0.963	0.913	-0.087	0.096	0.967
$\hat{\gamma}_{naive}$	-0.917	0.083	0.148	0.957	-0.920	0.080	0.089	0.960
$\hat{\alpha}_{BSIMEX}$	0.482	-0.018	0.139	0.892	0.500	0.000	0.049	0.873
$\hat{\beta}_{BSIMEX}$	0.982	-0.018	0.155	0.900	0.997	-0.003	0.062	0.803
$\hat{\gamma}_{BSIMEX}$	-0.980	0.020	0.140	0.906	-0.994	0.006	0.054	0.841
% risky obs. dropped	$r_y = 80 \quad r_x = 78$				$r_y = 80 \quad r_x = 74$			
Obs. dropped	19.827				185.193			

Table 3. Estimation results for Design 2. This table contains the results of the Monte-Carlo simulations for sample size $n = 100$ in the left part and the results for sample size $n = 1000$ in the right part of the table.

	$n = 100$				$n = 1000$			
	Mean	Bias	RMSE	RELSE	Mean	Bias	RMSE	RELSE
$\hat{\alpha}_{true}$	0.494	-0.006	0.101	1.002	0.501	0.001	0.032	0.988
$\hat{\beta}_{true}$	0.998	-0.002	0.104	1.007	0.998	-0.002	0.032	1.006
$\hat{\gamma}_{true}$	-1.004	-0.004	0.110	0.959	-0.999	0.001	0.032	1.015
$\hat{\alpha}_{naive}$	0.328	-0.172	0.216	0.931	0.343	-0.157	0.162	0.920
$\hat{\beta}_{naive}$	0.881	-0.119	0.180	1.006	0.884	-0.116	0.123	0.994
$\hat{\gamma}_{naive}$	-0.896	0.104	0.170	0.924	-0.889	0.111	0.117	0.982
$\hat{\alpha}_{BSIMEX}$	0.468	-0.032	0.175	0.815	0.497	-0.003	0.062	0.829
$\hat{\beta}_{BSIMEX}$	0.971	-0.029	0.182	0.858	0.994	-0.006	0.069	0.819
$\hat{\gamma}_{BSIMEX}$	-0.978	0.022	0.167	0.832	-0.990	0.010	0.068	0.774
% risky obs. dropped	$r_y = 82 \quad r_x = 80$				$r_y = 82 \quad r_x = 77$			
Obs. dropped	28.879				277.964			

The relative standard error, RELSE, is defined as the ratio of the average standard error of the estimator over the number of completed MC replications to the empirical standard deviation of the estimator. When the number of replications tends to infinity, the standard error of the estimates converges to the true standard error, for a finite number of observations n. A deviation of RELSE from 1 provides information about the accuracy of the estimation of the standard error based on the asymptotic distribution.

Table 4. Estimation results for Design 3. This table contains the results of the Monte-Carlo simulations for sample size $n = 100$ in the left part and the results for sample size $n = 1000$ in the right part of the table.

	$n = 100$				$n = 1000$			
	Mean	Bias	RMSE	RELSE	Mean	Bias	RMSE	RELSE
$\hat{\alpha}_{true}$	0.497	-0.003	0.093	1.081	0.498	-0.002	0.033	0.969
$\hat{\beta}_{true}$	0.997	-0.003	0.103	1.007	0.998	-0.002	0.032	1.027
$\hat{\gamma}_{true}$	-1.001	-0.001	0.103	1.008	-1.002	-0.002	0.032	1.006
$\hat{\alpha}_{naive}$	0.362	-0.138	0.187	0.993	0.373	-0.127	0.133	0.926
$\hat{\beta}_{naive}$	0.582	-0.418	0.432	1.013	0.586	-0.414	0.415	1.035
$\hat{\gamma}_{naive}$	-0.818	0.182	0.223	0.995	-0.822	0.178	0.182	1.024
$\hat{\alpha}_{BSIMEX}$	0.531	0.031	0.160	0.943	0.543	0.043	0.072	0.887
$\hat{\beta}_{BSIMEX}$	0.906	-0.094	0.220	0.946	0.920	-0.080	0.107	0.887
$\hat{\gamma}_{BSIMEX}$	-0.948	0.052	0.169	0.936	-0.966	0.034	0.072	0.825
% risky obs. dropped	$r_y = 80$ $r_x = 38$				$r_y = 78$ $r_x = 32$			
Obs. dropped	17.970				167.944			

Table 5. Estimation results for Design 4. This table contains the results of the Monte-Carlo simulations for sample size $n = 100$ in the left part and the results for sample size $n = 1000$ in the right part of the table.

	$n = 100$				$n = 1000$			
	Mean	Bias	RMSE	RELSE	Mean	Bias	RMSE	RELSE
$\hat{\alpha}_{true}$	0.505	0.005	0.098	1.031	0.500	0.000	0.031	1.023
$\hat{\beta}_{true}$	1.002	0.002	0.103	1.020	0.999	-0.00	0.033	0.984
$\hat{\gamma}_{true}$	-1.000	0.000	0.106	0.985	-1.000	0.000	0.033	0.983
$\hat{\alpha}_{naive}$	0.301	-0.199	0.247	0.922	0.299	-0.201	0.206	0.952
$\hat{\beta}_{naive}$	0.563	-0.437	0.455	0.986	0.563	-0.437	0.439	0.992
$\hat{\gamma}_{naive}$	-0.787	0.213	0.255	0.980	-0.786	0.214	0.218	0.951
$\hat{\alpha}_{BSIMEX}$	0.549	0.049	0.191	0.910	0.567	0.067	0.108	0.724
$\hat{\beta}_{BSIMEX}$	0.909	-0.091	0.246	0.939	0.931	-0.069	0.115	0.837
$\hat{\gamma}_{BSIMEX}$	-0.937	0.063	0.192	0.906	-0.960	0.040	0.100	0.669
% risky obs. dropped	$r_y = 81$ $r_x = 49$				$r_y = 81$ $r_x = 45$			
Obs. dropped	27.455				261.851			

As one can see from the tables, the bias and the RMSE are considerably reduced for all models compared to the case when applying the naive estimator. As expected, the bias is somewhat larger for the small sample than for $n = 1000$. The same can be said about the RMSE which decreases considerably with a larger sample size. The results also show that a higher variance of the measurement error u_i (Tables 4 and 5) yields less precise and more biased estimates which would be of little use for the data user in order to get information about the true data generating process. Considering the RELSE for small sample sizes the standard errors cannot be estimated with sufficient precision which was to be expected. However, what is surprising, is that the estimates of RELSE for $n = 1000$ become worse. We think that this is due to the low number of bootstrap replications we chose in order to reduce the computational burden.[9] We will increase the number of replications in future work. Overall we can conclude that the quality of the BSIMEX estimates comes close to those for the original data.

In order to analyze the degree of protection we give in addition the percentage of risky observations dropped for Y_i and X_{1i}^m which is denoted as r_y and r_x, respectively. For example, if the data collecting institution wants to blank sensitive observations above the 90%-quantile for a sample of size $n = 100$, the total amount of risky observations is 10. Then the measure gives the percentage of how many of the 10 risky observations are really dropped from the sample. Due to the random component v_i in the blanking process, data below the 90% quantile are also dropped while some risky data above but close to the 90%-quantile are retained in the sample. As can be seen from Table 2-5, we actually blank about 80% of the risky observations in the corresponding quantiles for the dependent variable. The protection due to the blanking process is therefore relatively high. Regarding the masked variable X_{1i}^m, for $\sigma_u^2 = 0.01$ the percentage of risky observations dropped is also relatively high. For the designs with $\sigma_u^2 = 0.5$, the amount of protection is lower. However, the additional protection due to the blanking procedure relatively to the protection due to the masking procedure is still considerable.

4 Conclusion

In this paper we propose a method for data disclosure limitation which combines blanking and masking by noise addition. In order to correct the estimates for the missing data and the presence of measurement errors we apply the SIMEX-method to the IPW-estimator. The user only needs the variance of the measurement error term and the conditional probability that the observation is blanked, in order to apply the proposed estimation method. The method can be applied to the estimation of both linear and nonlinear models. Monte Carlo evidence shows that the proposed method seems to be appropriate to yield estimates sufficiently close to those based on the original data set, even if the estimated variances are a little bit higher. The simulation results also reveal that the proposed data disclosure limitation method behaves well in protecting the sensitive observations and in offering additional protection relatively to the case when applying only blanking or masking by noise addition.

However, there remain further research topics. First, an extension of the Monte Carlo Study to investigate the finite sample properties in estimating nonlinear models is in progress. Second, a trade-off analysis between bias and efficiency for different specifications of the blanking error and the measurement error should be carried out. This would shed light on the relation between disclosure risk and

[9] We used only 50 bootstrap replications.

estimation quality. Finally, it would be worth to investigate if other combination of statistical estimation methods like for example SIMEX and imputation methods yield better results.

Acknowledgment

Financial support by the DFG is gratefully acknowledged. The usual disclaimer applies.

References

1. Brand, R.: Anonymität Von Betriebsdaten. Beiträge zur Arbeitsmarkt- und Berufsforschung, BeitrAB 237, IAB, Nürnberg (2000)
2. Carroll, R.J., Kuechenhoff, H., Lombard, F., Stefanski, L.A.: Asymptotics for the Simex Estimator in Structural Measurement Error Models. *Journal of the American Statistical Association* **91** (1996) 242–250
3. Carroll, R.J., Ruppert, D., Stefanski, L.A.: Measurement Error in Nonlinear Models. Chapman and Hall (1995)
4. Cook, J.R., Stefanski, L.A.: A Simulation Extrapolation Method for Parametric Measurement Error Models. *Journal of the American Statistical Association*, **89** (1994) 1314–1328
5. Domingo-Ferrer, J., Torra, V.: Disclosure Control Methods and Information Loss for Microdata. in *Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies*, Amsterdam: North-Holland,(2002) 93–112
6. Efron, B., Tibshirani, R.J.: An Introduction to the Bootstrap. Chapman and Hall, New York (1993)
7. Fuller, W.A.: Measurement Error Models. Wiley (1987)
8. Horvitz, D., Thompson, D.: A Generalization of Sampling Without Replacement from a Finite Population. *Journal of the American Statistical Association*, **47** (1952) 663–685
9. Lechner, S., Pohlmeier, W.: To Blank or Not to Blank? A Comparison of the Effects of Disclosure Limitation Methods on Nonlinear Regression Estimates. in Domingo-Ferrer J. and Torra V : *Privacy in Statistical Databases, CASC Project Final Conference, PSD 2004, LNCS 3050*, Springer (2004) 187–200
10. Pohlmeier, W., Ronning, G., Wagner, J.: Econometrics of Anonymized Micro Data. *Sonderband der Jahrbücher für Nationalökonomie und Statistik*, **225** (2005)
11. Rubin, D.B.: Inference and Missing Data. *Biometrika*, **63** (1976) 581 – 592
12. Willenborg, L., de Waal, T.: Statistical Disclosure Control in Practice. Springer Verlag, Lecture Notes in Statistics, Berlin **155** (2000)
13. Willenborg, L., de Waal, T.: Elements of Statistical Disclosure Control. Springer Verlag, Lecture Notes in Statistics, **111** (1996)
14. Wooldridge, J.M.: Econometric Analysis of Cross Section and Panel Data. MIT Press, Cambridge, MA (2002a)
15. Wooldridge, J.M.: Inverse Probability Weighted M-Estimators for Sample Selection, Attrition and Stratification. *Portuguese Economic Journal*, **1** (2002b) 117 – 139
16. Wooldridge, J.M.: Inverse Probability Weighted Estimation for General Missing Data Problems. *Working paper*, Department of Economics, Michigan State University (2003)

Why Swap When You Can Shuffle? A Comparison of the Proximity Swap and Data Shuffle for Numeric Data

Krish Muralidhar[1], Rathindra Sarathy[2], and Ramesh Dandekar[3]

[1] University of Kentucky, Lexington, KY, 40506, USA
krishm@uky.edu
[2] Oklahoma State University, Stillwater, OK, 74078, USA
sarathy@okstate.edu
[3] Energy Information Administration, Department of Energy, Washington DC, USA
Ramesh.Dandekar@eia.doe.gov

Abstract. The rank based proximity swap has been suggested as a data masking mechanism for numerical data. Recently, more sophisticated procedures for masking numerical data that are based on the concept of "shuffling" the data have been proposed. In this study, we compare and contrast the performance of the swapping and shuffling procedures. The results indicate that the shuffling procedures perform better than data swapping both in terms of data utility and disclosure risk.

Keywords: Confidentiality, Data masking, Privacy, Shuffling, Swapping.

1 Introduction

The need for protecting numerical data from disclosure has gained considerable importance in recent years. Government agencies which release data have always been interested in this problem. However, with the increase in the ability of organizations to gather, store, analyze, disseminate, and share data, there has also been a growing demand for commercial organizations to secure sensitive data from disclosure. Recent legislation worldwide has made this an important issue for all organizations that gather and store any sensitive information.

A host of techniques are available for protecting numerical data from disclosure. These include rounding or coarsening, perturbation, micro-aggregation, data swapping, and more recently, data shuffling. Muralidhar and Sarathy [13] provide a comprehensive discussion of the different techniques for protecting numerical data. With the exception of swapping and shuffling, most other data masking techniques involve the modification of the original values of the confidential variables. Many users find such modification of values to be objectionable [17] and hence are less likely to use the modified data. By contrast, by swapping or shuffling the original values, these two techniques leave the original data unmodified. Hence, these techniques are more likely to be accepted by users who find "data modification" objectionable. In addition, it is also easier to explain the concept of swapping or shuffling compared to other advanced techniques.

J. Domingo-Ferrer and L. Franconi (Eds.): PSD 2006, LNCS 4302, pp. 164 – 176, 2006.

Of the two techniques, data swapping has a longer history while the shuffling procedure is relatively new. Hence, there is a need to compare the contrast the performance of the two techniques in terms of disclosure risk and data utility. In this study, we perform such a comparison using simulated data. The remainder of this paper is organized as follows. In the next section, we present a brief description of data swapping and data shuffling. In the third section, we present the data utility and disclosure risk measures used in this study. In the fourth section, we describe the results of simulation experiments conducted to evaluate the relative performance of the two techniques. The final section presents the conclusions.

2 A Brief Description of the Techniques

2.1 Data Swapping

Data swapping was originally proposed by Dalenius and Reiss [3] for masking confidential categorical, rather than numerical, variables. Fienberg and McIntyre [6] provide an excellent discussion of the history of data swapping and its relationship to other methods. In this section, we focus on methods for swapping numerical variables. Reiss et al. [14] swapped numerical data using an optimization approach to maintain the first and second order moments. This approach is computationally difficult and its disclosure risk remains to be evaluated.

Moore [10] describes the best known procedure for swapping numerical data based on the data swapping algorithm proposed by Brian Greenberg in an unpublished manuscript. Let X represent the original, confidential numerical variables and Y, the masked variables (swapped or shuffled). The rank-based proximity swap (hereafter referred to simply as data swapping) for numerical variables can be described as follows [10]. Sort the data by the kth confidential variable Xk ϵ X. Let x(i),k represent the observation with rank (i) of the sorted variable Xk. Replace x(i),k with the observation whose rank is (j) (which now becomes y(j),k). Correspondingly, replace the observation with rank (j) (whose value is y(i),k) with x(j),k. Repeat the process for every i and k to result in Yk and repeat the process of every k to result in Y. The choice of ranks (i) and (j) in the swapping process depends on a masking parameter. For a confidential variable "a" (Xk, in our case) Moore [10] uses a masking parameter called the "swapping distance", defined as follows:

Determine a value P(a), with 0 < P(a) < 100. The intent of the procedure is to swap the value of ai with that of aj, so that the percentage difference of the indices, i and j, is less than P(a) of N. That is |i – j| < P(a)*N/100 (Moore 1996, 6).

The larger the value of P(a), the larger the value of |i – k|, and the greater the distance between the swapped values, and vice versa. The biggest advantage of data swapping is that the marginal distributions of the individual confidential variables are identical to those of the original variables. Assuming a uniform distribution, Moore [10] also shows an inverse relationship between swapping distance and data utility, and a direct relationship between swapping distance and disclosure risk, resulting in a trade-off between data utility and disclosure risk.

2.2 Data Shuffling

Unlike data swapping where the values of the confidential attributes are "exchanged" between records i and j, in data shuffling the value of the confidential variable of record i are assigned to that of j, the value of the confidential variable of record j is assigned to k, and so on, in a specific manner so as to maintain certain characteristics of the data. Currently, there are two approaches that can be used to shuffle the data, namely, based on Latin Hypercube Sampling (LHS) [4] and post perturbation reverse mapping approach [11, 13].

The LHS approach proposed by Dandekar et al. [4] uses Latin Hypercube Sampling [9] and the rank correlation refinement [8] to generate a new synthetic dataset that reproduces both the univariate and multivariate structure of the original dataset. The multivariate structure is reproduced in the sense that the rank order correlation of the masked dataset is the same as that of the original dataset. This approach also uses an iterative refinement approach to reduce the difference in the rank order correlations of the original and shuffled data. For a complete description of the procedure and its performance, please see [4]. Dandekar et al. [4] suggest generating the shuffled values using the inverse cumulative distribution function of the confidential variable. However, this procedure can be easily modified so that the original values are directly used in the newly generated dataset.

Muralidhar and Sarathy [11] independently proposed "data shuffling" as a data masking approach. The procedure proposed by Muralidhar and Sarathy [11, 13] involves two steps. The first step involves the generation of perturbed values of the confidential variables from the conditional distribution of the confidential variables, given the non-confidential variables. Once the perturbed values have been generated, the second step reverse maps the rank ordered values perturbed values to the rank ordered original values. For a complete description of the data shuffling procedure and its performance, please see Muralidhar and Sarathy [13].

In the absence of non-confidential variables, the two procedures reduce to two different approaches for generating an independent dataset with the same joint distribution (in the sense of rank order correlation) as the original dataset for the following reason. When there are no non-confidential variables, then the conditional distribution of the confidential variables given the non-confidential variables reduces to the generation of an independent dataset with the same joint distribution as the confidential variables. It is possible that the iterative refinement proposed in LHS could yield better results.

The major difference between the two procedures lies in the manner in which they address non-confidential variables. Data shuffling explicitly accounts for the presence of the non-confidential variables and generates the perturbed values based on the conditional distribution of the confidential variables given the non-confidential variables. Hence, in addition to maintaining relationships among confidential variables, data shuffling also maintains relationships between confidential and non-confidential variables. By contrast, the primary objective of LHS is to generate a new "synthetic" dataset when the entire dataset is to be perturbed and does not explicitly address the issue of non-confidential variables. When such variables are present, they remain unmodified while the confidential variables are masked, leading to attenuation in the relationship among confidential and non-confidential variables in the masked data. In this sense, data shuffling can be considered more general than the LHS procedure.

3 Criteria for Evaluating Performance of Data Shuffling and Data Swapping

In this section, we describe the specific measures used to evaluate the performance of data shuffling and data swapping. Consistent with prior studies, we use two major criteria, namely, data utility and disclosure risk in evaluating performance.

3.1 Data Utility

Data utility is the extent to which the results from analyzing the masked data are similar to the results from analyzing the original data. A variety of measures have been proposed for this purpose. However, both shuffling and swapping possess one important characteristic, namely, that the univariate distribution of the variables remains exactly the same.

Given that the univariate characteristics remain unchanged, it is natural then to consider the extent to which shuffling and swapping maintain multivariate characteristics. Both procedures rely on rank order correlation for masking the data. Pair-wise rank order correlation is a better measure of the relationship among variables since, unlike product moment correlation that captures only linear relationships, rank order correlation has the ability capture all monotonic relationships. In situations where it can be assumed that the data has a multivariate normal distribution, it is adequate to consider only product moment correlation since all relationships in such a case will be linear. However, in the general case multivariate normality cannot be assumed, it is preferable to use rank order correlation to measure relationships among variables. In order to provide the high data utility, the rank order correlation of the masked and original data should be the same. Hence, our first measure of data utility will be the extent to which the shuffling and swapping maintain rank order correlation. In practice, many users of the data will use product moment correlation in addition to (or in place of) rank order correlation, even if the dataset is not normal. For these users, it is important that the product moment correlation matrix for the masked data be very similar to that of the original data. Hence, it is necessary to evaluate the extent to which the product moment correlation of the masked data is similar to that of the original data.

Thus, a comparison of the utility of shuffling and swapping will be based on the comparison of the difference resulting from using the rank order or product moment correlation computed from the masked data, in place of the same matrices computed from the original data. Note that similar measures have been suggested and used in prior studies [5, 13].

3.2 Disclosure Risk

Disclosure risk is assessed using identity disclosure and value disclosure. Identity disclosure is the ability of an intruder to identify a particular released record as belonging to a particular individual. Value disclosure is the ability of an intruder to be able to estimate the value of a confidential attribute using the released data. Consistent with the prior literature, we will use a broad definition of disclosure to include "partial" disclosure where an intruder may not be able to identify an individual with certainty or the exact value of a confidential variable, but is able to identify an individual or estimate a value with greater certainty using the released data.

In assessing disclosure risk, we use the approach found in Fuller [7]. The intruder, using information on certain original variables, along with the masked data, attempts to first re-identify a masked record as belonging to a particular individual. Subsequently, the intruder attempts to estimate true confidential values. We also assume that the intruder possess accurate information on the aggregate characteristics of the entire, original dataset. These assumptions imply that an intruder is able to develop a prediction equation of the confidential variables, using the masked variables. The proportion of variability in them masked variables by the original variables, using this prediction equation, provides a measure of the risk of value disclosure. Note that this measure represents a lower bound on the risk of value disclosure since an intruder may use more sophisticated approaches to result in even greater level of disclosure [12].

Ideally, a masking procedure should result in masked variables that provide little or no information about the original variables, while maintaining the same relationship among the masked variables to be the same as that found in the original, confidential variables.

4 Description of the Simulation Experiments

4.1 Experiment 1

We conducted two different simulation experiments to compare the effectiveness of swapping and shuffling. The first simulation experiment consisted of generating a dataset of size n (= 30, 100, and 1000) generated from a bivariate normal distribution with a specified product moment correlation ρ (= 0.05, 0.25, 0.50, 0.75, 0.95). For both variables, the mean and variance were specified as 0 and 1, respectively. We masked the data using the shuffling procedure and three specifications of the proximity parameter (10%, 50%, and 100%) for the swapping procedure. The product moment and rank order correlation between the two variables for the original data, the shuffled data, and the three swapped datasets were computed. The difference in the respective correlation between the original and the masked data was computed and recorded. The process was then repeated 1000 times. The average difference and the variance of the difference from the 1000 replications were computed. The entire simulation was then repeated for all sample sizes and specified population correlation combination.

The above dataset was also used to assess value disclosure. The correlation between each of the original variable and the corresponding masked variables was computed and recorded. The proportion of variability explained was then computed as the square of the correlation coefficient. The average proportion of variability explained for each variable was computed as the average of the 1000 replications.

4.2 Experiment 2

The objective of the second simulation experiment was to assess the risk of identity disclosure. In this experiment, we generated a dataset of size n consisting of k variables from a multivariate normal distribution. The mean vector of the dataset was specified as 0 and the covariance matrix was specified as the identity matrix. Four sets of masked data (one shuffled and three swapped) were generated for this dataset. Record linkage was performed on the datasets using the procedure suggested by Fuller [7]. The number of records that were re-identified was recorded. The entire process was repeated 1000 times. The average percentage of records re-identified was computed. The experiment was conducted for four values of n (= 30, 100, 1000) and five values of k (= 2, 3, 4, 5, 6).

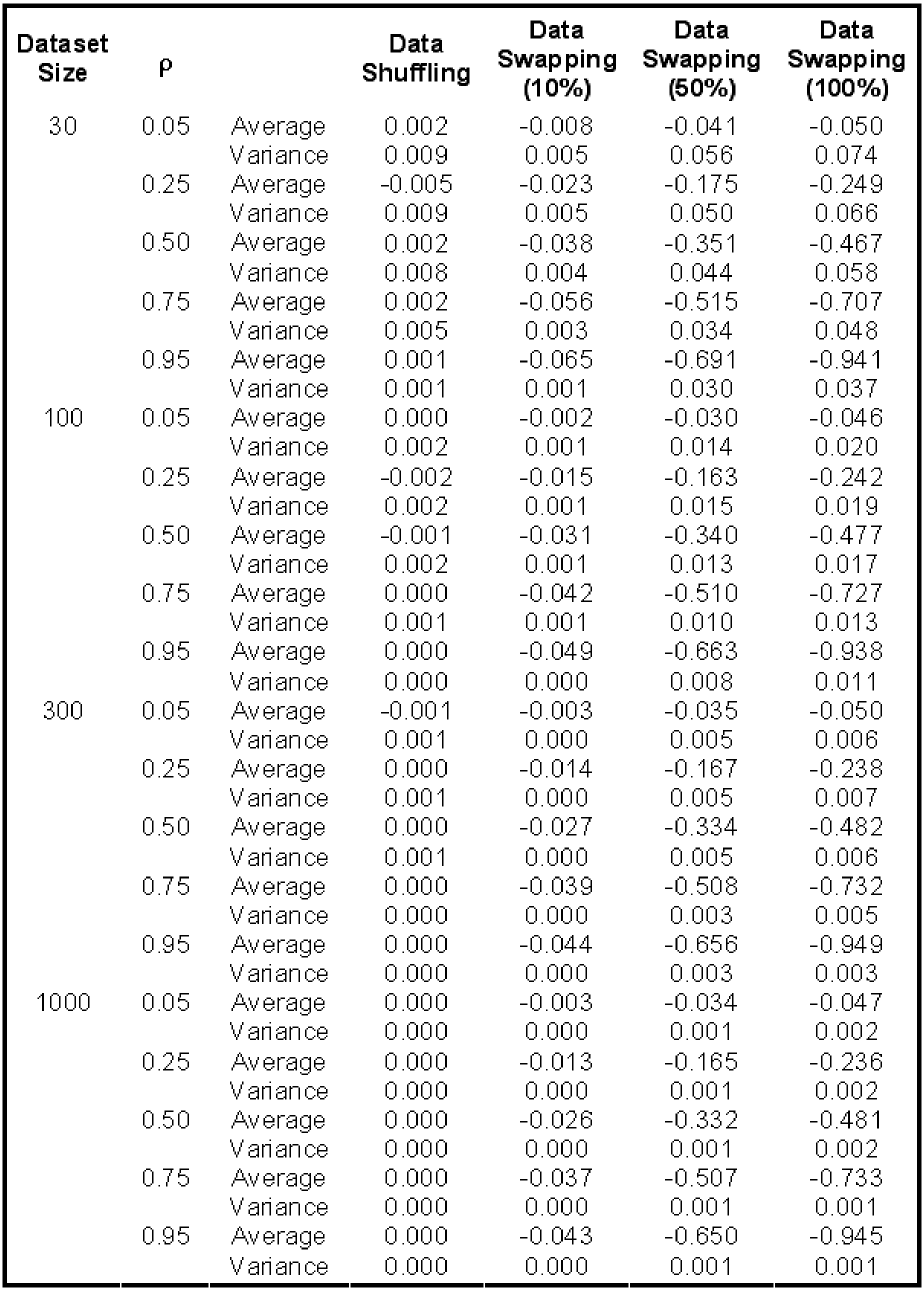
5 Discussion of the Results

5.1 Data Utility

Tables 1 and 2 summarize the results of the simulation experiment conducted to evaluate the relative performance of shuffling and swapping. Table 1 provides the results of the simulation experiment for each n and ρ combination, the average and variance in the difference in the rank order correlation between the masked and original variables for shuffling and the three swapping procedures. For data shuffling, even for a very small dataset (n = 30), the absolute average difference in the rank order correlation between the masked and original data is always less then 0.006. When n = 100, the largest absolute average difference is even small (less than 0.002). For n = 300 and 1000, the average difference is practically negligible. The variance of the difference is also small (no more than 0.008 for any n and ρ combination). These results indicate that, if the shuffled data is used in place of the original data, the rank order correlation obtained from the masked data is likely to be very close to that of the original data.

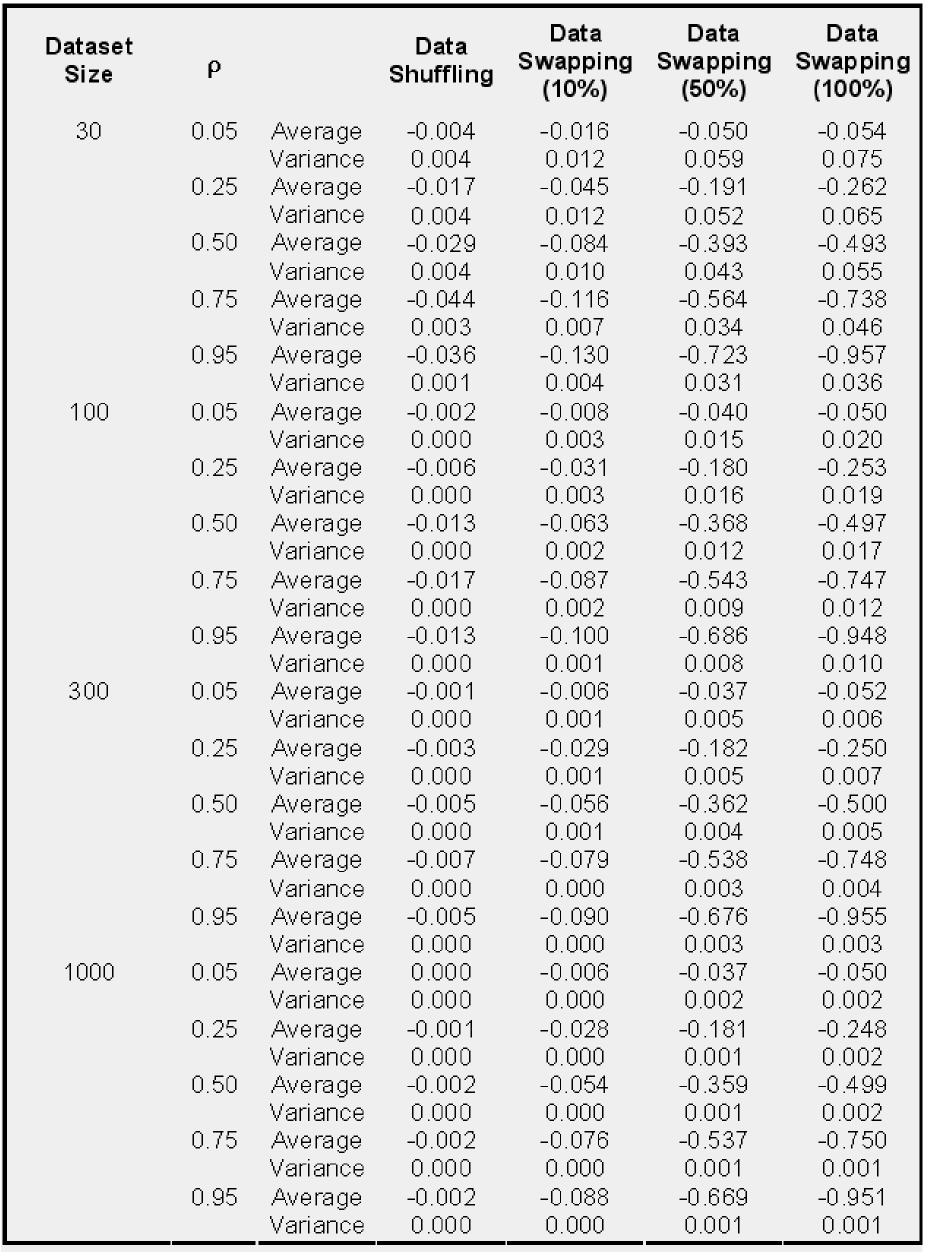
Data swapping procedures do not perform as well as data shuffling. Even when the data is swapped in close proximity (10%), there is a consistent attenuation (or reduction) in the rank order correlation for all n and ρ combinations. Even when proximity parameter is only 10%, for n = 30 and ρ = 0.95, the attenuation is consistent and of the order of –0.05. For the same experimental parameters, when swapping is performed with proximity parameter = 50% and ρ = 0.95, the average attenuation is -0.69 and for proximity parameter = 100%, the average attenuation is -0.94. For a bivariate normal population, when ρ = 0.95 the rank order correlation is of the order of 0.945. If we employ swapping with proximity parameter of 50%, the resulting rank order correlation is like to be of the order of only 0.25. This would lead users to conclude that the strength of the relationship between the two variables is much smaller than the actual relationship in the original dataset. More importantly, for exactly the same datasets, the results indicate that rank order correlation of the masked data is very close to that of the original data. Thus, if maintaining rank order correlation is used as the criteria for assessing data utility of the masked data, then the results of this experiment indicate that, in every case, data shuffling performs better than all three versions of the swapped data.

Table 1. Average and Variance of the Difference between the Rank Order Correlation of the Original and Masked Data



Dataset Size	ρ		Data Shuffling	Data Swapping (10%)	Data Swapping (50%)	Data Swapping (100%)
30	0.05	Average	0.002	-0.008	-0.041	-0.050
		Variance	0.009	0.005	0.056	0.074
	0.25	Average	-0.005	-0.023	-0.175	-0.249
		Variance	0.009	0.005	0.050	0.066
	0.50	Average	0.002	-0.038	-0.351	-0.467
		Variance	0.008	0.004	0.044	0.058
	0.75	Average	0.002	-0.056	-0.515	-0.707
		Variance	0.005	0.003	0.034	0.048
	0.95	Average	0.001	-0.065	-0.691	-0.941
		Variance	0.001	0.001	0.030	0.037
100	0.05	Average	0.000	-0.002	-0.030	-0.046
		Variance	0.002	0.001	0.014	0.020
	0.25	Average	-0.002	-0.015	-0.163	-0.242
		Variance	0.002	0.001	0.015	0.019
	0.50	Average	-0.001	-0.031	-0.340	-0.477
		Variance	0.002	0.001	0.013	0.017
	0.75	Average	0.000	-0.042	-0.510	-0.727
		Variance	0.001	0.001	0.010	0.013
	0.95	Average	0.000	-0.049	-0.663	-0.938
		Variance	0.000	0.000	0.008	0.011
300	0.05	Average	-0.001	-0.003	-0.035	-0.050
		Variance	0.001	0.000	0.005	0.006
	0.25	Average	0.000	-0.014	-0.167	-0.238
		Variance	0.001	0.000	0.005	0.007
	0.50	Average	0.000	-0.027	-0.334	-0.482
		Variance	0.001	0.000	0.005	0.006
	0.75	Average	0.000	-0.039	-0.508	-0.732
		Variance	0.000	0.000	0.003	0.005
	0.95	Average	0.000	-0.044	-0.656	-0.949
		Variance	0.000	0.000	0.003	0.003
1000	0.05	Average	0.000	-0.003	-0.034	-0.047
		Variance	0.000	0.000	0.001	0.002
	0.25	Average	0.000	-0.013	-0.165	-0.236
		Variance	0.000	0.000	0.001	0.002
	0.50	Average	0.000	-0.026	-0.332	-0.481
		Variance	0.000	0.000	0.001	0.002
	0.75	Average	0.000	-0.037	-0.507	-0.733
		Variance	0.000	0.000	0.001	0.001
	0.95	Average	0.000	-0.043	-0.650	-0.945
		Variance	0.000	0.000	0.001	0.001

Table 2. Average and Variance of the Difference between the Product Moment Correlation of the Original and Masked Data



Dataset Size	ρ		Data Shuffling	Data Swapping (10%)	Data Swapping (50%)	Data Swapping (100%)
30	0.05	Average	-0.004	-0.016	-0.050	-0.054
		Variance	0.004	0.012	0.059	0.075
	0.25	Average	-0.017	-0.045	-0.191	-0.262
		Variance	0.004	0.012	0.052	0.065
	0.50	Average	-0.029	-0.084	-0.393	-0.493
		Variance	0.004	0.010	0.043	0.055
	0.75	Average	-0.044	-0.116	-0.564	-0.738
		Variance	0.003	0.007	0.034	0.046
	0.95	Average	-0.036	-0.130	-0.723	-0.957
		Variance	0.001	0.004	0.031	0.036
100	0.05	Average	-0.002	-0.008	-0.040	-0.050
		Variance	0.000	0.003	0.015	0.020
	0.25	Average	-0.006	-0.031	-0.180	-0.253
		Variance	0.000	0.003	0.016	0.019
	0.50	Average	-0.013	-0.063	-0.368	-0.497
		Variance	0.000	0.002	0.012	0.017
	0.75	Average	-0.017	-0.087	-0.543	-0.747
		Variance	0.000	0.002	0.009	0.012
	0.95	Average	-0.013	-0.100	-0.686	-0.948
		Variance	0.000	0.001	0.008	0.010
300	0.05	Average	-0.001	-0.006	-0.037	-0.052
		Variance	0.000	0.001	0.005	0.006
	0.25	Average	-0.003	-0.029	-0.182	-0.250
		Variance	0.000	0.001	0.005	0.007
	0.50	Average	-0.005	-0.056	-0.362	-0.500
		Variance	0.000	0.001	0.004	0.005
	0.75	Average	-0.007	-0.079	-0.538	-0.748
		Variance	0.000	0.000	0.003	0.004
	0.95	Average	-0.005	-0.090	-0.676	-0.955
		Variance	0.000	0.000	0.003	0.003
1000	0.05	Average	0.000	-0.006	-0.037	-0.050
		Variance	0.000	0.000	0.002	0.002
	0.25	Average	-0.001	-0.028	-0.181	-0.248
		Variance	0.000	0.000	0.001	0.002
	0.50	Average	-0.002	-0.054	-0.359	-0.499
		Variance	0.000	0.000	0.001	0.002
	0.75	Average	-0.002	-0.076	-0.537	-0.750
		Variance	0.000	0.000	0.001	0.001
	0.95	Average	-0.002	-0.088	-0.669	-0.951
		Variance	0.000	0.000	0.001	0.001

The results of the simulation experiment to assess the extent to which the data masking procedures maintain product moment correlation is provided in Table 2. As in the previous table, Table 2 provides, for each n and ρ combination, the average
]

difference in the product moment correlation of the masked data and the original data and the variance of the difference. Carlson and Salabasis [2] have shown that replacing a related dataset with another (even independently generated) dataset results in attenuation of the product moment correlation. Hence, we expect all masking procedures used in this study to exhibit attenuation in correlation. However, as the size of the dataset increases, the attenuation should decrease and approach zero.

Data shuffling performs extremely well in maintaining product moment correlation as well. For all datasets of size 300 and 1000, the correlation attenuation is very small (–0.006). For dataset of size 100, the attenuation is of the order of –0.017 when ρ = 0.75 and smaller in all other cases. The largest differences are observed for n = 30 where for ρ = 0.75, the attenuation is of the order of –0.043. The variance of the difference is relatively small and never exceeds 0.005 in any case.

Among the swapped datasets, the best results are observed when the proximity parameter is set to 10%. The level of attenuation ranges from a high of –0.130 (n = 30, ρ = 0.95) to a low of –0.004 (n = 1000, ρ = 0.05). When the proximity parameter is 50%, the level of attenuation ranges from –0.034 (n = 1000, ρ = 0.05) to a high of –0.722 (n = 30, ρ = 0.95), and for the proximity parameter value of 100%, the attenuation ranges from –0.050 (n = 1000, ρ = 0.05) to –0.957 (n = 30, ρ = 0.95). In general, these results also support the theoretical derivations provided by Moore [10] regarding the reduction in correlation when data swapping is used.

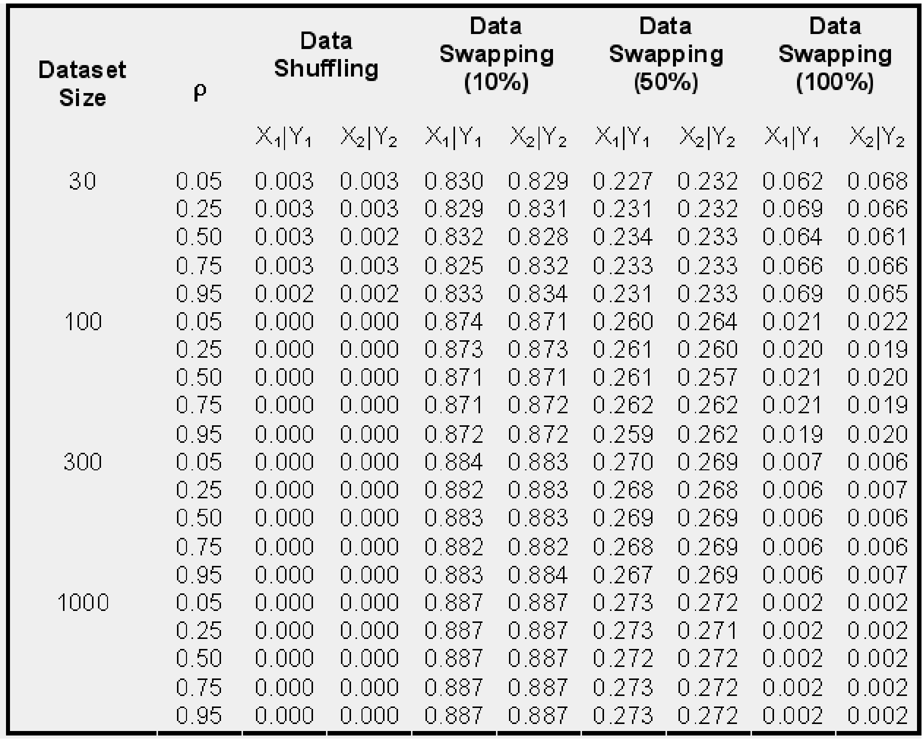
It is clear that users would be better off using shuffling as the making procedure instead of swapping, since based on the conclusive evidence in Table 2 where, in every case considered, the average attenuation in correlation from the shuffled data is smaller than that from the swapped data.

In summary, when data utility is evaluated either in terms of the ability to maintain rank order correlation or product moment correlation, the shuffled data performs better than all three swapped datasets for every combination of n and ρ. Hence, in terms of data utility, the performance of data shuffling is superior to the performance of data swapping. In the following section, we show that this is true for the ability to prevent risk of disclosure as well.

5.2 Disclosure Risk

Disclosure risk was evaluated using two alternate procedures, namely, risk of identity disclosure and the risk of value disclosure. The results of the experiment conducted to assess value disclosure are provided in Table 3. For each n and ρ combination, Table 3 provides the proportion of variability explained in the original variable (X1 or X2) using the corresponding masked variable (Y1 or Y2). As discussed earlier, one of the attractive features of data shuffling is that it is based on the conditional distribution approach and hence should provide the highest possible level of security [12]. The results in Table 3 verify this. Using the released variable Yi the proportion of variability explain in the corresponding confidential variable Xi is very small. Even when the size of the dataset is small (n = 30), the proportion of variability explained in Xi using shuffled Yi never exceeds 0.004. For all larger datasets, the proportion of variability explained is practically negligible and never exceeds 0.0004.

Table 3. Proportion of Variability in the Original Variable Explained by the Masked Variable



Dataset Size	ρ	Data Shuffling		Data Swapping (10%)		Data Swapping (50%)		Data Swapping (100%)	
		$X_1\|Y_1$	$X_2\|Y_2$	$X_1\|Y_1$	$X_2\|Y_2$	$X_1\|Y_1$	$X_2\|Y_2$	$X_1\|Y_1$	$X_2\|Y_2$
30	0.05	0.003	0.003	0.830	0.829	0.227	0.232	0.062	0.068
	0.25	0.003	0.003	0.829	0.831	0.231	0.232	0.069	0.066
	0.50	0.003	0.002	0.832	0.828	0.234	0.233	0.064	0.061
	0.75	0.003	0.003	0.825	0.832	0.233	0.233	0.066	0.066
	0.95	0.002	0.002	0.833	0.834	0.231	0.233	0.069	0.065
100	0.05	0.000	0.000	0.874	0.871	0.260	0.264	0.021	0.022
	0.25	0.000	0.000	0.873	0.873	0.261	0.260	0.020	0.019
	0.50	0.000	0.000	0.871	0.871	0.261	0.257	0.021	0.020
	0.75	0.000	0.000	0.871	0.872	0.262	0.262	0.021	0.019
	0.95	0.000	0.000	0.872	0.872	0.259	0.262	0.019	0.020
300	0.05	0.000	0.000	0.884	0.883	0.270	0.269	0.007	0.006
	0.25	0.000	0.000	0.882	0.883	0.268	0.268	0.006	0.007
	0.50	0.000	0.000	0.883	0.883	0.269	0.269	0.006	0.006
	0.75	0.000	0.000	0.882	0.882	0.268	0.269	0.006	0.006
	0.95	0.000	0.000	0.883	0.884	0.267	0.269	0.006	0.007
1000	0.05	0.000	0.000	0.887	0.887	0.273	0.272	0.002	0.002
	0.25	0.000	0.000	0.887	0.887	0.273	0.271	0.002	0.002
	0.50	0.000	0.000	0.887	0.887	0.272	0.272	0.002	0.002
	0.75	0.000	0.000	0.887	0.887	0.273	0.272	0.002	0.002
	0.95	0.000	0.000	0.887	0.887	0.273	0.272	0.002	0.002

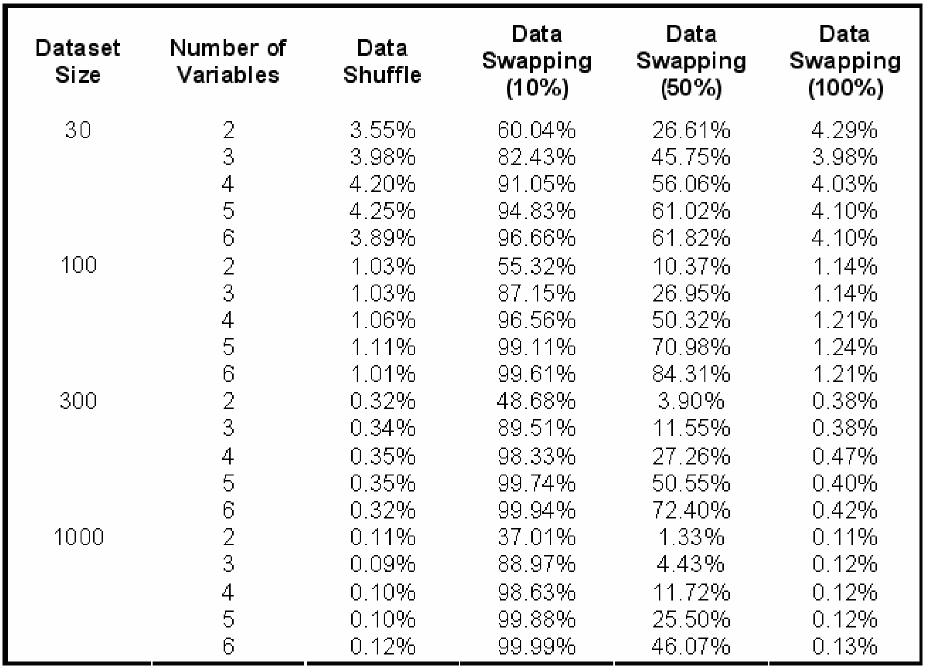
By contrast, the performance of data swapping is very poor. When the proximity parameter is 10% (that is records within close proximity are swapped), the proportion of variability explained is very high. In every case considered, the proportion of variability explained exceeds 0.820. This implies that an intruder would be able to get a very accurate estimate of the true value of the confidential variable using the released data. One interesting aspect is that, as the size of the dataset increases, the predictive ability of the intruder increases correspondingly. When n = 30 and the proximity parameter is set to 10%, the proportion of variability explained ranges from (0.825 to 0.834) while that for n = 1000 is consistently around 0.887. Similar results are observed when the proximity parameter is set to 50%, although the results are not quite as bad as those for proximity parameter = 10%. When the proximity parameter is set to 100%, the proportion of variability explained in Xi using the swapped Yi reduces dramatically. Another interesting aspect is that for larger sample sizes, the proportion of variability explained actually decreases.

Comparing the two procedures is relatively simple. In every case considered, the proportion of variability explained in Xi using the shuffled Yi is smaller than the proportion of variability explained in Xi using the swapped Yi. Thus, releasing the shuffled data provides an intruder with less information regarding the original variables, thereby resulting in lower level of disclosure risk.

The results for identity disclosure risk are very similar and are provided in Table 4. This table provides, for several n and for varying number of variables, the proportion of records re-identified using the procedure suggested by Fuller [7]. While other procedures for record linkage are available [18], Fuller's procedure is optimal for multivariate normal datasets that were used in the experiment. The results indicate that the shuffled data provides the greatest protection against identity disclosure. In almost every case, the proportion of records re-identified using the shuffled data is very close to the probability of re-identification by chance (1/n). For instance, when n = 30, the proportion of records re-identified for any number of variables is around 4% whereas the probability of re-identification by chance is approximately 3.33%. When n = 1000, the probability of re-identification by change is approximately 0.10% and the actual re-identification rate for the shuffle data is in the range 0.09% to 0.12%. Thus, we can conclude that the re-identification rate for the shuffled data is practically the same as the re-identification by chance. This again provides empirical evidence to support the conditional distribution approach used to generate the shuffled data.

The performance of data swapping varies widely. When the proximity parameter is set to 10%, the proportion of records re-identified is very high. With 6 confidential variables, an intruder would be able to identify at least 96% of the records (n = 30) to as high as 99.99% of the records (n = 1000). Even when there are only 2 confidential variables, an intruder would be able to identify as much as 60% (n = 30) and at least 37% (n = 1000). The re-identification results when the proximity parameter is set to 50% are a better than those observed when the proximity parameter is 10%. When the proximity parameter is set to 100%, the re-identification results are, in general, low.

Table 4. Percentage of Records Re-Identified



Dataset Size	Number of Variables	Data Shuffle	Data Swapping (10%)	Data Swapping (50%)	Data Swapping (100%)
30	2	3.55%	60.04%	26.61%	4.29%
	3	3.98%	82.43%	45.75%	3.98%
	4	4.20%	91.05%	56.06%	4.03%
	5	4.25%	94.83%	61.02%	4.10%
	6	3.89%	96.66%	61.82%	4.10%
100	2	1.03%	55.32%	10.37%	1.14%
	3	1.03%	87.15%	26.95%	1.14%
	4	1.06%	96.56%	50.32%	1.21%
	5	1.11%	99.11%	70.98%	1.24%
	6	1.01%	99.61%	84.31%	1.21%
300	2	0.32%	48.68%	3.90%	0.38%
	3	0.34%	89.51%	11.55%	0.38%
	4	0.35%	98.33%	27.26%	0.47%
	5	0.35%	99.74%	50.55%	0.40%
	6	0.32%	99.94%	72.40%	0.42%
1000	2	0.11%	37.01%	1.33%	0.11%
	3	0.09%	88.97%	4.43%	0.12%
	4	0.10%	98.63%	11.72%	0.12%
	5	0.10%	99.88%	25.50%	0.12%
	6	0.12%	99.99%	46.07%	0.13%

Comparing shuffling and swapping, as in the previous cases, for every combination of n and ρ, the results for the shuffled data are better than those observed for the swapped data. It should be noted however, that the results for the shuffled data and those for the swapped data when the proximity parameter is 100% are very close to each other (but shuffling is still better albeit by a small margin). Thus, in terms of identity disclosure, we reach the same results for all other criteria, that by providing lower rates of re-identification resulting in lowered disclosure risk, the shuffled data dominates the performance of all 3 sets of swapped data.

In summary, the conclusion that we reach from the assessing disclosure risk is that by both measures, namely, value disclosure and identity disclosure, releasing the shuffled data results in lower risk of disclosure that releasing the swapped data, regardless of the size of the dataset and the number of variables.

Overall, when we consider both data utility and disclosure risk, the results are straight-forward. On every criteria considered in this study, the results of the simulation experiments indicate that the shuffled data provide better results (better data utility or lower disclosure risk) than the swapped data. Hence, we can conclude that data shuffling is a better data masking procedure than data swapping.

6 Conclusion

The objective of this study was to investigate the relative performance of data shuffling and data swapping. We conducted several simulation experiments to perform this evaluation. The results of the simulation experiments are consistent and clear. Regardless of the characteristics of the dataset used in the experiments, and regardless of the criteria used for evaluation (data utility based on rank order or product moment correlation and disclosure risk based on value or identity disclosure), data shuffling provides better performance (higher data utility and lower disclosure risk) than data swapping. Hence, data shuffling should always be the preferred approach for data masking.

Finally, early research on disclosure control techniques, particularly those relating to numerical microdata seemed to imply an inherent, unavoidable trade-off between data utility and disclosure risk. The implication of this trade-off was that if a technique resulted in higher data utility, an increased level of disclosure risk was also unavoidable. However, recent studies [1, 12, 13] have provided strong evidence to indicate that this trade-off may not always exist and that some techniques may simultaneously provide higher data utility and lower disclosure risk. Muralidhar and Sarathy [13] theoretically proved that data shuffling provides better (or equal) data utility and lower (or equal) disclosure risk compared to data swapping (with any proximity parameter) which is verified by the empirical evidence in this study.

References

1. Burridge, J. 2003. Information Preserving Statistical Obfuscation, Statistics and Computing, 13 321-327.
2. Carlson, M. and M. Salabasis. 2002. A data swapping technique for generating synthetic samples: A method for disclosure control. Research in Official Statistics. 6 35-64.

3. Dalenius, T. and Reiss, S. P. 1982. Data-swapping: A Technique for Disclosure Control. Journal of Statistical Planning and Inference 6 73-85.
4. Dandekar, R.A., M. Cohen, and N. Kirkendall 2002. Sensitive Microdata Protection Using Latin Hypercube Sampling Technique. In Inference Control in Statistical Databases (J. Domingo-Ferrer, Editor), Springer-Verlag, New York.
5. Domingo-Ferrer, J., and V. Torra. 2001 Disclosure control methods and information loss for microdata, in: P. Doyle, J. Lane, J. Theeuwes, L. Zayatz (Eds.), Confidentiality, Disclosure and Data Access, North-Holland, Amsterdam, 91-110.
6. Fienberg, S.E. and J. McIntyre. 2005. Data swapping: Variations on a theme by Dalenius and Reiss. Journal of Official Statistics, 21 309-323.
7. Fuller, W.A. 1993. Masking procedures for microdata disclosure limitation. Journal of Official Statistics 9 383-406.
8. Iman, R.L. and W.J. Conover. 1982. A distribution free approach to inducing rank correlation among input variables. Communication in Statistics B11 311-334.
9. McKay, M.D., W.J. Conover, and R.J. Beckman. 1979. A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. Technometrics 21 239-245.
10. Moore, R.A. 1996. Controlled data swapping for masking public use microdatasets. U.S. Census Bureau Research Report 96/04.
11. Muralidhar, K. and R. Sarathy, 2002. "Application of the Two-step Data Shuffle to the 1993 AHS Data: A Report on the Feasibility of Applying Data Shuffling for Microdata Release," research report prepared for the Census Bureau (http://gatton.uky.edu/faculty/muralidhar/maskingpapers/).
12. Muralidhar, K. and R. Sarathy. 2003. A theoretical basis for perturbation methods. Statistics and Computing 13 329-335.
13. Muralidhar, K. and R. Sarathy. 2006. Data Shuffling - A New Masking Approach for Numerical Data. Management Science 52 658-670.
14. Reiss, S.P., M.J. Post, and T. Dalenius. 1982. Non-reversible privacy transformations. Proceedings of the ACM Symposium on Principles of Database Systems, Los Angeles, CA 139-146.
15. Sarathy R., K. Muralidhar, R. Parsa. 2002. Perturbing non-normal confidential variables: The copula approach. Management Science 48 1613-1627.
16. Sarathy, R. and K. Muralidhar. 2002. The Security of Confidential Numerical Data in Databases. Information Systems Research 389-403.
17. Wall Street Journal 2001. Bureau Blurs Data to Keep Names Confidential. February 14 2001. B1-B2.
18. Winkler W. E. 1995. Advanced methods for record linkage. In: Proceedings of the American Statistical Association Section on Survey Research Methods. 467-472.

Adjusting Survey Weights When Altering Identifying Design Variables Via Synthetic Data

Robin Mitra and Jerome P. Reiter

Duke University, Durham, NC 27708, USA
{rm51, jerry}@stat.duke.edu
http://www.stat.duke.edu

Abstract. Statistical agencies alter values of identifiers to protect respondents' confidentiality. When these identifiers are survey design variables, leaving the original survey weights on the file can be a disclosure risk. Additionally, the original weights may not correspond to the altered values, which impacts the quality of design-based (weighted) inferences. In this paper, we discuss some strategies for altering survey weights when altering design variables. We do so in the context of simulating identifiers from probability distributions, i.e. partially synthetic data. Using simulation studies, we illustrate aspects of the quality of inferences based on the different strategies.

Keywords: Disclosure; Multiple imputation; Swapping; Synthetic data; Weights.

1 Introduction

Survey design variables often contain identifying information, for example race in a survey that over-samples minorities or establishment size in a probability proportional to size sample of businesses. To limit disclosure risks, statistical agencies may need to alter these variables before releasing the data to the public. It also may be necessary to alter the survey weights, which typically are deterministic functions of the design variables. Failure to do so can leave identifying information on the file, effectively defeating the purpose of the masking [1]. For example, an unaltered weight could reveal that a person was part of a minority group or could disclose the size of the establishment. Not altering weights also could affect the quality of data analysts' estimates, because the weights may not be appropriate for making the released sample representative of the population.

In this paper, we discuss some strategies for adjusting survey weights when altering design variables to limit disclosure risks. We do so in the context of simulating identifiers from probability distributions, i.e. partially synthetic data. Using simulation studies, we illustrate aspects of the data quality and confidentiality of the different strategies. We also examine the performance of the strategies when swapping identifiers.

J. Domingo-Ferrer and L. Franconi (Eds.): PSD 2006, LNCS 4302, pp. 177–188, 2006.

2 Partially Synthetic Data and Weights

We first review partially synthetic data. Then, we describe some strategies to adjust weights when replacing design variables with synthetic values.

2.1 Partial Synthesis

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 imputations. Releasing partially synthetic data can preserve confidentiality, since identification of units and their sensitive data can be difficult when some released data are not actual, collected values. Furthermore, using appropriate data generation and estimation methods [2]—based on the concepts of multiple imputation [3] for missing data—analysts can make valid inferences for a variety of estimands using standard, complete-data statistical methods and software, at least for inferences congenial to the model used to generate the data. Provided the agency releases some description of this model, analysts can determine whether or not their questions can be answered using the synthetic data. See [4] and [5] for genuine applications of partially synthetic data.

Following the derivations of [2], we assume that the agency synthesizes some design variables, X, based on the observed data, $D = (X, Y_{obs})$, by drawing new values from the Bayesian posterior predictive distribution of $(X|D)$. Imputations are made independently for $i = 1, \ldots, m$ times to yield m different synthetic data sets. These synthetic data sets are released to the public.

From these synthetic data sets, some user of the publicly released data, henceforth abbreviated as the analyst, seeks inferences about some estimand Q. In each synthetic data set d_i, the 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 synthetic data set d_i. The analyst can obtain valid inferences for scalar Q by using the following quantities:

$$\bar{q}_m = \sum_{i=1}^{m} q_i/m \tag{1}$$

$$b_m = \sum_{i=1}^{m} (q_i - \bar{q}_m)^2/(m-1) \tag{2}$$

$$\bar{v}_m = \sum_{i=1}^{m} v_i/m. \tag{3}$$

The analyst can then use the $\bar{q}_m$ to estimate Q and $T_p = b_m/m + \bar{v}_m$ to estimate the variance associated with $\bar{q}_m$. For large sample sizes, inferences for scalar Q can be based on t-distributions with degrees of freedom $\nu_p = (m-1)(1 + r_m^{-1})^2$, where $r_m = (m^{-1}b_m/\bar{v}_m)$ [2].

2.2 Survey Weights in Partial Synthesis

In complex surveys, it is well known that analyses that fail to account for the survey design variables can yield biased inferences [6] [7]. To incorporate the design, analysts can use survey-weighted estimation, where the survey weight w_i for unit i equals the inverse of the unit's inclusion probability, multiplied possibly by adjustments for nonresponse and calibration. For example, a common survey-weighted estimator of the population mean of Y based on the sample S is

$$\bar{y}_w = \frac{\sum_{i \in S} w_i y_i}{\sum_{i \in S} w_i}. \tag{4}$$

Weighted estimates exist for regression coefficients, as well as for the variances of these estimators.

When synthesizing some sampling design variables X, it is necessary to adjust weights to reflect the new values. We consider two approaches: (i) recalculate the weights (RCAL) to be consistent with the synthetic values, effectively making the synthetic sample representative of the population, and (ii) copy and paste (CPP) the original weights of records whose original design variables match the synthetic ones. The RCAL method preserves some properties of the original sampling weights that CPP does not; for example, the sum of the RCAL weights equals the sum of the observed weights, whereas the sums are not necessarily equal for the CPP weights. Additionally, the CPP cannot be applied unless exact matches are available.

3 Simulation Studies

In this section, we use simulation studies to investigate the implications for data quality of using survey-weighted analyses based on weights from the RCAL and CPP methods. For comparisons, we also consider using unweighted (UNW) estimates and survey-weighted estimates based on the old weights (OLDW). The simulations include stratified sampling, probability proportional to size sampling, and two stage cluster sampling. We also apply the procedures on genuine data from the Survey of Youth in Custody [6]. Unless stated otherwise, all estimates and standard errors are calculated using the survey package in the R statistical software.

3.1 Stratified Sampling Simulation

We first generate a stratified population of size 20000. The four strata are formed by crossing two binary variables, X_1 and X_2. There are approximately 1000, 2000, 10000, and 7000 units in stratum one through four, respectively. We generate two survey variables, Y_1 and Y_2, from the following distributions:

$$y_{1h} \sim N(\mu_h, \sigma_h^2) \tag{5}$$

$$y_{2h} \sim N(\alpha_h + \beta_h y_{1h}, \tau_h^2) \tag{6}$$

where μ_h, σ_h^2, α_h, β_h, and τ_h^2 differ for each stratum h. The observed data are a stratified sample of 250 units from each stratum, so that the weight for all observations in stratum one equals 4, in stratum two equals 8, in stratum three equals 40, and in stratum four equals 28.[1]

For each sample, we synthesize (X_1, X_2, Y_2) from their joint posterior predictive distribution conditional on Y_1, which remains unaltered. To do so, we first simulate replacement values for X_1 by using a logistic regression conditional on Y_1. Second, we simulate replacement values for X_2 by using a logistic regression on (Y_1, X_1), using the synthetic X_1 for predictions. Third, we simulate replacement values for Y_2 using a linear regression conditional on (Y_1, X_1, X_2), using the synthetic X_1 and X_2 for predictions. We repeat this process independently $m = 5$ times to obtain five partially synthetic datasets for each D.

We run this simulation 1,000 times. In each replication, we obtain confidence intervals for the means of Y_1 and Y_2; the percentages of values of Y_1 greater than the population 50th, 80th and 95th percentiles, and likewise for Y_2; the two regression coefficients from the linear regression of Y_2 on Y_1, the four regression coefficients from the linear regression of Y_2 on (Y_1, X_1, X_2); and, the eight regression coefficients from the linear regression of Y_2 on (Y_1, X_1, X_2) and their interactions. The synthetic 95% confidence intervals are based on the methods in Section 2.1, with $\bar{v}_m$ equal to the design-based variance estimate as computed in the *R* software. Because of how *R* computes variances, the $\bar{v}_m$ is the same for the RCAL and CPP methods, although the point estimates of Q differ.

To illustrate RCAL and CPP in this setting, suppose one synthesized dataset comprises 200 records in each of stratum one and stratum two, and 300 records in each of stratum three and stratum four. Using RCAL, the new weight of all records in stratum one equals 5, in stratum two equals 10, in stratum three equals 100/3, and in stratum four equals 70/3. Using CPP, the weight of all records in stratum one remains at 4, in stratum two remains at 8, in stratum one remains at 40, and in stratum four remains at 28.

Figure 1 displays box plots of the percentages of the 95% confidence intervals that contain their corresponding population quantities. As expected, the coverage rates based on the observed data (OBS) are around 95%. Those based on RCAL and CPP also are near 95%. Coverage rates based on method OLDW do not match those based on the observed data. OLDW is particularly problematic for analyses involving Y_2. Coverage rates based on method UNW are too low for the means and proportions. This is not surprising, since unweighted means and percentages are known to be biased in unequal probability samples. Method UNW does provide coverage rates like those for unweighted analyses based on the observed data (OBSUNW).

As a check on the amount of alteration in the strata, for each sample we compare the modes of the m imputed values of the records' synthetic stratum indicators to their actual stratum indicators in the observed data. Approximately 45% of records can be placed in their original stratum by using this strategy, indicating a sizeable number of re-allocations of stratum memberships.

[1] The weights actually are slightly different from the integer values because the strata sizes are not precisely 1000, 2000, 10000, and 7000.

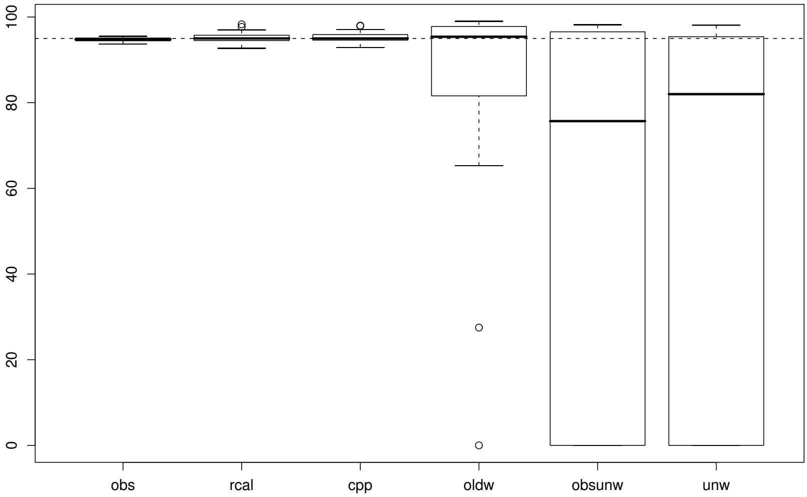


Fig. 1. Box plots of coverage rates for the twenty-three estimands in the stratified sampling simulation. The coverage rates based on *RCAL* or *CPP* are closest to the those based on the observed data.

3.2 PPS Sampling Simulation

We generate a population of size 20000 in which the design variable X is a size variable. We generate X from a log-normal distribution and add a constant so that all values are far from zero. The minimum size in the population equals 50, and the maximum size equals 412. The total of the size values equals 1,328,252. We then generate the survey variables Y_1 and Y_2 from

$$y_1 \sim N(1.3x, 32^2) \tag{7}$$
$$y_2 \sim N(1.2x + 0.9y_1, 32^2). \tag{8}$$

This results in correlations between X and Y_1 of 0.63, between X and Y_2 of 0.76, and between Y_1 and Y_2 of 0.82. We sample 1000 records from this population with probability proportional to the size variable X using the Hartley-Rao algorithm [8] For any observation i, the weight is $w_i = 1328252/(1000x_i)$.

We synthesize (X, Y_2) from their joint Bayesian posterior predictive distribution conditional on Y_1, which remains unaltered. To do so, we first simulate replacement values for X by using a generalized additive model (GAM) conditional on Y_1.[2] Second, we simulate replacement values for Y_2 by using a linear regression on (Y_1, X), using the synthetic X for predictions. We repeat this process independently $m = 5$ times to obtain five partially synthetic datasets for each D.

We run this simulation 1,000 times. In each replication, we obtain confidence intervals for the means of X, Y_1, and Y_2; the percentages of values of Y_1 greater

[2] For more details and code to implement this procedure, contact the second author.

than the population 50th, 80th and 95th percentiles, and likewise for X and Y_2; the six regression coefficients from the linear regression of Y_2 on Y_1, Y_2 on X, and Y_1 on X; and, the three regression coefficients from the linear regression of Y_2 on (Y_1, X). The synthetic 95% confidence intervals are based on the methods in Section 2.1, with $\bar{v}_m$ equal to the design-based variance estimate as computed in the *R* software.

The CPP method is not applicable here, because the simulated sizes do not match exactly with original sizes. The RCAL simply involves plugging in the synthesized values of the x_j in $1328252/(1000x_j)$.

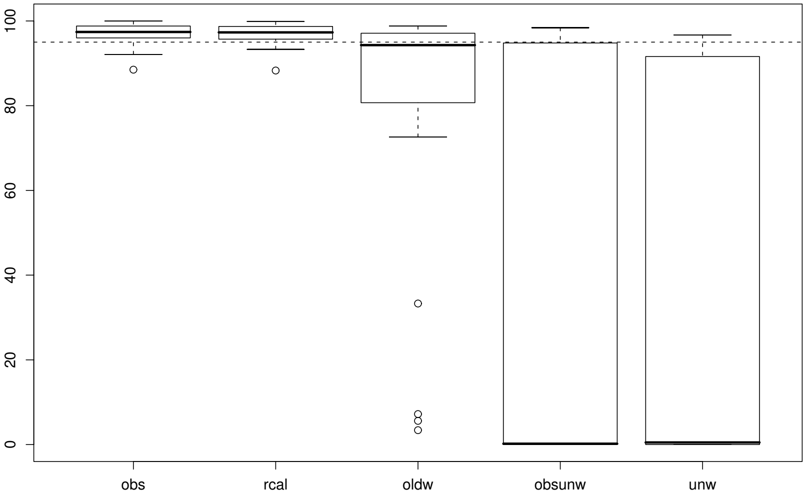


Fig. 2. Box plots of coverage rates for the twenty-one estimands in the PPS simulation. The coverage rates based on *RCAL* are closest to the those based on the observed data.

Figure 2 displays box plots of the percentages of the 95% confidence intervals that contain their corresponding population quantities. The coverage rates based on the observed data are slightly higher than 95%, because we did not specify the finite population correction in variance estimates. The coverage rates based on RCAL closely match those based on the observed data. Those based on OLDW do not match the observed data coverage rates, especially for analyses involving the size variable. Method UNW tends to have poor coverages for means and proportions, producing biased estimates of the population quantities as expected.

As a check on the amount of alteration in the size measures, for each record we compute the average of the five synthetic sizes. We then find the record in the population with the closest actual size to that average. Using this approach, approximately 0.2% of the respondents are correctly re-identified from the synthetic data. In contrast, releasing the old weights completely undoes the protection of the synthesis of size, since the original size can be backed out of the original weight.

3.3 Two Stage Cluster Sampling

We generate a population of 20000 units in which the data are grouped in 200 clusters. Twenty clusters have size 200; forty clusters have size 150; sixty clusters have size 100; and, eighty clusters have size 50. We generate the survey variables Y_1 and Y_2 from

$$y_1 \sim N(20 + \omega_c, 3^2) \tag{9}$$

$$y_2 \sim N(1.2y_1 + \delta_c, 20^2) \tag{10}$$

where ω_c and δ_c differ for each cluster c. We select observed data from this population using two-stage cluster sampling. We take a simple random sample of 40 clusters from the population of 200. For each sampled cluster, we take a simple random sample of 25 observations.

We synthesize the observed cluster indicators X and Y_2 from their joint posterior predictive distribution conditional on Y_1. First, we simulate replacement values for X by using a multinomial logistic regression model conditional on Y_1. Only the observed clusters in each sample are used to fit the model. Second, we simulate replacement values for Y_2 from a normal linear regression on (Y_1, X), using the synthetic X for predictions. We repeat this process independently $m = 5$ times to obtain five partially synthetic datasets for each D.

We run this simulation 1,000 times. In each replication, we obtain confidence intervals for the means of Y_1, and Y_2; the percentages of values of Y_1 greater than the population 50th, 80th and 95th percentiles, and likewise for Y_2; the regression coefficient of Y_1 from the linear regression of Y_2 on Y_1; and, the regression coefficient of Y_1 from the linear regression of Y_2 on (Y_1, X). The synthetic 95% confidence intervals are based on the methods in Section 2.1, with $\bar{v}_m$ equal to the design-based variance estimate as computed with (i) our own variance estimation code for means and percentages and (ii) the R survey package for regression coefficients.

For any record in cluster c, the weight equals the product of five and the inverse of the fraction of records sampled in that cluster. To apply RCAL, only the second term in the multiplication changes, depending on the new fraction of records in each cluster. To apply CPP, we use the same process illustrated in Section 3.1. As in the stratified sampling simulation, variance estimates of means and proportions are the same for methods RCAL and CPP.

Figure 3 displays box plots of the percentages of the 95% confidence intervals that contain their corresponding population quantities. The coverage rates based on the observed data are slightly lower than 95%, whereas the coverage rates based on RCAL and CPP are nearly 95%. The coverage rates based on OLDW are like those based on RCAL and CPP. In this constructed population, many weights do not change substantially after applying RCAL and CPP—even though cluster memberships change—due to the identical second stage sampling rate and the existence of many clusters of the same population size. This also explains why RCAL and CPP result in similar coverage rates. Coverage rates based on UNW are very low for means and proportions but close to 95% for the regression coefficients.

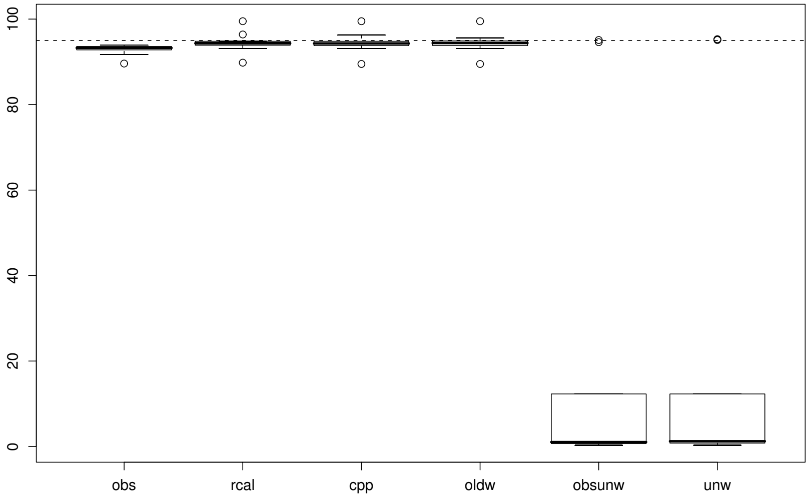


Fig. 3. Box plots of coverage rates for the ten estimands in the cluster sampling simulation. The coverage rates based on *RCAL*, *CPP*, and OLDW are close to the those based on the observed data.

As a check on the amount of alteration in the cluster indicators, we compare the units' modal synthesized cluster indicators to their corresponding observed indicators, like the strategy used in the stratified simulation,. Approximately 15% of units can be placed in their original cluster, indicating a sizeable number of re-allocations of cluster memberships.

3.4 Survey of Youth in Custody

We now examine the performance of RCAL, the method that performs best across all the simulations, on genuine data from the 1987 Survey of Youth in Custody. The survey interviewed youths in juvenile institutions about their family background, previous criminal history, and drug and alcohol use. The sampling frame comprises 206 facilities. The eleven facilities (strata 6 to 16) with more than 360 youths were treated as strata. The remaining facilities were divided in five strata (strata 1 to 5) based on size. These facilities were sampled with probability proportional to size, and residents within sampled facilities were sampled with predetermined sampling fractions. The sample contains 50 facilities and 2,621 youths.

To simplify the illustration, we deleted four facilities for which size was unknown and ignored the small amount of unit nonresponse. We re-specified the original survey weights to reflect the smaller number of facilities and clusters in this reduced dataset. We filled in the small number of missing item values using univariate re-sampling.[3]

[3] We recommend using the principled approach of multiple imputation to handle missing data, but the univariate re-sampling is adequate to illustrate the performance of the RCAL approach.

Table 1. Point and interval estimates based on observed and synthetic data for Survey of Youth in Custody

Variable	q_{obs}	Obs. 95% CI	$\bar{q}_5$	Syn. 95% CI
Avg. age	16.7	(16.6, 16.8)	16.8	(16.7, 16.9)
Avg. age at first arrest: Hispanics	13.0	(12.7, 13.2)	13.0	(12.6, 13.2)
Avg. age at first arrest: non-Hispanics	13.0	(12.9, 13.1)	13.0	(12.8, 13.1)
% with age at first arrest < 15	73.4	(71.3, 75.5)	73.1	(70.8, 75.4)
% with age at first arrest > 18	.39	(.16, .62)	.40	(.15, .64)
% used drugs	25.4	(23.4, 27.3)	25.2	(23.2, 27.1)
% females	7.4	(6.1, 8.6)	7.5	(6.1, 9.0)
Coefficients in logistic regression of ever violent on				
Intercept	1.36	(.80, 1.93)	1.33	(.73, 1.92)
Age at first arrest	$-.083$	$(-.126, -.041)$	$-.082$	$(-.127, -.037)$
Black	.46	(.25, .67)	.48	(.27, .69)
Asian	.33	$(-.72, 1.38)$	.76	$(-.28, 1.79)$
American Indian	$-.014$	$(-.551, .523)$	$-.088$	$(-.726, .549)$
Other	1.35	(.56, 2.15)	1.21	(.42, 2.00)

We consider facility membership to be potentially identifying information. Therefore, we generate new facility identifiers for all records in the dataset. For strata 6 to 16, we synthesize the stratum value for each observation using a multinomial regression estimated with records in strata 6 to 16 only. For purposes of illustration, we include main effects for all twenty-two predictors in the regression model except for race, education, and who the youth lived with before being institutionalized. These variables are excluded to enable the model to be identifiable, since there is multi-collinearity in the data. For strata 1 - 5, we synthesize the facility indicators using another multinomial regression estimated with records in strata 1 to 5 only. This model excludes from the synthesis model the youth's race, education, who they lived with, whether anyone in the family served time, the type of crime, and their alcohol use. More terms are dropped because the sample sizes in these facilities are small. A potentially more accurate synthesis model would incorporate informative prior distributions on the parameters of the logistic regression with all twenty-two predictors.

We create $m = 5$ partially synthetic data sets. We then recalculate the survey weights using the RCAL method, which involves recalculations like described in the cluster sampling simulation. Table 1 displays the observed and synthetic point estimates and 95% confidence intervals for a variety of estimands. All results are based on survey-weighted estimation based on the design. Generally, the observed and synthetic point estimates and confidence intervals are similar. The one possible exception is the regression coefficient for the indicator variable corresponding to Asian race; however, the observed and synthetic confidence intervals are relatively wide and overlap to an extent.

To check on the amount of alteration in the facility indicators, we apply the strategy used in the cluster sampling simulation—place the youth in its modal

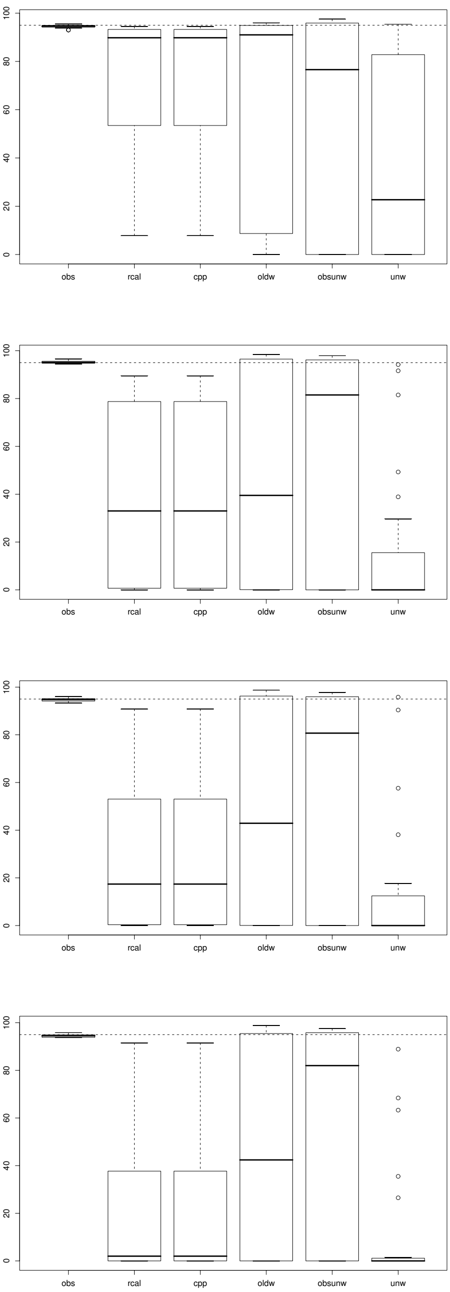


Fig. 4. Box plots of coverage rates for swapping percentages of 5%, 30%, 50%, and 100%, going from top panel to bottom panel. None of the methods based on the swapped data have satisfactory coverage properties.

imputed facility—and find that approximately 17% of youths can be placed in their original facility.

4 Extension to Data Swapping

In this section, we examine the performance of the weight adjustment procedures when swapping design variables. We use the stratified sampling simulation. Rather than synthesizing new stratum indicators, we assign some percentage of the stratum indicators to be randomly swapped, creating one masked dataset per observed dataset. Stratum indicators might be swapped with a like value, resulting in no change for the unit's stratum in the masked data.

We follow the simulation design in Section 3.1, except that we leave Y_2 unaltered. We consider swapping rates of 5%, 30%, 50% and 100%. Figure 4 displays box plots of the percentages of the 95% confidence intervals that contain their corresponding population quantities. The coverage rates get progressively worse as the degree of swapping increases. For the 5% swapping simulation, coverage rates based on RCAL or CPP are better than those based on OLDW or UNW, but they remain inadequate. With this version of swapping, the methods RCAL and CPP yield exactly the same weights and hence estimates, since there always are 250 records in each stratum.

Comparing swapping to partial synthesis, the coverage rates based on RCAL and CPP are much closer to nominal in the partially synthetic data than in the swapped data, even though we replaced the values of Y_2 in the former but not the latter. Using a swapping rate of 50% leaves approximately 45% of records' stratum indicators unchanged, which is the same percentage of records that can be placed in their correct stratum when using partially synthetic data. But, the partial synthesis clearly is more effective at preserving the statistical properties of the data.

5 Conclusions

The simulations in this paper illustrate the importance of survey weights when altering design variables to limit disclosure risks. Releasing the original weights can lead to biased inferences or compromise identity of respondents. At least for partially synthetic data, recalculating the weights to be consistent with released values can improve design-based estimation. Unfortunately, this approach does not appear to improve inferences sufficiently when using data swapping of design variables. Further research is needed to investigate the viability of the recalculation approach for more complicated multi-stage sampling schemes. Additionally, research is needed to see how adjustments for non-response and calibration interact with the recalculation approach.

Acknowledgments. This research was funded by the National Science Foundation grant ITR-0427889.

References

1. De Waal, A.G., Willenborg, L.C.R.J.: Statistical Disclosure Control and Sampling Weights. Journal of Official Statistics, **13** (1997) 417–434
2. Reiter, J.P.: Inference for Partially Synthetic, Public Use Microdata Sets. Survey Methodology, **29** (2003) 181–189
3. Rubin, D.B.: Multiple Imputation for Nonresponse in Surveys. John Wiley & Sons, New York (1987)
4. Kennickell, A.B.: Multiple Imputation and Disclosure Protection: The Case of the 1995 Survey of Consumer Finances. In: Alvey, W., Jamerson, B. (eds.): Record Linkage Techniques, 1997. National Academy Press, Washington, D.C. (1997) 248–267
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 North-Holland, Amsterdam (2001) 215–277
6. Lohr, S.L.: Sampling: Design and Analysis. Duxbury Press, New York (1999)
7. Reiter, J.P., Zanutto, E.L., Hunter, L.W.: Analytical Modeling in Complex Surveys of Work Practices. Industrial Labor Relations Review, **59** (2005) 82–100.
8. Valliant R., Dorfman A.H., Royall R.M.: Finite Population Sampling and Inference. John Wiley & Sons, New York (2000) 72–73.

Risk, Utility and PRAM

Peter-Paul de Wolf*

Department of Methods and Informatics
Statistics Netherlands
P.O. Box 4000
2270 JM Voorburg, The Netherlands
pwof@cbs.nl

Abstract. PRAM (Post Randomization Method) is a disclosure control method for microdata, introduced in 1997. Unfortunately, PRAM has not yet been applied extensively by statistical agencies in protecting their microdata. This is partly due to the fact that little knowledge is available on the effect of PRAM on disclosure control as well as on the loss of information it induces.

In this paper, we will try to make up for this lack of knowledge, by supplying some empirical information on the behaviour of PRAM. To be able to achieve this, some basic measures for loss of information and disclosure risk will be introduced. PRAM will be applied to one specific microdata file of over 6 million records, using several models in applying the procedure.

Keywords: Disclosure control, post randomisation method, information loss, disclosure risk.

1 Introduction

The Post Randomization Method (PRAM) was introduced in [1] as a method for disclosure protection applied to categorical variables in microdata files. In [2] and [3], the method and some of its implications were discussed in more detail.

PRAM produces a microdata file in which the scores on some categorical variables for certain records are changed with respect to the scores in the original microdata file. This is usually applied to identifying variables, i.e., variables that can be used to identify the respondent that corresponds to a record. This results in a microdata file with scores on identifying variables, that, with certain probability, are incorrect scores. Hence, the risk of identification of respondents is reduced: even in case one could make a link between a record in the microdata file and an individual, the possible incorrectness of the scores yields uncertainty on the correctness of the link.

Note that PRAM can be regarded as a form of misclassification, where the so called transition probabilities (i.e., the probabilities of changing a score into another score) completely determine the underlying probability mechanism. These

* The views expressed in this paper are those of the author and do not necessarily reflect the policy of Statistics Netherlands.

J. Domingo-Ferrer and L. Franconi (Eds.): PSD 2006, LNCS 4302, pp. 189–204, 2006.

transition probabilities are summarized in a Markov matrix called the PRAM-matrix. Contrary to the general situation, the probability mechanism that determines the misclassification is known in case of PRAM. Since the probability mechanism is known, some statistical analyses can still be performed legitimately, be it with a slight adjustment of the standard methods. See, e.g., [4], [5] and [6]. A similar situation of misclassification with known transition probabilities is the case of Randomized Response (see, e.g., [7] and [8]). In that case it has been known for some time, that unbiased estimates of population parameters can be obtained as well, see e.g., [9] and [10].

In order to let a user make legitimate inference, the transition probabilities should hence be supplied to him. On the other hand, making use of the literature on inference about misclassification mechanisms (see e.g., [10]), even without the exact transition probabilities a user could still perform sound analyses.

When applying Statistical Disclosure Control (SDC) methods, one has to deal with two competing goals: the microdata file has to be safe enough to guarantee the protection of individual respondents but at the same time the loss of information should not be too large. For a general discussion, see, e.g., [11]. Moreover, these competing mechanisms are often the core of the discussion of SDC methods themselves, see, e.g., [12] and [13]. However, quantifying the loss of information and the level of disclosure control can be done in several ways. We will introduce some basic measures to quantify the loss of information as well as a measure to determine the level of disclosure control in case of using PRAM.

In this paper we will apply PRAM to a microdata file of 6,237,468 records and discuss the effect of applying PRAM on the amount of information loss and the level of disclosure control, using different PRAM-matrices.

In Section 2 we will give a brief description of PRAM. Moreover, in this section we will introduce the notation concerning PRAM that we will be using throughout the rest of the paper. The aim of this paper is to investigate the effect of different PRAM-matrices on disclosure control as well as information loss. In Section 3 we will therefore define a measure to quantify the effect on disclosure control. Section 4 contains the definitions of the measures of information loss we used in our experiments. Both the effect on disclosure control and the effect on the amount of information, in the different experiments we performed, will be given in Section 5. Finally, in Section 6 we will briefly summarize the results and draw some conclusions.

2 A Short Description of PRAM

In this section we will briefly describe the theory involving PRAM, mainly to introduce the notation we will use throughout this paper. For details we refer to [2].

Let ξ denote a categorical variable in the original file to which PRAM will be applied and let X denote the same variable in the perturbed file. Moreover, assume that ξ, and hence X as well, has K categories, labeled $1, \dots, K$. The transition probabilities that define PRAM are denoted as

$$p_{kl} = \mathbb{P}(X = l \mid \xi = k), \tag{1}$$

i.e., the probability that an original score $\xi = k$ is changed into the score $X = l$, for all $k, l = 1, \ldots, K$. Using these transition probabilities as entries of a $K \times K$ matrix, we obtain a Markov matrix $\mathbf{P}$ that we will call the PRAM-matrix.

Applying PRAM now means that, given the score $\xi = k$ for record r, the score X for that record is drawn from the probability distribution $p_{k1}, \ldots, p_{kK}$. For each record in the original file, this procedure is performed independently of the other records.

To illustrate the ideas, suppose that the variable ξ is gender, with scores $\xi = 1$ if male and $\xi = 2$ if female. Applying PRAM with $p_{11} = p_{22} = 0.9$ on a microdata file with 110 males and 90 females, would yield a perturbed microdata file with, in expectation, 108 males and 92 females. However, in expectation, 9 of these males were originally female, and similarly, 11 of the females were originally male.

More generally, the effect of PRAM on one-dimensional frequency tables is that

$$\mathbb{E}\left(\boldsymbol{T}_X \mid \xi\right) = \mathbf{P}^t \, \boldsymbol{T}_\xi, \tag{2}$$

where $\boldsymbol{T}_\xi = (T_\xi(1), \ldots, T_\xi(K))^t$ denotes the frequency table according to the original microdata file and $\boldsymbol{T}_X$ the frequency table according to the perturbed microdata file. A conditionally unbiased estimator of the frequency table in the original file is then given by

$$\hat{\boldsymbol{T}}_\xi = \left(\mathbf{P}^{-1}\right)^t \boldsymbol{T}_X. \tag{3}$$

This can be extended to two-dimensional frequency tables, by vectorizing such tables. The corresponding PRAM-matrix is then given by the Kronecker product of the PRAM-matrices of the individual dimensions. Alternatively, one could use the two-dimensional frequency tables $\mathbf{T}_{\xi\eta}$ for the original data and $\mathbf{T}_{XY}$ for the perturbed data directly in matrix notation:

$$\hat{\mathbf{T}}_{\xi\eta} = \left(\mathbf{P}_X^{-1}\right)^t \mathbf{T}_{XY} \mathbf{P}_Y^{-1}. \tag{4}$$

3 Measure of Disclosure Control

In this section we will define the measure we used to specify the effects of the different PRAM-matrices on the level of (statistical) disclosure control.

A frequently used rule to determine the safety of microdata files is the so called threshold rule: whenever a certain combination of scores on identifying variables occurs less than a certain threshold, that combination is considered to be unsafe. As an example consider the case that the combination of gender, occupation and age is to be checked for the threshold rule. Moreover, assume that the threshold is chosen to be 50. Then, if only 43 female surgeons of age 57 exist in the population, each record that corresponds to a female surgeon of age 57 is considered to be an unsafe record. Even though the threshold rule is defined in terms of population frequencies, in practice one often only has the sample file at hand. In that case the rule is usually applied to that sample file, with an appropriately adjusted threshold.

In case of using PRAM as an SDC-method this rule does not make any sense: since the perturbed file is the result of a probabilistic experiment, the unsafe records would vary over each realization. To deal with this problem, an alternative approach was suggested in [14]. In that approach, the disclosure risk is considered, i.e., the probability that given a score k in the perturbed file, the original score was k as well. By Bayes rule this can be calculated using

$$R_{\mathrm{PRAM}}(k) = \mathbb{P}(\xi = k \mid X = k) = \frac{\mathbb{P}(X = k \mid \xi = k)\mathbb{P}(\xi = k)}{\sum_{l=1}^{K} \mathbb{P}(X = k \mid \xi = l)\mathbb{P}(\xi = l)}. \quad (5)$$

Assuming that PRAM is applied to (the combination of) variable(s) ξ and using the appropriate notation, one could estimate this by

$$\hat{R}_{\mathrm{PRAM}}(k) = \frac{p_{kk} T_{\xi}(k)}{\sum_{l=1}^{K} p_{lk} T_{\xi}(l)}. \quad (6)$$

Note that we used $T_{\xi}(k)/n$ as an estimate of $\mathbb{P}(\xi = k)$, where n is the size of the original microdatafile.

In order to link this PRAM-risk to the traditional threshold rule, we suggest to use the following definition: a record is considered to be safe, whenever

$$\hat{R}_{\mathrm{PRAM}}(k) \leq \frac{T_{\xi}(k)}{\tau}, \quad (7)$$

where τ is the threshold used in the threshold rule for the original microdata file. Note that a safe record according to the original threshold rule applied to the original file, will be considered to be safe according to this rule as well. Hence, we need to verify this inequality for unsafe combinations only. Moreover, the number of unsafe records according to (7) only depends on the PRAM-matrix used and the original frequencies, i.e., is independent of the realization.

4 Measures of Information Loss

In this section we will briefly define the measures of information loss we will use in our experiments.

4.1 Entropy Based Information Loss

In [12] two measures of information loss, based on entropy arguments were introduced, EBIL and IL. The major difference between the two is that the latter measure makes use of both the original file and the perturbed file.

We can write these measures in the following way:

$$\mathrm{EBIL}(\mathbf{P}, \mathcal{G}) = -\sum_{l=1}^{K} \sum_{k=1}^{K} T_X(l) p_{lk}^{\leftarrow} \log p_{lk}^{\leftarrow} \quad (8)$$

$$\mathrm{IL}(\mathbf{P}, \mathcal{F}, \mathcal{G}) = -\sum_{l=1}^{K} \sum_{k=1}^{K} T_{\xi,X}(k,l) \log p_{lk}^{\leftarrow} \quad (9)$$

where $T_{\xi,X}(k,l)$ denotes the number of records with score $\xi = k$ in the original file $\mathcal{F}$ and $X = l$ in the perturbed file $\mathcal{G}$ and $p^{\leftarrow}_{lk} = \mathbb{P}(\xi = k \mid X = l)$. Since intuitively $T_{\xi,X}(k,l)$ should be close to $T_X(l)p^{\leftarrow}_{lk}$, we see that EBIL and IL will not differ much whenever the number of records is large enough, relative to the number of categories K.

Using similar arguments as in the derivation of the estimator of the PRAM-risk, the probabilities $p^{\leftarrow}_{lk}$ can be estimated by

$$\hat{p}^{\leftarrow}_{lk} = \frac{p_{kl}T_\xi(k)}{\sum_{m=1}^{K} p_{ml}T_\xi(m)}. \tag{10}$$

4.2 Frequency Table Based Information Loss

Often frequency tables are calculated for certain (crossings of) variables, as a first step in investigating a microdata file. Applying PRAM obviously effects these frequency tables, whenever one of the variables to which PRAM is applied is part of such a frequency table. Therefore, some measures of information loss will be defined, based on comparison of the original frequency tables with the estimated frequency tables, using an estimate that corrects for the fact that PRAM has been applied.

The first measure is the median of the relative differences between the counts in the table $\boldsymbol{T}_\xi$ based on the original file and the counts in the estimate $\hat{\boldsymbol{T}}_\xi$ based on the perturbed file:

$$\mathrm{RD}_d = \mathrm{Median}\left\{ \left| \frac{T_\xi(k) - \hat{T}_\xi(k)}{T_\xi(k)} \right|, \quad k = 1, \ldots, K \right\}, \tag{11}$$

where d denotes the dimension of the frequency table. In this paper we will only consider $d = 1, 2$. We used the median as a summarising measure of the information loss, since small cell counts and empty cells may lead to extreme situations when adjusting for PRAM.

Note that the relative difference is infinite whenever $T_\xi(k) = 0$ and $\hat{T}_\xi(k) \neq 0$. In our experiments, this only occurred in case of two dimensional tables, with large numbers of categories for both variables. Hence, for $d = 2$, we will additionally calculate the maximum relative difference mRD_d over all finite relative differences and count the number of occurrences of infinity.

Another way to measure information loss, is to use the additional variance introduced by applying PRAM, when estimating one-dimensional frequency tables, i.e., the variance of the estimator (3). Obviously, the conditional variance-covariance matrix of $\hat{\boldsymbol{T}}_\xi$ in equation (3) is given by

$$\not{\Sigma}_{\hat{\boldsymbol{T}}_\xi} = \mathrm{Var}(\hat{\boldsymbol{T}}_\xi \mid \xi) = \mathrm{Var}\left((\mathbf{P}^{-1})^t \boldsymbol{T}_X \mid \xi\right) = \left(\mathbf{P}^{-1}\right)^t \mathrm{Var}(\boldsymbol{T}_X \mid \xi)\mathbf{P}^{-1}. \tag{12}$$

We will use the formulas to calculate $\mathrm{Var}(\boldsymbol{T}_X \mid \xi)$ as given in [2]. To obtain a single figure as a measure of information loss, we will use the median of the coefficients of variation of the categories of the one-dimensional frequency table. I.e., we will use

$$\mathrm{CV} = \mathrm{Median}\left\{ \frac{\sqrt{\rlap{\,\,|}\Sigma_{\hat{\boldsymbol{T}}_\xi}(k,k)}}{T_\xi(k)}, \quad k = 1, \ldots, K \right\}. \tag{13}$$

Additionally, we will calculate the maximum coefficient of variation mCV over the K categories. In our experiments we have that $T_\xi(k) > 0$ for all categories k of all one-dimensional variables ξ, i.e., the coefficients of variation we consider, are all finite.

4.3 Linear Regression Based Information Loss

A second type of statistical analysis that is often used to explore a microdata file, is linear regression. Since PRAM effects categorical variables, a way to measure the loss of information, is to consider a linear regression on a categorical variable and to compare the regression coefficients estimated using the original file with those estimated using the perturbed file.

In this paper we will consider a linear regression model, with income as the dependent variable and a perturbed variable as explanatory variable. I.e., we will use the model

$$Y = \mathbb{E}\left(\sum_{k=1}^{K} \beta_k \delta(k) \right), \tag{14}$$

with Y the dependent variable income and $\delta(k)$ a dummy variable corresponding to the k-th category of variable ξ on which PRAM is applied. The regression coefficients $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_K)^t$ are estimated, based on the original microdata file, by

$$\boldsymbol{\beta} = \left[\mathrm{diag}\left(T_\xi(1), \ldots, T_\xi(K)\right)\right]^{-1} \boldsymbol{T}_\xi^y, \tag{15}$$

where $T_\xi^y(k) = \sum_{r \in \mathcal{F}} Y_r \delta_{\xi,r}(k)$, the sum of the response on income over all records with score $\xi = k$. When PRAM is applied to ξ, the regression coefficients β_k can be estimated using

$$\tilde{\boldsymbol{\beta}} = \left[\mathrm{diag}\left(\hat{T}_\xi(1), \ldots, \hat{T}_\xi(K)\right)\right]^{-1} \left(\mathbf{P}^{-1}\right)^t \boldsymbol{T}_X^y, \tag{16}$$

where $\hat{\boldsymbol{T}}_\xi$ is given in (3) and $\boldsymbol{T}_X^y$ is the analogous of $\boldsymbol{T}_\xi^y$ based on the perturbed file. The measure for the loss of information is then given by

$$\mathrm{LRD} = \mathrm{Median}\left\{ \left| \frac{\beta_k - \tilde{\beta}_k}{\beta_k} \right|, \quad k = 1, \ldots, K \right\}. \tag{17}$$

Additionally, we will calculate the maximum relative difference mLRD over the K regression coefficients.

5 The Empirical Results

In our experiments we used one microdata file of 6 237 468 records, representing a complete population and containing the categorical variables Gender (with 2 categories), Marital status (with 8 categories of which one is 'unknown'), Year of birth (with 89 categories), Place of Residence (with 130 categories of which one is 'unknown') and the continuous variable Income.

To check the effect of the different PRAM-matrices on the disclosure control, we will use the notion of unsafe records as given in (7), with $\tau = 100$. We will check two instances of combinations of identifying variables: 'Place of Residence × Marital status × Gender' (RMG) and 'Place of Residence × Marital status × Year of birth' (RMY). RMG consists of 1 806 combined categories (excluding the codes 'unknown'), of which 535 occur less than 100 times in the original microdata file (i.e., are rare combinations), whereas RMY consists of 80 367 combinations, with 20 072 rare occurrences.

5.1 Univariate PRAM

Firstly, we will apply PRAM to one categorical variable at a time. Appendix A shows the PRAM-matrices that we used for each categorical variable.

To measure the effect on disclosure control, we will count the number of unsafe combinations, as defined in (7), that will be left after applying each PRAM-matrix. Obviously, in case of applying PRAM to the variable Gender, we will not consider RMY, since the number of unsafe combinations in RMY will not be changed in that case. Similarly, we won't consider RMG in case of applying PRAM to Year of birth. In Tables 1 and 2 the results for the different PRAM-matrices are given.

The results on applying PRAM to Gender show that the number of unsafe combinations left after applying PRAM, increases with the transition probability p_{kk}. Indeed, since a large value of p_{kk} yields a high probability that an observed score equals the original score, this is what one would expect. The same effect is apparent comparing M2 with M3, M5 with M6, and Y1 with Y4 and Y8.

In most cases, increasing the number of nonzero elements in the PRAM matrix decreases the number of unsafe combinations. Except in case of M3, M4 and M6, where the number of unsafe combinations in RMY increases. Moreover, in case of fully filled matrices, the exact distribution of the probability mass over the off-diagonal elements does not seem to matter much. See e.g., R1 and R2: they only differ in the distribution of the mass over the off-diagonal elements within each block, whereas the number of unsafe combinations is virtually the same.

In Tables 3–6 the results concerning the measures of loss of information as given in Section 4 are given. In the columns marked '# Inf', the number of infinite relative differences is shown. The results showed that, indeed, the measures EBIL and IL did not differ very much (maximum difference of 0.15%). Therefore, we will only state one of them (IL) in the following tables.

Table 1. Number of unsafe combinations after applying G- and R-matrices

	RMG-unsafe
Matrix	$\tau = 100$
G1	464
G2	472
G3	481
G4	489
G5	506
G6	522
G7	528

	RMG-unsafe	RMY-unsafe
Matrix	$\tau = 100$	$\tau = 100$
R1	482	18 797
R2	480	18 780

Table 2. Number of unsafe combinations after applying Y- and M-matrices

	RMY-unsafe
Matrix	$\tau = 100$
Y1	18 277
Y2	18 236
Y3	17 844
Y4	19 086
Y5	19 054
Y6	17 907
Y7	19 295
Y8	19 309
Y9	17 122
Y10	17 116
Y11	16 651
Y12	17 971
Y13	16 642
Y14	17 971
Y15	17 971

	RMG-unsafe	RMY-unsafe
Matrix	$\tau = 100$	$\tau = 100$
M1	236	12 821
M2	52	11 628
M3	136	16 475
M4	130	16 951
M5	4	16 531
M6	9	17 634
M7	432	16 124
M8	433	15 805
M9	432	15 805

From the results it is also clear that the stated median relative differences are quite small. However, very large maxima are found as well. These extreme values are linked with cells with very small original frequency counts: for these cells a small absolute difference can be a large relative difference.

To put the number of infinite relative differences shown in the results into perspective, the number of empty cells in the frequency tables concerned, are 3 in $G \times Y$, 232 in $Y \times M$ and 2 547 in $Y \times R$.

Increasing the number of nonzero elements in the PRAM matrix, does not have a clear effect on the measures of loss of information: in some instances of the PRAM-matrices, the loss of information increases, whereas in other cases it decreases for the same measure of loss of information. Moreover, using one instance of a PRAM-matrix, the effect on the different measures is not the same either.

Table 3. Measures of loss of information for applying PRAM to Gender

				$G \times M$		$G \times Y$						
Matrix	IL	RD_1 (%)	mRD_1 (%)	RD_2 (%)	mRD_2 (%)	RD_2 (%)	mRD_2 (%)	# Inf	CV(%)	mCV(%)	LRD(%)	mLRD(%)
G1	1 824 380	0.06	0.07	4.91	411	5.46	667	3	0.41	0.48	0.39	0.58
G2	1 784 786	0.23	0.26	0.71	78	2.00	333	3	0.20	0.24	0.15	0.29
G3	1 717 559	0.14	0.16	0.91	6	1.40	225	3	0.13	0.15	0.04	0.06
G4	1 620 977	0.16	0.18	0.54	22	1.13	213	3	0.10	0.11	0.07	0.12
G5	1 329 907	0.11	0.13	0.23	33	0.56	133	3	0.06	0.06	0.06	0.07
G6	866 217	0.02	0.02	0.57	19	0.42	119	3	0.03	0.04	0.00	0.01
G7	529 517	0.00	0.00	0.16	4	0.21	34	3	0.02	0.02	0.01	0.01

Table 4. Measures of loss of information for applying PRAM to Marital status

				$M \times G$		$M \times Y$						
Matrix	IL	RD_1 (%)	mRD_1 (%)	RD_2 (%)	mRD_2 (%)	RD_2 (%)	mRD_2 (%)	# Inf	CV(%)	mCV(%)	LRD(%)	mLRD(%)
M1	1 716 579	1.29	118	0.55	1 965	37.31	50 542	204	2.50	1 975	0.91	3 912
M2	1 714 697	2.33	3 319	0.52	5 016	53.11	18 013	204	2.01	1 762	0.36	13
M3	1 114 651	0.59	318	0.72	851	19.45	11 142	204	0.96	683	0.50	401
M4	1 048 860	0.52	84	0.57	2 169	21.40	21 438	204	0.85	2 068	0.56	3 876
M5	942 280	0.25	1 142	0.23	1 984	22.36	14 464	204	0.97	3 230	0.58	407
M6	791 340	0.18	3 477	0.54	6 151	21.20	18 589	204	0.80	2 675	0.53	70
M7	1 260 776	1.04	302	0.91	322	1.83	25 482	83	0.97	329	0.24	13
M8	962 267	0.08	79	0.12	838	1.55	21 665	84	0.79	776	0.66	5 588
M9	961 093	0.29	52	0.73	511	1.93	20 128	83	0.99	329	0.69	143

Table 5. Measures of loss of information for applying PRAM to Year of birth

				$Y \times G$			$Y \times M$						
Matrix	IL	RD$_1$ (%)	mRD$_1$ (%)	RD$_2$ (%)	mRD$_2$ (%)	# Inf	RD$_2$ (%)	mRD$_2$ (%)	# Inf	CV(%)	mCV(%)	LRD(%)	mLRD(%)
Y1	2 563 469	0.82	545	1.45	339	3	11.36	818	148	1.43	421	0.94	2 740
Y2	3 301 840	0.78	91	1.12	277	3	10.12	931	149	0.99	241	0.92	519
Y3	4 405 751	0.70	904	0.87	2 217	3	18.76	889	212	0.92	824	0.82	624
Y4	1 985 020	0.45	278	0.60	130	3	3.33	348	105	0.63	164	0.59	276
Y5	2 446 654	0.40	109	0.64	112	3	4.58	321	131	0.58	133	0.48	98
Y6	3 684 799	0.81	1 211	0.78	2 548	3	23.43	5 178	232	0.67	1 033	0.84	8 291
Y7	1 350 261	0.41	163	0.65	100	2	1.89	343	79	0.41	102	0.35	68
Y8	1 725 785	0.33	69	0.63	154	3	2.45	269	95	0.51	129	0.50	149
Y9	3 592 735	0.55	13 356	0.65	10 232	3	29.30	11 590	232	0.70	17 791	0.21	5 028
Y10	3 582 004	0.51	5 955	0.57	12 419	3	35.39	18 499	232	0.70	17 894	0.96	12 957
Y11	3 856 311	0.58	1 085	0.77	957	3	6.40	2 944	136	0.81	1 202	0.78	950
Y12	3 752 656	0.39	190	0.79	293	3	6.35	3 659	136	0.67	222	0.72	232
Y13	3 853 803	0.36	1 450	0.69	1 182	3	5.51	1 819	136	0.82	1 231	1.16	680
Y14	3 754 083	0.56	1 943	0.64	1 531	3	5.57	5 067	136	0.67	1 202	0.55	503
Y15	3 751 564	0.45	454	0.86	839	3	5.64	4 765	136	0.67	1 231	1.34	949

Table 6. Measures of loss of information for applying PRAM to Place of Residence

				$R \times G$		$R \times Y$						
Matrix	IL	RD$_1$ (%)	mRD$_1$ (%)	RD$_2$ (%)	mRD$_2$ (%)	RD$_2$ (%)	mRD$_2$ (%)	# Inf	CV(%)	mCV(%)	LRD(%)	mLRD(%)
R1	2 389 452	0.25	3	0.46	7	3.94	670	1610	0.42	5.64	0.36	1.91
R2	2 369 788	0.33	6	0.43	10	3.92	739	1610	0.42	5.81	0.37	3.87

5.2 Multivariate PRAM

In order to observe the effect of applying PRAM to several categorical variables at the same time, we applied certain combinations of the previously mentioned PRAM-matrices simultaneously. The combinations we used are given in Appendix A.

In Table 7 the number of unsafe combinations after applying PRAM to three categorical variables at the same time are given. Again, only two combinations of identifying variables are considered: RMG and RMY. On average we see that applying PRAM in a multivariate way, the number of unsafe combinations that is left, is smaller compared to the univariate applications. However, we would like to stress the notion that only comparing the unsafe combinations is not fair: this should always be related to the amount of information that is lost.

Table 7. Number of unsafe combinations after applying PRAM

	RMG	RMY
Name	$\tau = 100$	$\tau = 100$
PRAM1	429	13 365
PRAM2	100	12 442
PRAM3	22	9 189
PRAM4	127	14 046
PRAM5	427	11 656
PRAM6	429	13 135
PRAM7	0	14 102

We will not consider all measures of loss of information in case of multivariate PRAM, but only state the results concerning the measure IL, and the results on two-dimensional relative differences in case PRAM is applied to both variables. I.e., in case of PRAM1 (PRAM applied to Year of birth, Gender and Marital Status), we will consider the two-dimensional frequency tables $Y \times G$, $Y \times M$ and $G \times M$. Tables 8 and 9 show the numerical results.

As we expected, the loss of information according to IL is larger compared to the univariate results. This is due to the fact that there are many more categories to consider. Since the definition of IL consist of sums of terms including $\log p^{\leftarrow}_{lk}$, this yields a large value for this measure.

If we want to compare univariate PRAM with multivariate PRAM, we will have to take both the level of disclosure control as well as the amount of information that is lost into account. We expect that applying multivariate PRAM with the same level of information loss as a univariate application, will yield a higher level of disclosure control.

The closest values for IL in case of univariate and multivariate PRAM are the ones corresponding to Y3 and PRAM4. If we then look at the number of unsafe RMY-combinations, we see that PRAM4 has 3 798 unsafe combinations less (about 21%), even though the IL value is 12% larger than in case of Y3. I.e., even though the loss of information is larger, the level of disclosure control

Table 8. Measures of loss of information for multivariate PRAM, part 1

		$M \times G$	
Name	IL	RD$_2$ (%)	mRD$_2$ (%)
PRAM1	5 825 419	2.11	672
PRAM2	6 241 220	3.40	3 725
PRAM3	6 683 416	4.34	25 977
PRAM4	4 927 149	0.70	4 088
PRAM5	5 932 505	2.90	2 036
PRAM6	5 197 464	1.39	1 019
PRAM7	5 751 979	5.70	78 263

Table 9. Measures of loss of information for multivariate PRAM, part 2

		$Y \times M$			$Y \times G$	
Name	RD$_2$ (%)	mRD$_2$ (%)	# Inf	RD$_2$ (%)	mRD$_2$ (%)	# Inf
PRAM1	62.99	39 439	232	1.44	1 841	3
PRAM2	126.10	16 464	232	3.69	2 824	3
PRAM3	198.05	27 607	232	2.72	4 906	3
PRAM4	97.18	17 503	232	0.77	310	3
PRAM5	18.00	22 939	148	1.73	2 086	3
PRAM6	14.67	17 967	148	0.99	962	3
PRAM7	280.55	28 459	232	7.22	225 259	3

is higher as well. Moreover, since the univariate application with Y3 has no effect on the unsafe combinations in RMG but the multivariate application with PRAM4 does have, PRAM4 outperforms Y3 in that sense as well. Similarly, considering the information loss according to the relative differences, the overall loss of information is larger for the multivariate cases. This is not surprising: both variables in the frequency tables have been perturbed in the multivariate setting, whereas in the univariate setting only one of the spanning variables is perturbed. Hence, more cells are affected more seriously.

However, if we take, e.g., the table $G \times M$ in case of G2 and PRAM4, the median relative differences are 0.71 and 0.70 respectively, whereas the number of unsafe combinations in RMG is reduced from 472 for G2 to 127 for PRAM4. Additionally, in case of PRAM4 the number of unsafe combinations in RMY is reduced as well (from 20 072 to 14 046), whereas in case of G2 there is no effect on the number of unsafe combination in RMY. So, with more or less the same loss of information the multivariate case has a much higher level of disclosure control.

6 Summary and Conclusions

PRAM is a method to deal with disclosure control when disseminating microdata. This method was introduced in 1997, but has not yet been applied extensively. This is partly due to the fact that there is little knowledge available on the effect of PRAM on disclosure control or on the loss of information it induces.

The method is defined in terms of transition probabilities, summarized in a PRAM-matrix. In this paper we investigated the effect of different distributions of the transition probabilities, on the level of disclosure control as well as on the amount of information that is lost when applying PRAM. Several instances of PRAM-matrices have been applied to a specific microdatafile, both in a univariate as well as a multivariate setting. Several measures of loss of information have been calculated along with a measure for the level of disclosure control. The different instances resulted in different effects. In most cases, increasing the number of non-zero elements resulted in a decrease of unsafe combinations. However, its effect on the measures of loss of information was not unambiguous: some measures gave rise to an increase of loss of information, whereas others yielded a decrease. This indicates that it might be desirable to let the choice of PRAM matrix (or matrices) depend on the intended use of the microdatafile. To compare the results of the univariate and the multivariate application of PRAM, we should take into account both the effect on the level of disclosure control as well as on the loss of information. Indeed, one should only compare situations with either a comparable level of disclosure control or a comparable amount of loss of information. The results indicate that it seems possible to achieve the same level of disclosure control, with a lower loss of information, when applying PRAM in a multivariate way. Or, equivalently, to achieve the same amount of loss of information, with a higher level of disclosure control.

In our experiments, we used block matrices, with equal diagonal elements within each block. An obvious alternative would be to allow for a variation in the diagonal elements within each block. These diagonal elements might be chosen depending on the disclosure risk associated with that category. However, since that risk is related to combinations of categories of several variables, this becomes quite complicated, especially when applying PRAM in a multivariate way. This is a topic for further research.

References

1. Kooiman, P., Willenborg, L., Gouweleeuw, J. (1997): A method for disclosure limitation of microdata. Research paper 9705, Statistics Netherlands, Voorburg.
2. Gouweleeuw, J., Kooiman, P., Willenborg, L., de Wolf, P.P. (1998): Post randomisation for statistical disclosure control: Theory and implementation. Journal of Official Statistics **14**(4) 463–478.
3. de Wolf, P.P., Gouweleeuw, J., Kooiman, P., Willenborg, L. (1998): Reflections on pram. In: Statistical Data Protection, Luxembourg, Office for Official Publications of the European Communities 337–349.
4. van den Hout, A. (1999): The analysis of data perturbed by pram. Delft Univsersity Press, Delft.
5. van den Hout, A., van der Heijden, P.G.M. (2002): Randomized response, statistical disclosure control and misclassification: a review. International Statistical Review **70**(2) 269–288.
6. Ronning, G., Rosemann, M., Strotmann, H. (2004): Estimation of the probit model using anonymized micro data. Paper prepared for the 'European Conference on Quality and Methodology in Official Statistics (Q2004)', Mainz, 24–26 May 2004.

7. Warner, S.L. (1965): Randomized response: a survey technique for eliminating evasive answer bias. Journal of the American Statistical Association **60** 63–69.
8. Chen, T.T. (1979): Analysis of randomized response as purposively misclassified data. In: Proceedings of the section on survey research methods, American Statistical Association 158–163.
9. Press, S.J. (1968): Estimating from misclassified data. Journal of the American Statistical Association **63** 123–133.
10. Kuha, J., Skinner, C. (1997): Categorical data analysis and misclassification. In Lyberg, Biemer, Collins, de Leeuw, Dippo, Schwarz, Trewin, eds.: Survey measurement and process quality. Wiley, New York 633–670.
11. Fienberg, S.E. (1994): Conflict between the needs for access to statistical information and demands for confidentiality. Journal of Official Statistics **10**(2) 115–132.
12. Domingo-Ferrer, J., Torra, V. (2001a): Disclosure control methods and information loss for microdata. In Doyle, P., Lane, J., Theeuwes, J., Zayatz, L., eds.: Confidentiality, Disclosure and Data Acces: Theory and Practical Applications for Statistical Agencies. Elsevier, North-Holland 91–110.
13. Domingo-Ferrer, J., Torra, V. (2001b): A quantitative comparison of disclosure control methods for microdata. In Doyle, P., Lane, J., Theeuwes, J., Zayatz, L., eds.: Confidentiality, Disclosure and Data Acces: Theory and Practical Applications for Statistical Agencies. Elsevier, North-Holland 111–133.
14. Rienstra, M. (2003): Aanzet tot een nieuwe beveiligingsregel bij gebruik van pram. Internal report (in Dutch) 283-03-SOO, Statistics Netherlands, Voorburg.

Appendix A: PRAM-Matrices

In this appendix we first define the notation that we used to describe the PRAM-matrices of our experiments.

- Band matrices $n\mathbf{B}(p;b)$, with p the value of the diagonal elements, b the bandwidth (i.e., the number of entries p_{kl} with $|k-l|<b$) and n the size of the square matrix. Note that we should choose $b \leq n-1$. The probability mass $(1-p_{kk})$ is distributed equally over the off-diagonal elements in the band. E.g., a $4\mathbf{B}(0.6;2)$ matrix would look like

$$\begin{pmatrix} 0.6 & 0.4 & 0 & 0 \\ 0.2 & 0.6 & 0.2 & 0 \\ 0 & 0.2 & 0.6 & 0.2 \\ 0 & 0 & 0.4 & 0.6 \end{pmatrix}$$

- Fully filled matrices, with equal off-diagonal elements, denoted by $n\mathbf{E}(p)$, with n the size of the square matrix and p the value of the diagonal elements. E.g., a $3\mathbf{E}(0.8)$ matrix would look like

$$\begin{pmatrix} 0.8 & 0.1 & 0.1 \\ 0.1 & 0.8 & 0.1 \\ 0.1 & 0.1 & 0.8 \end{pmatrix}$$

Table 10. PRAM matrices for Gender (G) and Place of Residence (R)

Name	Description
G1	$2\mathbf{E}(0.55)$
G2	$2\mathbf{E}(0.6)$
G3	$2\mathbf{E}(0.65)$
G4	$2\mathbf{E}(0.7)$
G5	$2\mathbf{E}(0.8)$
G6	$2\mathbf{E}(0.9)$
G7	$2\mathbf{E}(0.95)$

Name	Description
R1	$\text{Block}(14; \mathbb{1}; 7\mathbf{E}(0.8); 8\mathbf{E}(0.8); 4\mathbf{E}(0.8); 10\mathbf{E}(0.8);$ $2\mathbf{E}(0.8); 15\mathbf{E}(0.8); 6\mathbf{E}(0.8); 17\mathbf{E}(0.8);$ $24\mathbf{E}(0.8); 6\mathbf{E}(0.8); 18\mathbf{E}(0.8); 11\mathbf{E}(0.8); \mathbb{1})$
R2	$\text{Block}(14; \mathbb{1}; 7\mathbf{F}(0.8); 8\mathbf{F}(0.8); 4\mathbf{F}(0.8); 10\mathbf{F}(0.8);$ $2\mathbf{F}(0.8); 15\mathbf{F}(0.8); 6\mathbf{F}(0.8); 17\mathbf{F}(0.8);$ $24\mathbf{F}(0.8); 6\mathbf{F}(0.8); 18\mathbf{F}(0.8); 11\mathbf{F}(0.8); \mathbb{1})$

Table 11. PRAM matrices for Year of birth (Y) and Marital status (M)

Name	Description
Y1	$89\mathbf{B}(0.6; 2)$
Y2	$89\mathbf{B}(0.6; 3)$
Y3	$89\mathbf{B}(0.6; 7)$
Y4	$89\mathbf{B}(0.75; 2)$
Y5	$89\mathbf{B}(0.75; 3)$
Y6	$89\mathbf{B}(0.75; 21)$
Y7	$89\mathbf{B}(0.8; 1\frac{1}{2})^*$
Y8	$89\mathbf{B}(0.8; 2)$
Y9	$89\mathbf{E}(0.75)$
Y10	$89\mathbf{F}(0.75)$
Y11	$\text{Block}(3; 24\mathbf{E}(0.6); 61\mathbf{E}(0.75); 4\mathbf{E}(0.6))$
Y12	$\text{Block}(3; 24\mathbf{B}(0.6; 5); 61\mathbf{B}(0.75; 21); 4\mathbf{B}(0.6; 2))$
Y13	$\text{Block}(3; 24\mathbf{F}(0.6); 61\mathbf{F}(0.75); 4\mathbf{F}(0.6))$
Y14	$\text{Block}(3; 24\mathbf{E}(0.6); 61\mathbf{B}(0.75; 21); 4\mathbf{E}(0.6))$
Y15	$\text{Block}(3; 24\mathbf{F}(0.6); 61\mathbf{B}(0.75; 21); 4\mathbf{F}(0.6))$

Name	Description
M1	$\text{Block}(2; \mathbb{1}; 7B(0.6; 3))$
M2	$\text{Block}(2; \mathbb{1}; 7\mathbf{B}(0.6; 4))$
M3	$\text{Block}(2; \mathbb{1}; 7\mathbf{B}(0.8; 4))$
M4	$\text{Block}(2; \mathbb{1}; 7\mathbf{B}(0.8; 5))$
M5	$\text{Block}(2; \mathbb{1}; 7\mathbf{F}(0.75))$
M6	$\text{Block}(2; \mathbb{1}; 7\mathbf{F}(0.8))$
M7	$\text{Block}(3; \mathbb{1}; 4\mathbf{E}(0.8); 3\mathbf{E}(0.6))$
M8	$\text{Block}(3; \mathbb{1}; 4\mathbf{F}(0.8); 3\mathbf{F}(0.6))$
M9	$\text{Block}(3; \mathbb{1}; 4\mathbf{F}(0.8); 3\mathbf{E}(0.6))$

* Non-zero elements at p_{kk}, p_{kk+1}, $k = 1, \ldots, K-1$, p_{KK} and p_{K-1K}.

- Fully filled matrices, with the off-diagonal elements depending on the corresponding frequencies in the original microdata file, denoted by $n\mathbf{F}(p)$, with n the size of of the square matrix and p the value of the diagonal elements. The off-diagonal elements are determined using a method defined in [14] (note that this formula requires that $n \geq 3$):

$$p_{kl} = \frac{(1 - p_{kk})\left(\sum_{i=1}^{K} T_\xi(i) - T_\xi(k) - T_\xi(l)\right)}{(n-2)\left(\sum_{i=1}^{K} T_\xi(i) - T_\xi(k)\right)} \qquad (18)$$

E.g., with $\boldsymbol{T}_\xi = (5576, 24, 632)^t$, the matrix $3\mathbf{F}(0.6)$ would look like

$$\begin{pmatrix} 0.6000 & 0.3854 & 0.0146 \\ 0.0407 & 0.6000 & 0.3593 \\ 0.0017 & 0.3983 & 0.6000 \end{pmatrix}$$

Table 12. Combinations of PRAM-matrices

Name	Combination		
PRAM1	Y6	G5	M7
PRAM2	Y11	G2	M3
PRAM3	Y14	G2	M2
PRAM4	Y12	G7	M4
PRAM5	Y13	G4	M9
PRAM6	Y15	G6	M8
PRAM7	Y10	G1	M6

Note that $1\mathbf{E}(1)$ is a special case that we will denote by $\mathbb{1}$. The three basic types can be combined into block-matrices. We will denote these block matrices by $\mathrm{Block}(m; \mathrm{type}_1; \cdots; \mathrm{type}_m)$, with m the number of blocks and following m the matrix type for each block. Note that, using this construction, the diagonal elements of a PRAM-matrix will be constant within each block, but may vary between the blocks.

Finally, tables 10, 11 and 12 show the PRAM matrices that we used in our experiments, using the notation given above.

Distance Based Re-identification for Time Series, Analysis of Distances

Jordi Nin[1] and Vicenç Torra[2]

[1] DAMA-UPC, Computer Architecture Dept.
Universitat Politècnica de Catalunya,
Campus Nord UPC, C/Jordi Girona 1-3
08034 Barcelona, Catalonia, Spain
nin@ac.upc.edu
http://www.dama.upc.edu
[2] IIIA-CSIC
Campus UAB s/n
08193 Bellaterra Catalonia, Spain
vtorra@iiia.csic.es

Abstract. Record linkage is a technique for linking records from different files or databases that correspond to the same entity. Standard record linkage methods need the files to have some variables in common. Typically, variables are either numerical or categorical. These variables are the basis for permitting such linkage.

In this paper we study the problem when the files to link are formed by numerical time series instead of numerical variables. We study some extensions of distance base record linkage in order to take advantage of this kind of data.

Keywords: Re-identification algorithms, time series, privacy statistical databases, time series distances, record linkage.

1 Introduction

Everyday, thousands of data about specific entities, such as people or business, are stored in databases. The integration of two or more databases or files is of increasing importance, and difficulty, due to the growth of these stored data.

In the last years, many researchers have developed methods for schema and record matching [9]. One of them is *e.g.* record linkage [13,14], a technique for linking records of different files or databases that correspond to the same entity.

An important use of record linkage algorithms is for risk assessment in privacy preserving data mining (PPDM) [1] and statistical disclosure control (SDC) [12]. In this case, record linkage methods permit to evaluate whether a protection mechanism provides enough protection to providers of sensitive information (to know whether a protection method guarantees avoids the disclosure of sensitive information to data providers).

An increasing percentage of this stored information has an implicit or explicit time component. This is the case of *e.g.*, income or stock prices. Similarly, data

J. Domingo-Ferrer and L. Franconi (Eds.): PSD 2006, LNCS 4302, pp. 205–216, 2006.

accumulation through consecutive years (*e.g.*, economical data from companies or census data from individuals) can also be considered from this point of view. Standard record linkage algorithms have been designed for non-temporal variables and they need to solve some key questions for their application to time series as *e.g.* data normalization and distance selection.

In the setting of risk assessment for PPDM, we can consider the scenario of an attacker (a person who should have no access to the sensible information) that obtains different files from the same population but corresponding to the same information in different years. Joining these files the attacker would obtain a new file where variables are time series instead of non-temporal variables. So, the attacker can exploit the (explicit) time information of this new file to obtain sensible information.

In this paper we study the record linkage methods for this kind of scenario. More specifically, we consider that the files to link are defined in terms of numerical time series (variables are time series) instead of numerical non-temporal variables. We study some extensions of distance base record linkage in order to take advantage of this kind of data.

The structure of the paper is as follows. In section 2 we describe some of the preliminaries required in the rest of the paper. In particular, this section describes standard re-identification methods and some distance functions for time series. Then, in Section 3, we propose our method for time series re-identification, in this section we describe our method to normalize time series and our method for re-identifying time series. In Section 4 we describe the experiments done and the results obtained. The paper finishes with some conclusions and research lines for future work.

2 Preliminaries

2.1 Re-identification Methods

Two main approaches have been used for re-identification in the case of numerical and categorical variables. See [10,13,14] for more details:

Probabilistic Record Linkage: For each pair of records (a,b), where a is a record of file A and b is a record in file b, an index is computed using some conditional probabilities. Then, this index is used to classify each pair (a,b) as either a linked pair (LP) or a non-linked pair (NP).

Distance-based Record Linkage: Records of two files A and B are compared with respect to a given distance measure, and then each record in A is linked to the nearest record in B using such distance measure.

Our research on record linkage for time series follows the second approach. That is, it is a method extending distance-based record linkage.

2.2 Time Series

In this section we focus on numerical time series. Formally speaking, numerical time series are defined by pairs $\{(v_k, t_k)\}$ for $k = 1, \ldots, N$ where t_k corresponds to the temporal variable and v_k is the numerical variable that depends on time (dependent variable). Naturally, $t_{k+1} > t_k$. Income and sport statistics are examples of time series, as they depend on time.

We can define in the same way ordinal or categorical time series replacing v_k for a categorical or ordinal variable. Weather forecast (*e.g.* sunny, cloudy, raining) and restaurant category (*e.g.* one Michelin star, two Michelin star, three Michelin star) over time are examples of categorical and ordinal time series respectively.

In the literature we can find a great variety of time series distances measures. See [11,6,3] for more details.

In the remaining part of this section we describe the time series distances used in this paper.

Minkowski distance. The Minkowski distance, that is a generalization of the Euclidean distance, is defined as

$$d_{MK}(x,v) = \sqrt[q]{\sum_{k=1}^{N}(x_k - v_k)^q}$$

In the above definition, q is a positive integer. We can define a normalized version if the measured values are normalized via division by the maximum value in the time series.

Short time series distance. The short time series distance (STS distance), was defined by Möller-Levet et al. in [6]. This distance corresponds to the square root of the sum of the squared differences of the slopes, and is defined as follows:

$$d_{STS}(x,v) = \sqrt[2]{\sum_{k=1}^{N}\left(\frac{v_{k+1} - v_k}{t_{k+1} - t_k} - \frac{x_{k+1} - x_k}{t_{k+1} - t_k}\right)^2}$$

Two cross correlation based distance. Pearson's correlation coefficient is defined as

$$cc = \frac{\sum_{k=1}^{N}(x_k - \mu_x)(v_k - \mu_v)}{S_x S_v}$$

where μ is the mean of each time series and S is the standard deviation.

Two cross correlation based distances are defined below

$$d_{cc} = \left(\frac{1 - cc}{1 + cc}\right)^{\beta}$$

Dynamic Time Warping distance. Now, we review the Dynamic Time Warping algorithm, or DTW algorithm in short. Suppose we have two numerical time series x and v, of length n and m respectively.

$$\{(x_k, t_k)\} \text{ for } \text{k=1}, \ldots, n$$

$$\{(v_k, t_k)\} \text{ for } \text{k=1}, \ldots, m$$

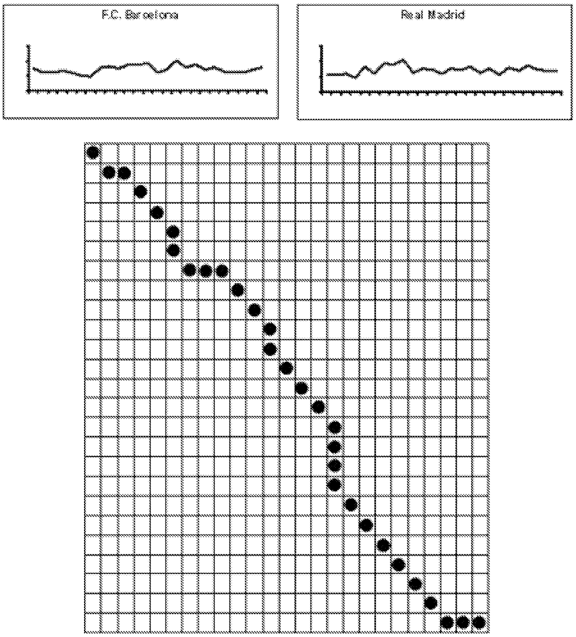


Fig. 1. An example of wrapping path between goals for of F.C. Barcelona and Real Madrid since 1982 to 2006

For aligning these two time series using the DTW algorithm, we construct a bi-dimensional $n \times m$ matrix where the element (i^{th},j^{th}) contains the distance between the two points x_i and v_j. To compute the distance between these two points, the Euclidean distance is often used (d(x_i ,v_j)=(($x_i - v_j$)^2)). Then, each matrix element (i,j) corresponds to the alignment between the points x_i and v_j.

A warping path, w, is a route from element (0,0) to element (n,m) formed by contiguous cells with some particular constraints. In general, the path represent a relation between x and v and has several constraints:

Boundary conditions: w_0=(0,0) and w_k=(n,m). A warping path requires starting and finishing in opposite diagonal corners of the matrix.

Continuity: Given w_k=(i,j) then w_{k+1}=(i',j'), where $i' - i \leq 1$ and $j' - j \leq 1$. This fits the allowable steps to adjacent cells including diagonally adjacent cells.

Monotonicity: Given w_k=(i,j) then w_{k+1}=(i',j'), where $i' - i \geq 0$ and $j' - j \geq 0$. This avoids cycles in the warping path.

There are many warping paths that satisfy the above restrictions, the warping paths grow exponentially with respect to their length, but we are interested only in the path which minimizes the following warping cost

$$DTW(x, v) = min \left(\frac{1}{K} \sqrt{\sum_{k=1}^{K} w_k} \right)^{\beta}$$

K is used to compensate the fact that warping paths may have different lengths. The path with minimum warping cost can be calculated very efficiently using dynamic programming. Euclidean distance (for time series) can be considered as a particular case of DTW when time series have the same length.

Example 1. An example of warping path is illustrated in Figure 1. The warping path represents the distance between the number of goals of F.C. Barcelona and Real Madrid. In this case, we obtain a DTW distance equal to 2021. This value represents the distance of an optimal alignment between goals for of F.C. Barcelona and Real Madrid. We understand an optimal alignment as the alignment that makes the smallest distance between the two teams.

3 Re-identification for Time Series

3.1 Time Series Normalization

It is usual to normalize data files before applying record linkage methods. This is so to avoid the scale problems of raw data. The following two alternatives are usually considered:

- Ranging: Raw Data are translated into the $[0, 1]$ interval using this expression $x' = \frac{(x - \min(v))}{(\max(v) - \min(v))}$ where x is the original value and $\max(v)$ and $\min(v)$ are the maximum and minimum values for the corresponding variable.
- Standardization: Raw data are normalized by translating mean equals zero and the standard deviation equals one: $x = \frac{(x - \mu_v)}{S_v}$ (where μ_v and S_v are, respectively, the mean and the standard deviation of the corresponding variable v).

This kind of pre-processing when applied independently for each component of the time series causes the lost of the temporal information of the series. For this reason, we apply another type of normalization using all the elements included into the time series. In our experiments we had used the following normalization

$$v'_k = \frac{(v_k - \mu_v)}{S_v}$$

where μ_v and S_v are the mean and the standard deviation of the elements of the corresponding time series.

Now we illustrate with a simple example (uses index prices for some food products) the impact of the normalization of the time series, comparing the normalization by component (each component treated as a variable) and the normalization of the series as a whole. As it is shown in the example, the normalization by component can distort completely the shape of the time series.

Table 1. Data extracted from Spanish National Statistics Institute

	Index of prices								
	1993	1994	1995	1996	1997	1998	1999	2000	2001
Bread	106,5	110,3	114,9	117,9	119,3	121	122,2	124,1	129
Oil	102,7	119,8	147,8	178,7	130,8	116,2	133,6	123,5	114,4
Fruits and vegetables	95,6	101,9	110,8	116,4	114,2	119	124,6	126,4	133,9
Potatoes	101,1	133,6	162,8	123,8	121,3	140,4	149,8	148,6	177,6

Table 2. Data normalized with the standard component-wise procedure

	Index of prices								
	1993	1994	1995	1996	1997	1998	1999	2000	2001
Bread	1,00	0,26	0,08	0,02	0,31	0,20	0,00	0,02	0,23
Oil	0,65	0,56	0,71	1,00	1,00	0,00	0,41	0,00	0,00
Fruits and vegetables	0,00	0,00	0,00	0,00	0,00	0,12	0,09	0,12	0,31
Potatoes	0,50	1,00	1,00	0,12	0,43	1,00	1,00	1,00	1,00

Example 2. Let us consider the price index of four different foods in nine years. We can observe in Table 1 the original raw values and their tendency in the period 1993 - 2001 and in Tables 2 and 3 the normalized data values after standard (component-wise) and time series (data altogether) normalization process, respectively.

Figure 2 shows that different normalizations produce different outcomes, and that the standard component-wise normalization causes important divergences on the tendency of the time series between the original time series and the normalized one. For example, in the case of the bread, when comparing charts (a) and (b), we observe that in the original data the bread price tendency was to increase every year but that after normalization the bread price has a decreasing tendency. This is a negative effect of normalization over the data.

To avoid this effect of component-wise normalization, we propose the use of specific normalization procedures for time serie: normalization of all the series.

3.2 Approach

As we have explained in the introduction, we consider in this paper the re-identification problem when the files include time series and not only simple variables. In this case the application of the methods outlined in Section 2.1 is problematic. To solve these situations, we extend distance-based record linkage to this kind of variables.

From now on, we will consider that we have two files A and B, and that these files have one or more time series. Then, if we want to apply distance-based record linkage methods we need to change traditional distance function to the ones for time series.

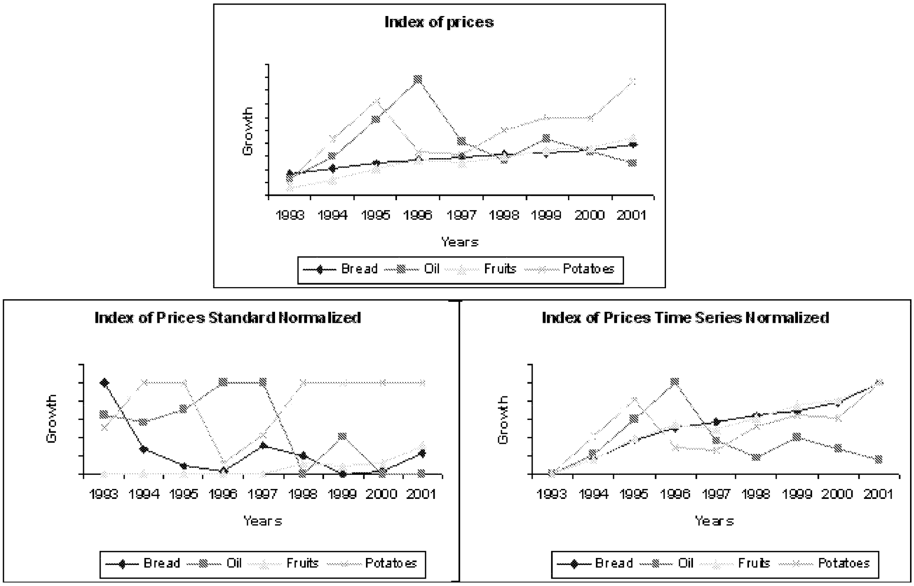


Fig. 2. Graphical representation of the effects of time series normalization, (a) Top: it represents the original data without normalization, (b) bottom left: it represents normalized data with independent normalization, (c) bottom right: represents normalized data with time series normalization

Table 3. Data normalized with the time series procedure

	Index of prices								
	1993	1994	1995	1996	1997	1998	1999	2000	2001
Bread	106,5	110,3	114,9	117,9	119,3	121	122,2	124,1	129
Oil	102,7	119,8	147,8	178,7	130,8	116,2	133,6	123,5	114,4
Fruits and vegetables	95,6	101,9	110,8	116,4	114,2	119	124,6	126,4	133,9
Potatoes	101,1	133,6	162,8	123,8	121,3	140,4	149,8	148,6	177,6

In short, our approach works as follows:

1. Apply time series normalization in both files.
2. Select a time series distance function D with a particular parameter p.
3. Apply record linkage to normalized files, with the selected time series distance.

4 Experiments

4.1 Data

In order to check our approach we have done some test with real data that can be obtained freely from different data sources. First we have downloaded

Table 4. Details of time series examples

	Forecasters	Ibex35	Soccer
Records	3003	35	176
Time series	1	2	8
Time series size	10	220	25
Record size	10	440	200
Series description	Financial information	opening prices Volume transactions	years FIFA points League position Goals for Goals against Matches win Matches dice Matches loose

from [5] a file (the so-called forecasters problem) with 3003 time series of different lengths between 14 and 64 elements, we have re-sampled all time series to 10 elements. Secondly, we have used the Stock Exchange information of the thirty five most important Spanish companies. These companies are ranked in the so-called Ibex35 stock market. We have downloaded the information about prices from June, 21Th 2005 to April, 28Th 2006 from [2]. And finally, we have downloaded information about all soccer teams of the nine most important European domestic leagues from [7]. As said, information about these three testbeds is publicly available. Data details are given in Table 4.

4.2 Time Series Masking

For applying our approach, we have protected the original data with a time series masking method based on microaggregation procedures described in [8]. We have applied this method with $k \in \{2, 3, 6, 9, 12\}$.

We have applied time series microaggregation method splitting the original time series in n masked series to check if our method improve its results when the number of time series grows.

Forecasters problem: We have split original time series in $n \in \{1, 2\}$ time series. So, in this case we have two different problems, one with one time series and other with two time series.

Ibex 35 problem: We have split original time series in $n \in \{2, 4, 20\}$ time series. So, in this case we have three different problems with $4, 8$ and 40 time series.

Soccer problem: We have not split original time series, in this case we have one problem with eight time series.

Once we have made protected files, we apply our time series DBRL method. We have tested four Minkowski distances with parameters q from the set $Q = \{2, 3, 4, 5\}$, four cross correlation based distance with parameter β from the set

Table 5. Results of forecasters problem

number of time series	1					2				
k	2	3	6	9	12	2	3	6	9	12
DBRL Standard	1180	850	421	283	215	348	253	207	174	223
Minkowski $q = 2$	1291	780	325	224	162	218	153	155	165	206
Minkowski $q = 3$	552	315	140	95	55	53	53	29	25	25
Minkowski $q = 4$	1240	750	300	205	159	176	130	131	149	192
Minkowski $q = 5$	633	394	174	112	65	59	46	20	23	23
Cross correlation $\beta = 1$	533	216	73	46	33	105	66	27	22	21
Cross correlation $\beta = 2$	0	1	0	0	0	0	0	0	0	0
Cross correlation $\beta = 3$	533	216	73	46	33	105	66	27	22	21
Cross correlation $\beta = 4$	0	1	0	0	0	0	0	0	0	0
STS	1000	531	185	106	78	282	165	133	125	80
DTW	1271	754	300	189	139	164	114	97	115	127

$\beta = \{1, 2, 3, 4\}$, short time series distance and Dynamic Time Warping distance in the forecasters test. In the case of DTW, the DBRL has not been computed for Ibex35 and Soccer time series as the computational time was very high.

4.3 Results

In this section we compare our method with standard DBRL (all variables in the series treated separately for both normalization and computing the distance), if we observe the results in Tables 5,6 and 7 we observe that in most of the cases we can find a time series distance, like STS distance, that equals or improves the results obtained from standard DBRL. This result is strengthened because the standard DBRL with the standard component-wise normalization might change the shape of the series (as shown in Example 2).

In the case of the soccer problem, we can observe improvements between 17% and 39% if we compare standard DBRL with the method based on STS or Minkowski distances.

With the Ibex35 problem, we obtain similar results with both approaches. We think that the main reason for this situation is because the problem has very few records to re-identify.

Finally, in the forecasters problem we obtain similar or worst results with our approach than with standard DBRL. In our opinion this is reasonable because all time series are increasing and the difference between points t and $t+1$ is very large. In this kind of series a standard normalization is possible because we can understand every point of a time series as a stand alone variable.

4.4 Analysis of Distances for Time Series

Although the best distance is the one that re-identifies more records, other aspects should be taken into account when selecting a distance as *e.g.* computational cost, number of records to be compared, size of the time series. We analyse below the distances we have considered.

Table 6. Results of Ibex35 problem

number of time series	4					8					40				
k	2	3	6	9	12	2	3	6	9	12	2	3	6	9	12
DBRL Standard	19	11	10	7	7	20	10	10	10	10	18	11	9	12	13
Minkowski $q = 2$	20	13	6	4	2	19	13	6	4	2	18	15	7	7	6
Minkowski $q = 3$	10	6	2	1	1	12	3	2	2	1	13	4	2	4	2
Minkowski $q = 4$	20	12	8	4	2	19	13	9	4	2	18	15	10	7	9
Minkowski $q = 5$	11	4	5	2	1	15	3	4	3	1	14	5	3	6	5
Cross correlation $\beta = 1$	0	0	1	1	1	0	0	0	3	3	1	0	0	0	0
Cross correlation $\beta = 2$	0	0	0	0	0	1	2	1	1	3	17	17	6	5	5
Cross correlation $\beta = 3$	0	0	1	1	1	0	0	0	2	3	2	0	0	1	0
Cross correlation $\beta = 4$	0	0	0	0	0	2	3	2	1	2	16	12	4	4	4
STS	21	14	10	4	3	20	14	9	4	3	20	19	10	7	8

Table 7. Results of Soccer problem

number of time series	8				
k	2	3	6	9	12
DBRL Standard	144	123	83	64	54
Minkowski $q = 2$	154	142	134	133	123
Minkowski $q = 3$	6	2	3	6	5
Minkowski $q = 4$	140	117	91	88	81
Minkowski $q = 5$	14	6	11	9	6
Cross correlation $\beta = 1$	0	0	0	1	0
Cross correlation $\beta = 2$	0	2	7	9	4
Cross correlation $\beta = 3$	0	0	0	1	0
Cross correlation $\beta = 4$	0	5	7	9	3
STS	174	168	121	86	62

In the case of having time series of different length, we can not use Minkowski or STS distances and we have to use DTW distance. However, this distance, even in the case of using dynamic programming, has a computational cost that is higher than the one of Minkowski, STS or Two cross correlation based. For this reason, when all time series have the same length we do not recommend to use DTW distance.

When there are time series with the same shape as *e.g.* food prices (prices with similar inflation), it is more reasonable to use the Minkowski distance instead of the STS distance. This is so because the Minkowski distance is based on the distance between data components while the STS distance is based on the shape of the time series. If all time series have the same shape, *e.g.* they are all increasing, STS distance is not a good election. Correct links would be hidden among incorrect ones as the degree of similarity between series will be similar.

When all time series have similar values, as *e.g.* prices of different brands of a single product, it is more appropriate to use the STS distance instead of the Minkowski distance. This is so because the Minkowski distance will obtain values near to zero for all time series pairs.

In all the experiments done so far, Minkowski, STS and Two cross correlation based distances have been applied to all scenarios. Instead, the DTW distance has only been applied to some of them because its computational cost is very high with large time series. We have only calculated this distance in the forecasters problem that has the shortest time series.

5 Conclusions and Future Work

In this paper, we have introduced a method for re-identification time series based on specific distances for this type of data.

We have tested our method in three different problems and we have obtained for most of the cases better results than those obtained using the standard method.

The comparison between time series distances shows that the best distances for our problems are STS, Minkowski distance and DTW if we can afford its computational cost. In addition, we have considered the problems related with a bad normalization process, using a specific time series normalization procedure.

As future work we include the analysis of more time series distances and make more test with other kind of problems (other data).

Acknowledgments

This work was partly funded by the Spanish Ministry of Education and Science under project SEG2004-04352-C04-02 ”PROPRIETAS”.

References

1. Agrawal, R., Srikant, R., (2000), Privacy Preserving Data Mining, Proc. of the ACM SIGMOD Conference on Management of Data, 439-450.
2. Stock Exchange web, Sabadell Bank, http://www.bsmarkets.com/
3. Chu, S., Keogh, E., Hart, D., Pazzani, M. (2002), Iterative Deepening Dynamic Time Warping for Time Series. The Second SIAM International Conference on Data Mining Chicago, USA, April 11-13, 2002.
4. Domingo-Ferrer, J., Torra, V., (2001), A Quantitative Comparison of Disclosure Control Methods for Microdata, Confidentiality, Disclosure, and Data Access: Theory and Practical Applications for Statistical Agencies, Elsevier Science, Pages 111-133, 2001.
5. Internation institute of forecasters, http://www.forecasters.org/
6. Möller-Levet, C. S., Klawonn, F., Cho, K.-H., Wolkenhauer, O., (2003), Fuzzy clustering of short time series and unevenly distributed sampling points, Proceedings of the 5th International Symposium on Intelligent Data Analysis, Berlin, Germany, August 28-30, 2003.
7. Football statistics web, http://www.histora.org
8. Nin J., Torra V., (2006), Extending microaggregation procedures for time series protection, RSCTC 2006, Lecture Notes in Artificial Intelligence, in press.

9. Rahm, E., Bernstein, P. A., (2001), A survey of approaches to automatic schema matching, The VLDB Journal, 10 334-350.
10. Torra, V., Domingo-Ferrer, J., (2003), Record linkage methods for multidatabase data mining, in V. Torra (Ed.), Information Fusion in Data Mining, Springer, 101-132.
11. Warren Liao, T., (2005), Clustering of time series data - a survey, Pattern Recognition, 38 1857-1874.
12. Willenborg, L., de Waal, T., (2001), *Elements of Statistical Disclosure Control*, Lecture Notes in Statistics, Springer-Verlag.
13. Winkler, W. E., (2003), Data Cleaning Methods, Proc. SIGKDD 2003, Washington.
14. Winkler, W. E., (2004), Re-identification methods for masked microdata, Privacy in Statistical Databases 2004, Lecture Notes in Computer Science 3050 216-230.

Beyond k-Anonymity: A Decision Theoretic Framework for Assessing Privacy Risk

Guy Lebanon[1], Monica Scannapieco[2,⋆],
Mohamed R. Fouad[1], and Elisa Bertino[1]

[1] Purdue University, USA
lebanon@stat.purdue.edu, {mrf, bertino}@cs.purdue.edu
[2] ISTAT and Università di Roma "La Sapienza", Italy
scannapi@istat.it

Abstract. An important issue any organization or individual has to face when managing data containing sensitive information, is the risk that can be incurred when releasing such data. Even though data may be sanitized, before being released, it is still possible for an adversary to reconstruct the original data by using additional information that may be available, for example, from other data sources. To date, however, no comprehensive approach exists to quantify such risks. In this paper we develop a framework, based on statistical decision theory, to assess the relationship between the disclosed data and the resulting privacy risk. We relate our framework with the k-anonymity disclosure method; we make the assumptions behind k-anonymity explicit, quantify them, and extend them in several natural directions.

1 Introduction

The problem of data privacy is today a pressing concern for many organizations and individuals. The release of data may have some important advantages in terms of improved services and business, and also for the society at large, such as in the case of homeland security. However, unauthorized data disclosures can lead to violations to the privacy of individuals, can result in financial and business damages, as in the case of data pertaining to enterprises; or can result in threats to national security, as in the case of sensitive GIS data [6]. Preserving the privacy of such data is a complex task driven by various goals and requirements. Two important privacy goals are: (i) preventing identity disclosure, and (ii) preventing sensitive information disclosure. Identity disclosure occurs when the released information makes it possible to identify entities either directly (e.g., by publishing identifiers like SSNs) or indirectly (e.g., by linkage with other sources). Sensitive information typically includes information that must be protected by law, for example medical data, or is required by the subjects described by the data. In the latter case, data sensitivity is a subjective measure and may differ across entities.

[⋆] This work was performed while visiting research assistant at CERIAS and Department of Computer Sciences, Purdue University, USA. It is partially supported by the project "ESTEEM" (http://www.dis.uniroma1.it/∼esteem/index.html).

J. Domingo-Ferrer and L. Franconi (Eds.): PSD 2006, LNCS 4302, pp. 217–232, 2006.

To date, an important practical requirement for any privacy solution is the ability to quantify the *privacy risk* that can be incurred in the release of certain data. However, most of the work related to data privacy has focused on how to transform the data so that no sensitive information is disclosed or linked to specific entities. Because such techniques are based on *data transformations* that modify the original data with the purpose of preserving privacy, they are mainly focused on the tradeoff between data privacy and data quality (see e.g. [9,2,3,5]). Conversely, few approaches exist to quantify privacy risks and thus to support informed decisions. Duncan et al. [4] describes a framework, called Risk-Utility (R-U) confidentiality map, which addresses the tradeoff between data utility and disclosure. Lakshmanan et al. [8] is an approach to the risk analysis for disclosed anonymized data that models a database as series of transactions and the attacker's knowledge as a belief function. Our model is fundamentally different from both works; indeed, we deal exactly with relational instances, rather than with generic files or data frequencies; also, we incorporate the concept of data sensitivity into our framework and we consider generic disclosure procedures, not only anonymization like in [8].

The goal of the work presented in this paper is to propose a comprehensive framework for the estimation of privacy risks. The framework is based on statistical decision theory and introduces the notion of a *disclosure rule*, that is a function representing the data disclosure policy. Our framework estimates the *privacy risk* by means of a function that takes into account a given disclosure rule and (possibly) the knowledge that can be exploited by the attacker. It is important to point out that our framework is able to assess privacy risks also when no information is available about the knowledge, referred to as *dictionary*, that the adversary may exploit. The privacy risk function incorporates both identity disclosure and sensitive information disclosure. We introduce and analyze different shapes of the privacy risk function. Specifically, we define the risk in the classical decision theory formulation and in the Bayesian formulation.We prove several interesting results within our framework: we show that, under reasonable hypotheses, the estimated privacy risk is an upper bound for the true privacy risk; we analyze the computational complexity of evaluating the privacy risk function, and we propose an algorithm for efficiently finding the disclosure rule that minimizes the privacy risk. We finally gain insight by showing that the privacy risk is a quantitative framework for exploring the assumptions and consequences of k-anonymity.

2 Privacy Risk Framework

As private information in databases is being disclosed, undesired effects occur such as privacy breaches, and financial loss due to identity theft. To proceed with a quantitative formalism we assume that we obtain a numeric description, referred to as loss, of that undesired effect. The loss may be viewed as a function of (i) whether the disclosed information enables identification and (ii) the sensitivity of the disclosed information. The first argument of the loss function encapsulates whether the disclosed data can be tied to a specific entity or not. Consider for example the case of a hospital disclosing a list of the ages of patients, together with data indicating whether they are healthy or not. Even though this data is sensitive and if there is a little chance that the disclosed information can be tied to specific individuals, no privacy loss occurs as the data is anonymous. The second argument of the loss function, the sensitivity of the disclosed information, may be high as is often the case for sensitive medical data. On the other hand, other disclosed information such as gender, may be only marginally private or not private at all. It is important to note that a precise quantification of the sensitivity of the disclosed information may depend on the entity to whom the data relates. For example, data such as annual income and past medical history may be very sensitive to some and only marginally sensitive to others.

Let T be a relation with a relational scheme $T(A_1, \ldots, A_m)$, where each attribute A_i is defined over the domain $\mathrm{Dom}_i \cup \{\perp, \S\}$, with the only exception of A_1 as detailed later. The relation T stores the records that are considered for disclosure and has some values either missing or suppressed for privacy preservation. Specifically, a null value is denoted by $\perp$ whereas a suppressed value is denoted by $\S$. Furthermore, we denote the different attribute values of a specific record $\boldsymbol{x}$ in T using a vector notation $(x_1, \ldots, x_m)$. The first attribute x_1 corresponds to a unique record identifier that can be neither $\perp$ nor $\S$. The set of all possible records may be written as

$$\mathcal{X} = (\mathrm{Dom}_1) \times (\mathrm{Dom}_2 \cup \{\perp, \S\}) \times \cdots \times (\mathrm{Dom}_m \cup \{\perp, \S\}).$$

If T has cardinality n, it can be seen as a subset of $\mathcal{X}^n$ which we may think of as a matrix whose rows are the different records. We refer to the i^{th} record in such a relation as $\boldsymbol{x}_i$ and its j^{th} attribute as x_{ij}. [1]

2.1 Disclosure Rules and Privacy Risk

Statistical decision theory [10] offers a natural framework for measuring the quantitative effect of the information disclosure phenomenon. The uncertainty is encoded by a parameter θ abstractly called "a state of nature" which is typically unknown. However, it is known that θ belongs to a set Θ, usually a finite or infinite subset of $\mathbb{R}^l$. The decisions are being made based on a sample of observations $(x_1, \ldots, x_n)$, $x_i \in \mathcal{X}$ and are represented via a function $\delta : \mathcal{X}^n \to \mathcal{A}$ where $\mathcal{A}$ is an abstract action space. The function δ is referred to as a decision policy or decision rule. A key element of statistical decision theory is that the state of nature θ governs the distribution p_θ that generates the observed data.

Instead of decision rules $\delta : \mathcal{X}^n \to \mathbb{A}$, we introduce *disclosure rules* defined as follows.

[1] Note that throughout the paper, records and vectors are denoted by ***bold italic*** symbols whereas variables and attributes are denoted by only *italic* symbols.

Definition 1. *A disclosure rule δ is a function $\delta : \mathcal{X} \to \mathcal{X}$ such that*

$$[\delta(\boldsymbol{z})]_j = \begin{cases} \perp & z_j = \perp \\ \S & \textit{the } j^{th} \textit{ attribute is suppressed} \\ z_j & \textit{otherwise} \end{cases}$$

The state of nature θ that influences the disclosure outcome is the side information used by the attacker in his identification attempt. Such side information θ is often a public data resource composed of identities and their attributes, for example a phone book. The distribution over records p_θ is taken to be the empirical distribution $\tilde{p}$ over the data that is to be disclosed $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$, defined below.

Definition 2. *The empirical distribution $\tilde{p}$ on $\mathcal{X}$ associated with a set of records $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ is*

$$\tilde{p}(\boldsymbol{z}) = \frac{1}{n} \sum_{i=1}^{n} 1_{\{\boldsymbol{z}=\boldsymbol{x}_i\}}$$

where $1_{\{\boldsymbol{z}=\boldsymbol{x}_i\}}$ is 1 if $\boldsymbol{z} = \boldsymbol{x}_i$ and 0 otherwise.

The empirical distribution is used for defining the risk associated with a disclosure rule δ using the mechanism of expectation. Note that the expectation with respect to $\tilde{p}$ is simply the empirical mean $E_{\tilde{p}}(f(\boldsymbol{x}, \theta)) = \frac{1}{n}\sum_{i=1}^{n} f(\boldsymbol{x}_i, \theta)$. The loss and risk functions in the privacy adaptation of statistical decision theory are defined below.

Definition 3. *The loss function $\ell : \mathcal{X} \times \Theta \to [0, \infty]$ measures the loss incurred by disclosing the data $\delta(\boldsymbol{z}) \in \mathcal{X}$ due to possible identification based on $\theta \in \Theta$.*

Definition 4. *The risk of the disclosure rule δ in the presence of side information θ is the average loss of disclosing the records $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$: $R(\delta, \theta) = E_{\tilde{p}(\boldsymbol{z})}(\ell(\delta(\boldsymbol{z}), \theta)) = \frac{1}{n}\sum_{i=1}^{n} \ell(\delta(\boldsymbol{x}_i), \theta)$.*

Definition 5. *The Bayes risk of the disclosure rule δ is $R(\delta) = E_{p(\theta)}(R(\delta, \theta))$ where $p(\theta)$ is a prior probability distribution on Θ.*

It is instructive at this point to consider in detail the identification process and its possible relations to the loss function. We use the term *identification attempt* to refer to the process of trying to identify the entity represented by the record. We refer to the subject performing the identification attempt as the *attacker.* The attacker performs an identification attempt based on the disclosed record $\boldsymbol{y} = \delta(\boldsymbol{x}_i)$ and additional side information θ referred to as a *dictionary.* The role of the dictionary is to tie a record $\boldsymbol{y}$ to a list of possible candidate identities consistent with the record $\boldsymbol{y}$, i.e. having the same values on common fields. For example, consider $\boldsymbol{y}$ being `(first-name,surname,phone#)` and the dictionary being a phone book. The attacker needs only considering dictionary entities that are consistent with the disclosed record. Recall that some of the attributes `(first-name,surname,phone#)` may be replaced with $\perp$ or § symbols due to missing information or due to the disclosure process, respectively. In this example, if all the attribute values are revealed and the available side information is an up-to-date phone book, it is likely that only one entity exists in the dictionary that is consistent with the revealed information. On the other hand, if the attribute value for `phone#` is suppressed, by replacing it with § symbol, the phone-book θ may or may not yield a single consistent entity, depending on the popularity of the (`first-name,surname`) combination. From the attacker's stand point, missing values are perceived the same way as suppressed values. Thus, in the rest of the paper and for the sake of notational simplicity, both missing and suppressed values will be denoted by the symbol $\perp$.

Note that the loss function $\ell(\delta(\boldsymbol{x}_i), \theta)$ measures the loss due to disclosing $\delta(\boldsymbol{x}_i)$ in the presence of the side information – in this case the dictionary θ. Specifying the loss is typically entity and problem dependent. We can, however, make some progress by decomposing the loss into two parts: (i) the ability to identify the entity represented by $\delta(\boldsymbol{x}_i)$ based on the side information θ and (ii) the sensitivity of the information in $\delta(\boldsymbol{x}_i)$. The identification part is formalized by the random variable Z defined as follows.

Definition 6. *Let $\rho(\delta(\boldsymbol{x}_i), \theta)$ denote the set of individuals in the dictionary θ consistent with the record $\delta(\boldsymbol{x}_i)$. Moreover, let the random variable $Z(\delta(\boldsymbol{x}_i))$ be a binary variable that takes value 1 if $\delta(\boldsymbol{x}_i)$ is identified and 0 otherwise.*

Assuming a uniform selection of entries in the dictionary by the attacker, we have

$$p_{Z(\delta(\boldsymbol{x}_i))}(1) = \begin{cases} |\rho(\delta(\boldsymbol{x}_i), \theta)|^{-1} & \rho(\delta(\boldsymbol{x}_i), \theta) \neq \emptyset \\ 0 & \text{otherwise} \end{cases}$$

and $p_{Z(\delta(\boldsymbol{x}_i))}(0) = 1 - p_{Z(\delta(\boldsymbol{x}_i))}(1)$.

2.2 Sensitivity

The sensitivity of disclosed data is formalized by the following definition.

Definition 7. *The sensitivity of a record is measured by a function $\Phi : \mathcal{X} \to [0, +\infty]$ where higher values indicate higher sensitivity.*

We allow Φ to take on the value $+\infty$ in order to model situations where the information in the record is so private that its disclosure is prohibited under any positive identification chance.

The sensitivity $\Phi(\delta(\boldsymbol{x}_i))$ measures the adverse effect of disclosing the record $\delta(\boldsymbol{x}_i)$ if the attacker correctly identifies it. We make the assumption (whose relaxation is straightforward) that if the attacker does not correctly identify the disclosed record, there is no adverse effect. The adverse effect is therefore a random variable with two possible outcomes: $\Phi(\delta(\boldsymbol{x}_i))$ with probability $p_{Z(\delta(\boldsymbol{x}_i))}(1)$ and 0 with probability $p_{Z(\delta(\boldsymbol{x}_i))}(0)$. It is therefore natural to account for the uncertainty resulting from possible identification by defining the loss $\ell(\boldsymbol{y},\theta)$ as the expectation of the adverse effect resulting from disclosing $\boldsymbol{y}=\delta(\boldsymbol{x}_i)$

$$\begin{aligned}\ell(\boldsymbol{y},\theta) &= E_{p_{Z(\boldsymbol{y})}}(\Phi(\boldsymbol{y})Z(\boldsymbol{y}))\\ &= p_{Z(\boldsymbol{y})}(1)\cdot\Phi(\boldsymbol{y})+p_{Z(\boldsymbol{y})}(0)\cdot 0=\frac{\Phi(\boldsymbol{y})}{|\rho(\boldsymbol{y},\theta)|}\end{aligned}$$

where the last equality holds if the dictionary selection probabilities are uniform and $\rho(\boldsymbol{y},\theta)\neq\phi$.

The risk $R(\delta,\theta)$ with respect to the distribution $\tilde{p}$ that governs the record set $\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n$ becomes

$$R(\delta,\theta)=E_{\tilde{p}(\boldsymbol{x})}(\ell(\delta(\boldsymbol{z}),\theta))=\frac{1}{n}\sum_{i=1}^{n}\frac{\Phi(\delta(\boldsymbol{x}_i))}{|\rho(\delta(\boldsymbol{x}_i),\theta)|}$$

and the Bayes risk under the prior $p(\theta)$ becomes (if Θ is discrete replace the integral below by a sum)

$$R(\delta)=E_{p(\theta)}(R(\delta,\theta))=\frac{1}{n}\sum_{i=1}^{n}\Phi(\delta(\boldsymbol{x}_i))\int_{\Theta}\frac{p(\theta)d\theta}{|\rho(\delta(\boldsymbol{x}_i),\theta)|}.$$

We now provide more details concerning records $\boldsymbol{x}_i$ and their space $\mathcal{X}$, that will be useful in the following. As introduced, a record $\boldsymbol{x}_i\in\mathcal{X}$ has attribute values $(x_{i1},\ldots,x_{im})$ where each attribute $x_{ij}, j=2,\ldots,m$ either takes values in a domain Dom_j or is unavailable, in which case we denote it by $\perp$. The first attribute is $x_{i1}\in\mathrm{Dom}_1$. We assume that $[\delta(\boldsymbol{x}_i)]_1=x_{i1}$, i.e. $[\delta(\boldsymbol{x}_i)]_1$ cannot have $\perp$ values. This assumption is for notational purposes only and in reality the disclosed data should be taken to be $[\delta(\boldsymbol{x}_i)]_2,\ldots,[\delta(\boldsymbol{x}_i)]_m$. Notice also that the primary key of the relation can be distinct from the introduced record identifier, and can be one or more attributes defined over $\mathrm{Dom}_2,\ldots,\mathrm{Dom}_m$. We make the assumption $[\delta(\boldsymbol{x}_i)]_1=x_{i1}$ in order to allow a possible dependency of $\Phi(\delta(\boldsymbol{x}_i))$ on the identifier $x_{i1}=[\delta(\boldsymbol{x}_i)]_1$ which enables the flexibility needed to treat attribute values related to different entities differently. For example, a certain entity, such as a specific person, may wish to protect certain attributes such as religion or age that may be less private for a different person. Possible expressions for the Φ function are provided in the Appendix.

3 Tradeoff Between Disclosure Rules and Privacy Risk

In evaluating disclosure rules δ we have to balance the following tradeoff. On one hand, disclosing private information incurs the privacy risk $R(\delta,\theta)$. On the other hand, disclosing information serves some purpose, or else no information would ever be disclosed. Such disclosure benefit may arise from various reasons such as increased productivity due to the sharing of commercial data.

We choose to represent this tradeoff by specifying a set of disclosure rules Δ that are acceptable in terms of their disclosure benefit. From this set, we seek to choose the rule that incurs the least privacy risk $\delta^* = \arg\min_{\delta \in \Delta} R(\delta, \theta)$. Notice that this framework is not symmetric in its treatment of the disclosure benefit and privacy risk and emphasizes the increased importance of privacy risk in the tradeoff.

It is difficult to provide a convincing example of a set Δ without specifying in detail the domain and the disclosure benefit. Nevertheless, we specify below several sets of rules that serve to illustrate the decision theoretic framework of this paper. The basic principle behind these rules is that the more attribute values are being disclosed, the greater the disclosure benefit is. The details of the specific application will eventually determine which set of rules is most appropriate.

The three sets of rules below are parameterized by a positive integer k. The set Δ_1 consists of rules that disclose a total of k attribute values for all records combined

$$\Delta_1 = \\ \{\delta : \delta(\boldsymbol{x}_1), \ldots, \delta(\boldsymbol{x}_n) \text{ contain a total of } k \text{ non } \perp \text{ entries}\}.$$

The second set Δ_2 consists of rules that disclose a certain number of attribute values for each record

$$\Delta_2 = \{\delta : \forall i \;\; \delta(\boldsymbol{x}_i) \text{ contains } k \text{ non } \perp \text{ entries}\}.$$

The third set Δ_3 consists of rules that disclose a certain number of attribute values for each attribute

$$\Delta_3 = \{\delta : \forall j \;\; \{[\delta(\boldsymbol{x}_i)]_j\}_{i=1}^n \text{ contains } k \text{ non } \perp \text{ entries}\}.$$

The set Δ_1 may be applicable in situations where the disclosure benefit is influenced simply by the number of disclosed attribute values. Such a situation may arise if there is a need for computing statistics on the joint space of represented entities-attributes without an emphasis on either dimension. The set Δ_2 may be applicable when the disclosure benefit is tied to per-entity data, for example discovering association rules in grocery store transactions. A rule $\delta \in \Delta_2$ guarantees that there are sufficient attributes disclosed for each entity to obtain meaningful conclusions. Similarly, the set Δ_3 may be useful in cases where there is an emphasis on per-attribute data.

Disclosure rules $\delta \in \Delta$ are evaluated on the basis of the risk functions $R(\delta, \theta)$, $R(\delta)$. In some cases, the attacker's dictionary is publicly available. We can then treat the "true" side information θ^{true} as known, and the optimal disclosure rule is the minimizer of the risk

$$\delta^* = \arg\min_{\delta \in \Delta} R(\delta, \theta^{\text{true}}). \tag{1}$$

If the attacker's side information is not known, but we can express a prior belief $p(\theta)$ describing the likelihood of $\theta^{\text{true}} \in \Theta$, we may use the Bayesian approach and select the minimizer of the Bayes risk

$$\delta^*_B = \arg\min_{\delta \in \Delta} E_{p(\theta)}(R(\delta, \theta)). \tag{2}$$

If there is no information concerning θ^{true} whatsoever, a sensible strategy is to select the minimax rule δ^*_M that achieves the least worst risk, i.e. δ^*_M satisfies

$$\sup_{\theta \in \Theta} R(\delta^*_M, \theta) = \inf_{\delta \in \Delta} \sup_{\theta \in \Theta} R(\delta, \theta). \tag{3}$$

Notice that in all cases above we try to pick the best disclosure rule in terms of privacy risk, out of a set Δ of disclosure rules that are acceptable in terms of the amount of revealed data. The rules $\delta^*, \delta^*_B, \delta^*_M$ are useful, respectively, if we know θ^{true}, we have a prior over it, or we have no knowledge whatsoever.

An alternative situation to the one above is that the database is trying to estimate (or minimize) the privacy risk $R(\delta, \theta^{\text{true}})$ based on side information $\hat{\theta} \neq \theta^{\text{true}}$ available to the database. In such cases we can use $R(\delta, \hat{\theta})$ as an estimate for $R(\delta, \theta^{\text{true}})$ but we need to find a way to connect the two risks above by leveraging on a relation between $\hat{\theta}$ and θ^{true}.

A reasonable assumption is that the database dictionary $\hat{\theta}$ is specific to the database while the attacker's dictionary θ^{true} is a more general-purpose dictionary. We can then say that θ^{true} contains the records in $\hat{\theta}$ as well as additional records. Following the same reasoning we can also assume that for each record that exists in both dictionaries, $\hat{\theta}$ will have more attribute values that are not $\perp$. For example, consider a database of employee records for some company. $\hat{\theta}$ would be the database dictionary and θ^{true} would be a general-purpose dictionary such as a phone-book. It is natural to assume that θ^{true} will contain additional records over the records in $\hat{\theta}$ and that the non-$\perp$ attributes in θ^{true} (e.g. `first-name,surname,phone#`) will be more limited than the non-$\perp$ attributes in $\hat{\theta}$. After all, some of the record attributes are private and would not be disclosed in order to find their way into the attacker's dictionary (resulting in more $\perp$ symbols in the θ^{true}).

Under the conditions specified above we can show that the true risk is bounded from above by $R(\delta, \hat{\theta})$ and that the chosen rule $\arg\min_{\delta \in \Delta} R(\delta, \hat{\theta})$ has a risk that is guaranteed to bound the true privacy risk. This is formalized below.

We consider dictionaries θ as relational tables, where $\boldsymbol{\theta}_i = (\theta_{i1}, \ldots, \theta_{iq})$ is a record of a relation $T_\theta(A_1, \ldots, A_q)$, with A_1 corresponding to the record identifier.

Definition 8. *We define the relation $\preceq$ between dictionaries $\theta = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_{l_1})$ and $\eta = (\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_{l_2})$ by saying that $\theta \preceq \eta$ if for every $\boldsymbol{\theta}_i$, $\exists \boldsymbol{\eta}_v$ such that $\boldsymbol{\eta}_{v1} = \boldsymbol{\theta}_{i1}$ and $\eta_{vk} \neq \perp \Rightarrow \theta_{ik} = \eta_{vk}$. The relation $\preceq$ constitutes a partial ordering on the set of dictionaries Θ.*

Theorem 1. *If $\hat{\theta}$ contains records that correspond to$\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ and $\hat{\theta} \preceq \theta^{true}$, then*

$$\forall \delta \quad R(\delta, \theta^{true}) \leq R(\delta, \hat{\theta}).$$

Proof. For every disclosed record $\delta(\boldsymbol{x}_i)$ there exists a record in $\hat{\theta}$ that corresponds to it and since $\hat{\theta} \preceq \theta^{\text{true}}$ there is also a record in θ^{true} that corresponds to it. As a result, $\rho(\delta(\boldsymbol{x}_i), \hat{\theta})$ and $\rho(\delta(\boldsymbol{x}_i), \theta^{\text{true}})$ are non-empty sets.

For an arbitrary $\boldsymbol{a} \in \rho(\delta(\boldsymbol{x}_i), \hat{\theta})$ we have $\boldsymbol{a} = \hat{\boldsymbol{\theta}}_v$ for some v and since $\hat{\theta} \preceq \theta^{\text{true}}$ there exists a corresponding record $\boldsymbol{\theta}_k^{\text{true}}$. The record $\boldsymbol{\theta}_k^{\text{true}}$ will have the same or more $\perp$ symbols as $\boldsymbol{a}$ and therefore $\boldsymbol{\theta}_k^{\text{true}} \in \rho(\delta(\boldsymbol{x}_i), \theta^{\text{true}})$. The same argument can be repeated for every $\boldsymbol{a} \in \rho(\delta(\boldsymbol{x}_i), \hat{\theta})$ thus showing that $\rho(\delta(\boldsymbol{x}_i), \hat{\theta}) \subseteq \rho(\delta(\boldsymbol{x}_i), \theta^{\text{true}})$ or $|\rho(\delta(\boldsymbol{x}_i), \theta^{\text{true}})|^{-1} \leq |\rho(\delta(\boldsymbol{x}_i), \hat{\theta})|^{-1}$.

The probability of identifying $\delta(\boldsymbol{x}_i)$ by the attacker is thus smaller than the identification probability based on $\hat{\theta}$. It then follows that for all i, $\ell(\delta(\boldsymbol{x}_i), \theta^{\text{true}}) \leq \ell(\delta(\boldsymbol{x}_i), \hat{\theta})$ as well as $R(\delta, \theta^{\text{true}}) \leq R(\delta, \hat{\theta})$.

Solving (1) in the general case requires evaluating $R(\delta, \theta^{\text{true}})$ for each $\delta \in \varDelta$ and selecting the minimum. The reason is that the dictionary θ controlling the identification distribution $p_{Z(\delta(\boldsymbol{x}_i))}(1) = |\rho(\delta(\boldsymbol{x}_i), \theta)|^{-1}$ is of arbitrary shape. A practical assumption, that is often made for high dimensional distributions, is that the distribution underlying θ factorizes into a product form

$$\frac{|\rho(\delta(\boldsymbol{x}_i), \theta)|}{N} = \prod_j \frac{|\rho_j([\delta(\boldsymbol{x}_i)]_j, \theta)|}{N} \quad \text{or}$$
$$|\rho(\delta(\boldsymbol{x}_i), \theta)| = \prod_j \alpha_j([\delta(\boldsymbol{x}_i)]_j, \theta)$$

for some appropriate functions α_j. In other words, the appearances of y_j for all $j = 1, \ldots, m$ in θ are independent random variables. Returning to the phonebook example, the above assumption implies that the popularity of first names does not depend on the popularity of last names, e.g.,

$$\begin{aligned} &p(\text{first-name} = \texttt{Mary} | \text{surname} = \texttt{Smith}) = \\ &p(\text{first-name} = \texttt{Mary} | \text{surname} = \texttt{Johnson}) = \\ &p(\text{first-name} = \texttt{Mary}). \end{aligned}$$

The independence assumption does not hold in general, as attribute values may be correlated, for instance, by integrity constraints; we plan to relax it in future work.

First we analyze the complexity of evaluating the risk function $R(\delta, \theta)$. This would depend on the complexity of computing $\varPhi$, denoted by $C(\varPhi)$, and the complexity of computing $|\rho(\delta(\boldsymbol{x}_i), \theta)|$ which is $O(Nm)$, where N is the number of records in the dictionary θ. Solving $\arg\min_{\delta \in \varDelta} R(\delta, \theta)$ by enumeration requires $O(n)\big(C(\varPhi) + O(Nm)\big) \cdot |\varDelta|$ computations.

We have $|\varDelta_1| = \binom{nm}{k}$, $|\varDelta_2| = \binom{m}{k}^n$, $|\varDelta_3| = \binom{n}{k}^m$ and $C(\varPhi)$ typically being $O(m)$ for the additive and multiplicative forms. In a typical setting where $k \ll m$ we have for $\varDelta_2$ and a linear or multiplicative $\varPhi$, a minimization complexity of $O(nNm^{kn+1})$.

The complexities above are computed for the naive enumeration algorithm. A much more efficient algorithm for obtaining $\arg\min_{\delta \in \varDelta} R(\delta, \theta)$ for $\varDelta_2$ and $\varPhi_5$ under the assumption of dictionary independence is described below.

If we define $C_1(\boldsymbol{y}) = \{j : j > 1, y_j = \perp\}, C_2(\boldsymbol{y}) = \{j : j > 1, y_j \neq \perp\}$ we have

$$\begin{aligned}
\ell(\boldsymbol{y}, \theta) &= \frac{\prod_{j \in C_2(\boldsymbol{y})} e^{w_{j,y_1}}}{|\rho(\boldsymbol{y}, \theta)|} = \frac{\prod_{j \in C_2(\boldsymbol{y})} e^{w_{j,y_1}}}{\prod_{k>1} \alpha_k(y_k, \theta)} \\
&= \prod_{j \in C_2(\boldsymbol{y})} \frac{e^{w_{j,y_1}}}{\alpha_j(y_j, \theta)} \cdot \prod_{l \in C_1(\boldsymbol{y})} \frac{1}{\alpha_l(\perp, \theta)} \\
&= \prod_{j \in C_2(\boldsymbol{y})} e^{w_{j,y_1}} \frac{\alpha_j(\perp, \theta)}{\alpha_j(y_j, \theta)} \cdot \prod_{l=2}^{m} \frac{1}{\alpha_l(\perp, \theta)}.
\end{aligned}$$

To select the disclosure of k attributes that minimizes the above loss it remains to select the set $C_2(\boldsymbol{y})$ of k indices that minimizes the loss. This set corresponds to the k smallest $\{e^{w_{j,y_1}} \frac{\alpha_j(\perp,\theta)}{\alpha_j(y_j,\theta)}\}_{j=2}^{m}$ and leads to the following algorithm.

Algorithm 1. `MinRisk`

```
foreach $i = 1, \ldots, n$
    foreach $j = 2, \ldots, m$
        set $\gamma_j := e^{w_{j,x_{i1}}} \frac{\alpha_j(\perp,\theta)}{\alpha_j(x_{ij},\theta)}$
    identify the $k$ smallest elements in $\{\gamma_j\}_{j=2}^{m}$
    set $\delta(\boldsymbol{x}_i)$ to disclose the attributes corresponding to these $k$ elements
```

Theorem 2. *The algorithm* ***MinRisk*** *for solving* $\arg\min_{\delta \in \Delta} R(\delta, \theta)$ *requires* $O(nNm)$ *computations.*

Proof. For each record $\boldsymbol{y} = \boldsymbol{x}_i$ we compute the following. The set $\{\gamma_j = e^{w_{j,y_1}} \frac{\alpha_j(\perp,\theta)}{\alpha_j(y_j,\theta)}\}_{j=2}^{m}$ can be obtained in $O(Nm)$. Moreover, the set corresponding to the k smallest elements in $\{\gamma_j\}_{j=2}^{m}$ can be obtained in two steps: (i) Get the k^{th}-smallest element in $\{\gamma_j\}_{j=2}^{m}$, γ' (this requires $O(m)$ computations), then (ii) scan the set $\{\gamma_j\}_{j=2}^{m}$ for elements $< \gamma'$ (again, this requires $O(m)$ computations). Hence the overall complexity of the above procedure is $O(n)\big(O(Nm) + O(m)\big) = O(nNm)$.

4 Privacy Risk and *k*-Anonymity

k-Anonymity [9] has recently received considerable attention by the research community [11,1]. Given a relation T, k-anonymity ensures that each disclosed record can be indistinctly matched to at least k individuals in T. It is enforced by considering a subset of the attributes called *quasi-identifiers*, and forcing the disclosed values of these attributes to appear at least k times in the database. k-anonymity uses two operators to accomplish this task: suppression and generalization. We ignore the role of generalization operators in this paper as our privacy framework is cast solely in terms of suppression at attribute-level. However, it is straightforward to extend the privacy risk framework to include generalization operators leading to a more complete analogy with k-anonymity, and we plan to do it in future work.

In its original formulation, k-anonymity does not seem to make any assumptions on the possible external knowledge that could be used for entity identification and does not characterize the privacy loss. However, k-anonymity does make strong implicit assumptions whose absence eliminates any motivation it might possess. Following the formal presentation of k-anonymity in the privacy risk context, we analyze these assumptions and possible relaxations.

Since the k-anonymity requirement is enforced on the relation T, the anonymization algorithm considers the attacker's dictionary as equal to the relation $T = \theta$. Representing the k-anonymity rule by δ_k^* we have that the k-anonymity constraints may be written as

$$\forall i \quad |\rho(\delta_k^*(\boldsymbol{x}_i), T)| \geq k. \tag{4}$$

The sensitivity function is taken to be constant $\Phi \equiv c$ as k-anonymity considers only the constraints (4) and treats all attributes and entities in the same way. As a result, the loss incurred by k-anonymity δ_k^* is bounded by $\ell(\delta_k^*(\boldsymbol{x}_i), T) \leq c/k$ where equality is achieved if the constraint $|\rho(\delta_k^*(\boldsymbol{x}_i), T)| = k$ is met. On the other hand, any rule δ_0 that violates the k-anonymity requirement for some $\boldsymbol{x}_i$ will incur a loss higher (under $\theta = T$ and $\Phi \equiv c$) than the k-anonymity rule

$$\ell(\delta_0(\boldsymbol{x}_i), T) = \frac{c}{|\rho(\delta_0(\boldsymbol{x}_i), T)|} \geq \ell(\delta_k^*(\boldsymbol{x}_i), T).$$

We thus have the following result presenting k-anonymity as optimal in terms of the privacy risk framework.

Theorem 3. *Let δ_k^* be a k-anonymity rule and δ_0 be a rule that violates the k-anonymity constraint, both with respect to $\boldsymbol{x}_i \in T$. Then*

$$\ell(\delta_k^*(\boldsymbol{x}_i), T) \leq c/k < \ell(\delta_0(\boldsymbol{x}_i), T).$$

As the above theorem implies, the k-anonymity rule minimizes the privacy loss per example $\boldsymbol{x}_i$ and may be seen as $\arg\min_{\delta \in \Delta} R(\delta, T)$ where Δ is a set of rules that includes both k-anonymity rules and rules that violate the k-anonymity constraints. The assumptions underlying k-anonymity, in terms of the privacy risk framework are

1. $\theta^{\text{true}} = T$
2. $\Phi \equiv c$
3. Δ is under-specified.

The first assumption may be taken as an indication that k-anonymity does not assume any additional information regarding the attacker's dictionary. As we showed earlier, the resulting risk $R(\delta_k^*, T) \leq c/k$ may be seen as a bound on the true risk $R(\delta_k^*, \theta^{\text{true}})$ under some assumptions. Alternatively, the privacy framework also introduces the mechanisms of the minimax rule and the Bayes rule if additional information is available such as the set Θ of possible dictionaries or even a prior on Θ. Moreover, the attacker's dictionary θ is often a standard public resource. In such cases the constraints (4) should be taken with respect to θ, rather than T.

The second assumption $\Phi \equiv c$ is somewhat questionable. The privacy risk framework measures the loss as the expectation of the data sensitivity, as measured by Φ, with respect to the identification probability. Taking the sensitivity function Φ to be a constant ignores the role of the sensitivity of the disclosed data in the framework. The loss measured would depend only on the identification probability and not on the types of attributes that are being disclosed. In other words, privacy loss becomes synonymous with identification. This leads to the paradoxical situation where the disclosure of a sensitive attribute such as the type of medical situation diagnosed (e.g. HIV positive) may lead to lower risk than the disclosure of a less sensitive attribute such as the precise date of the most recent doctor visit (assuming that the precise date of the most recent doctor visit leads to greater identification chance).

The third assumption implies that the set Δ may be specified in several ways. Recall that the risk minimization framework is based on the assumption that there is a tradeoff in disclosing private information. On one hand the disclosed data incurs a privacy loss and on the other hand disclosing data serves some benefit. The risk minimization framework $\arg\min_{\delta\in\Delta} R(\delta,\theta)$ assumes that Δ contains a set of rules acceptable in terms of their disclosure benefit, and from which we select the one incurring the least risk. k-Anonymity ignores this tradeoff and the set of candidate rules Δ may be specified in several ways, for example $\Delta = \Delta_0 \cup \{\delta^*_k\}$ where Δ_0 contains rules that violate the k-anonymity constraints.

In light of the above, k-anonymity may be modified in several directions. If we possess some information concerning the attacker's dictionary we can do a better job using $\delta^*, \delta^*_B, \delta^*_M$, as well as upper-bound the true risk using $\hat{\theta}$ (see Section 5 for an explanation of these concepts). We can alter Φ to account for the different sensitivities of different attributes, perhaps even allowing Φ to be entity-dependent. Finally, a more careful consideration of the disclosure benefit may lead to a better definition of the rule set Δ allowing the preference of some k-anonymity rule over others. As mentioned earlier, our discussion is in terms of the suppression operator alone. Nevertheless, the same arguments and conclusions apply to k-anonymity using both suppression and generalization operators.

5 Conclusion

In this paper we have described a novel framework for assessing privacy risk in a variety of situations. We consider optimal disclosure rules in the contexts of exact knowledge, partial knowledge, and no knowledge with respect to the attacker's side information. We discuss several forms for expressing the largely ignored role of data sensitivity in the privacy risk. We have shown that the estimated privacy risk is an upper bound for the true privacy risk, under some reasonable hypotheses on the relationships between the attacker's dictionary and the database dictionary. We have also provided a computationally efficient algorithm for minimizing the privacy risk under some hypotheses. Finally, we have proved the generality of our framework by showing that k-anonymity is a special case of it, and we have highlighted, in our decision theory based formulation, the particular assumptions underlying k-anonymity.

At first glance it may appear that the privacy risk framework requires knowledge that is typically unavailable or somewhat undesirable assumptions. After all, it seems possible to use k-anonymity without making such compromising assumptions. This is a misleading interpretation as any attempt at forming a sensible privacy policy or characterizing the result of private data disclosure requires such assumptions. In particular, assumptions have to be made concerning the attacker's resources and the data sensitivity. Existing algorithms such as k-anonymity typically make such assumptions implicitly. However, in order to obtain a coherent view of privacy it is essential to make these assumptions explicit, and discuss their strengths and weaknesses.

References

1. G. Aggarwal, T. Feder, K. Kenthapadi, R. Motwani, P. Panigrahy, D. Thomas, and A. Zhu. Anonymizing tables. In *Proc. of ICDT 2005.*
2. Avrim Blum, Cynthia Dwork, Frank McSherry, and Kobbi Nissim. Practical privacy: The sulq framework. In *Proc. of PODS 2005.*
3. I. Dinur and K. Nissim. Revealing information while preserving privacy. In *Proc. of PODS 2003.*
4. G.T. Duncan, S.A. Keller-McNulty, and L.S. Stokes. Disclosure Risk vs. Data Utility: The R-U Confidentiality Map. Technical Report 121, National Institute of Statistical Sciences (NISS), December 2001.
5. A. Evfimievski, J. Gehrke, and R. Srikant. Limiting privacy breaches in privacy preserving data mining. In *Proc. of PODS 2003.*
6. Guidelines for Providing Appropriate Access to Geospatial Data in Response to Security Concerns. Federal geographic data committee, 2005. `http://fgdc.er.usgs.gov/fgdc/homeland/access_ guidelines.pdf`.
7. M.A. Jaro. UNIMATCH: A record linkage system, user's manual. In *Washington DC: U.S. Bureau of the Census*, 1978.
8. L.V.S. Lakshmanan, R.T. Ng, and G. Ramesh. To do or not to do: the dilemma of disclosing anonymized data. In *Proc. of SIGMOD 2005.*
9. P. Samarati and L. Sweeney. Generalizing data to provide anonymity when disclosing information (abstract). In *Proc. of PODS 1998.*
10. A. Wald. *Statistical Decision Functions.* Wiley, 1950.
11. S. Zhong, Z. Yang, and R.N. Wright. Privacy-enhancing k-anonymization of customer data. In *Proc. of PODS 2005.*

Appendix

Possible Expression for the Φ Function

In the following, we review several possible expressions for the function Φ, which models the sensitivity of a record that is involved in a disclosure process, if the attacker correctly identifies it. Since Φ is defined on $\mathcal{X}$, the set of all possible records, defining it by a lookup table is often impractical. We therefore consider several options leading to compact and efficient representation. Given a disclosed record $\boldsymbol{y} = \delta(\boldsymbol{x}_i)$, the simplest meaningful form for Φ is a linear combination of non-negative weights w_j over the disclosed attributes

$$\Phi_1(\boldsymbol{y}) = \sum_{j>1:y_j \neq \perp} w_j$$

where w_j represents the sensitivity associated with the corresponding attribute A_j. A weight of $+\infty$ represents the most sensitive information that may only be disclosed if there is zero chance of it leading to identification (we define $0\cdot\infty = 0$).

Alternatively, we may assume that attributes sensitivities vary from record to record, but are identical for each record. In this case a linear form would yield

$$\Phi_2(\boldsymbol{y}) = \sum_{j>1:y_j \neq \perp} w_{y_1} = w_{y_1} \times (\#\text{ of disclosed attributes})$$

where w_{y_1} is the weight associated with the release of each attribute value of the record $\boldsymbol{y}$. Incorporating different weights for different attribute values and different records yields the linear form

$$\Phi_3(\boldsymbol{y}) = \sum_{j>1:y_j \neq \perp} w_{j,y_1}$$

where w_{j,y_1} represents the sensitivity of attribute j in record $\boldsymbol{y}$. These weights may be represented by a two dimensional table of numbers. A possible special case is to assume the decomposition of attribute-record weights $w_{j,y_1} = w_j w'_{y_1}$ leading to a representation of the weight table as two vectors of weights $(w_1, \ldots, w_m)$, $(w'_1, \cdots, w'_n)$.

An extension of the linear representation of Φ is accomplished through k-order interactions. In k-order interaction we use additional weights to capture interactions of k attributes that are not accounted for in the linear forms above. k-order interactions take into account cases in which the simultaneous disclosure of multiple attribute values needs to be weighted differently when compared to the independent disclosure of each single value. An example of $k = 2$-order interaction yields the form

$$\Phi_4(\boldsymbol{y}) = \sum_{j>1:y_j \neq \perp} w_{j,y_1} + \sum_{j>1:y_j \neq \perp} \sum_{h>j:y_h \neq \perp} w_{j,h,y_1}.$$

As k increases in magnitude, the class of functions represented by Φ becomes richer. Reaching $k = m$ would provide arbitrary flexibility with respect to the functional form of Φ (however, as mentioned above, the representation and computation are impractical for large m). If the simple linear form is not sufficient to capture the user-specified privacy values, it is likely that increasing k to 2 or 3 will bring the functional form of Φ quite close to the user-specified values.

In some cases, a multiplicative rather than linear form is preferred. In this case, a convenient form is

$$\Phi_5(\boldsymbol{y}) = \exp\left(\sum_{j>1:y_j\neq\perp} w_{j,y_1}\right) = \prod_{j>1:y_j\neq\perp} e^{w_{j,y_1}}$$

or its k-order extensions analogous to the linear forms above. The multiplicative form Φ_5 has the advantage that if we increase one privacy weight w_{ij} while fixing the other weights, Φ increases exponentially rather than linearly. Since the disclosure of extremely private information should not be considered even if the remaining attributes are non-private, the multiplicative form Φ_5 is the most appropriate in many settings.

Experiments

The goals of our experiments are 3-fold: (i) to validate the risk associated with different dictionaries, (ii) to assess the impact of different parameters on the privacy risk, and (iii) to use the proposed framework to assess the relationship between the estimated risk and the true risk.

We conducted our experiments on a real Wal-Mart database: An `item description` table of more than 400,000 records each with more than 70 attributes is used in the experiments. Part of the table is used to represent the disclosed data whereas the whole table is used to generate different dictionary. Throughout all our experiments, the risk components are computed as follows. First, the identification risk is computed with the aid of the Jaro distance function[7] that is used to identify dictionary items consistent with a released record to a certain extent (we used 80% similarity threshold to imply consistency.) Second, the sensitivity of the disclosed data is assessed by means of random weights that are generated using a uniform random number generator.

The impacts of the number of disclosed attributes per record, k, and the dictionary size on the privacy risk are reported in Figure 1 (left). As k increases (i.e. extra data is being disclosed) and by fixing the dictionary size, the possibility of identifying the entity, to which the data pertain, increases, thus increasing the privacy risks. We increase k from 25% to 100% of the total number of attributes. On the other hand, by fixing the number of data attributes that are disclosed, the relation between the risk and dictionary size is inversely related. The larger the size of the dictionary the attacker uses, the lower the probability that the entity be identified. Different dictionaries are generated from the original table with sizes varying from 10% to 100% of the size of the whole table. Moreover, the experimental data show that the multiplicative model for sensitivity is always superior in terms of the modeled risk to the additive model.

The relationship between the true risk $R(\delta, \theta^{\text{true}})$ and the estimated risk $R(\delta, \hat{\theta})$ is reported in the scatter plot in Figure 1 (right). As we proved before, $R(\delta, \hat{\theta})$ is always an upper bound of $R(\delta, \theta^{\text{true}})$ (all the points occur above the line $y = x$). Note that, as the size of the true dictionary becomes significantly larger than the size of the estimated dictionary, the points seem to trace a steeper line which means that the estimated risk becomes a looser upper bound for the true risk.

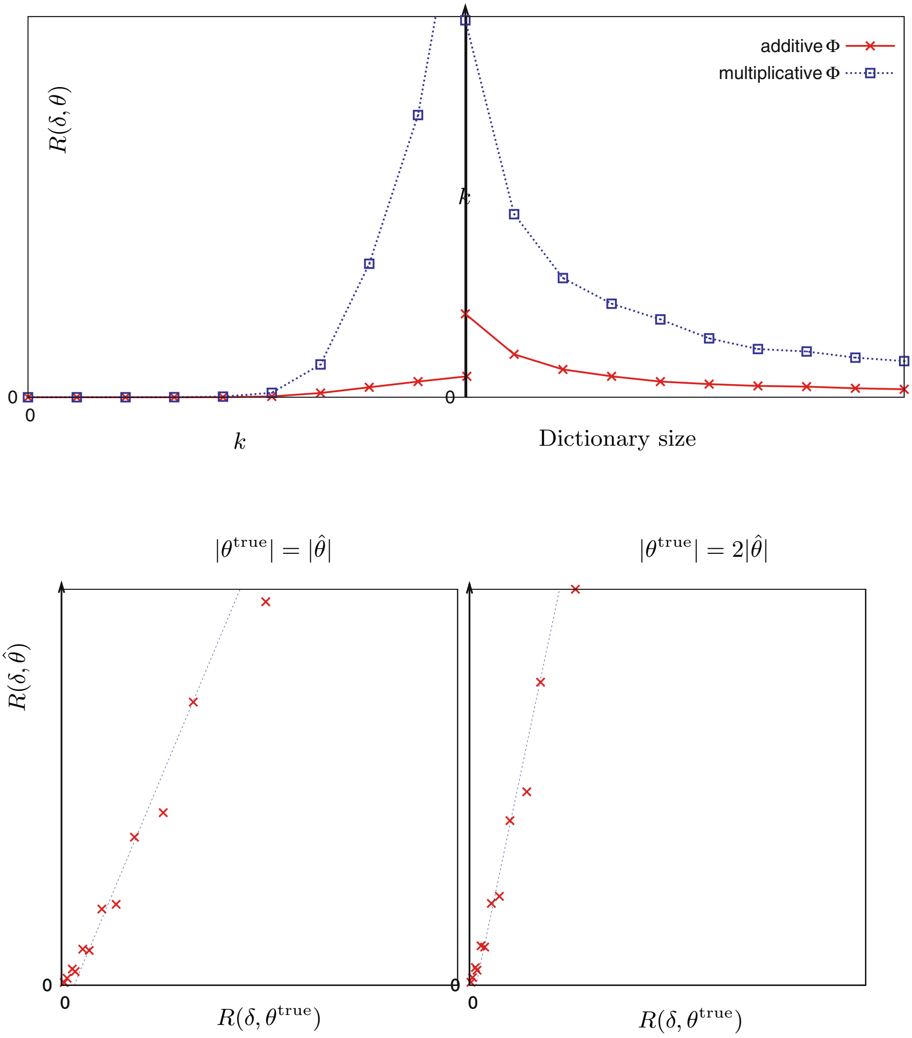


Fig. 1. The risk associated with different dictionaries and k values (left) and the relationship between the true risk and the estimated risk (right)

Using Mahalanobis Distance-Based Record Linkage for Disclosure Risk Assessment

Vicenç Torra[1], John M. Abowd[2], and Josep Domingo-Ferrer[3]

[1] IIIA-CSIC, Campus UAB, E-08193 Bellaterra, Catalonia
vtorra@iiia.csic.es

[2] Edmund Ezra Day Professor of Industrial and Labor Relations, Director, Cornell Institute for Social and Economic Research (CISER), 391 Pine Tree Road, Ithaca, NY 14850, USA
john.abowd@cornell.edu

[3] Universitat Rovira i Virgili, Dept. of Computer Engineering and Maths, Av. Països Catalans 26, E-43007 Tarragona, Catalonia
josep.domingo@urv.cat

Abstract. Distance-based record linkage (DBRL) is a common approach to empirically assessing the disclosure risk in SDC-protected microdata. Usually, the Euclidean distance is used. In this paper, we explore the potential advantages of using the Mahalanobis distance for DBRL. We illustrate our point for partially synthetic microdata and show that, in some cases, Mahalanobis DBRL can yield a very high re-identification percentage, far superior to the one offered by other record linkage methods.

Keywords: Microdata protection, Distance-based record linkage, Mahalanobis distance.

1 Introduction

A microdata set V can be viewed as a file with n records, where each record contains p attributes on an individual respondent. The attributes in the original unprotected dataset can be classified in four categories which are not necessarily disjoint:

- *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 in V have been removed/encrypted.
- *Quasi-identifiers.* Borrowing the definition from [3,13], a quasi-identifier is a set of attributes in V that, in combination, can be linked with external information to re-identify (some of) the respondents to whom (some of) the records in V refer. Examples of quasi-identifier attributes are birth date, gender, job, zipcode, etc. Unlike identifiers, quasi-identifiers cannot be removed from V. The reason is that any attribute in V potentially belongs to a quasi-identifier (depending on the external data sources available to the user of V).

J. Domingo-Ferrer and L. Franconi (Eds.): PSD 2006, LNCS 4302, pp. 233–242, 2006.

Thus one would need to remove all attributes (!) to make sure that the dataset no longer contains quasi-identifiers.

- *Confidential outcome attributes.* These are attributes which contain sensitive information on the respondent. Examples are salary, religion, political affiliation, health condition, etc.
- *Non-confidential outcome attributes.* Those are attributes which contain non-sensitive information on the respondent. Note that attributes of this kind cannot be neglected when protecting a dataset, because they can be part of a quasi-identifier. For instance, "Job" and "Town of residence" may reasonably be considered non-confidential outcome attributes, but their combination can be a quasi-identifier, because everyone knows who is the doctor in a small village.

Disclosure risk assessment is needed to measure the safety in a masked microdata being considered for release. The standard procedure is to use quasi-identifier attributes to perform record linkage between the masked dataset and an external identified data source. Each correctly linked pair yields a re-identification. To be more specific, the disclosure model considered in this paper is depicted in Figure 1 and is described next:

- We assume that the released microdata set (on the right-hand side in Figure 1) contains records with quasi-identifier attributes Y' and confidential outcome attributes X. Attributes Y' are masked, synthetic or partially synthetic versions of original quasi-identifier attributes.
- A snooper has obtained an external identified microdata set (on the left-hand side in Figure 1) which consists of one or several identifier attributes Id and several quasi-identifier attributes Y. Attributes Y are original (unmasked) versions of attributes Y' in the released dataset.
- The snooper attempts to link records in the external identified dataset with records in the released masked dataset. Linkage is done by matching quasi-identifier attributes Y and Y'. The snooper's goal is to pair identifier values with confidential attribute values (*e.g.* to pair citizens' names with health conditions).

1.1 Contribution and Plan of This Paper

We offer here an empirical comparison of various record linkage methods for re-identification. The masked datasets have been generated using the IPSO family of partially synthetic data generators [2] in the same way described in [9]. The range of record linkage methods tried is broader than in [9] and includes distance-based record linkage (DBRL) based on the Mahalanobis distance. This latter method yields surprising good results when there are strongly correlated attributes among in the quasi-identifiers.

Section 2 describes the record linkage methods used. The IPSO synthetic data generators are briefly recalled in Section 3. Section 4 specifies the two test datasets used as original datasets in the empirical study. Section 5 describes the experiments that were carried out. Conclusions are drawn in Section 6.

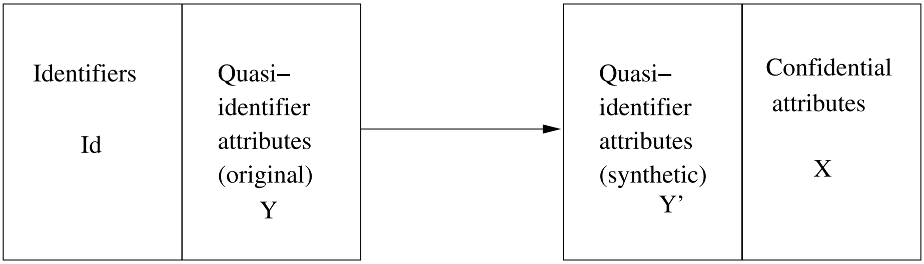

Identifiers
Id
Quasi–identifier attributes (original)
Y
Quasi–identifier attributes (synthetic)
Y'
Confidential attributes
X

Fig. 1. Re-identification scenario

2 Record Linkage Methods Used

We list below the record linkage methods implemented. For additional details and notation see [9,8]. In what follows, when the distance between pairs of records (a, b) where $a \in A$ and $b \in B$ is considered, we assume that files A and B are defined, respectively, on attributes $V_1^A, \ldots, V_n^A$ and $V_1^B, \ldots, V_n^B$. Accordingly, the actual values of a and b are, respectively, $a = (V_1^A(a), \ldots, V_n^A(a))$ and $b = (V_1^B(b), \ldots, V_n^B(b))$. The following record linkage methods were considered:

DBRL1: Attribute-standardizing implementation of distance-based record linkage. The Euclidean distance was used. Accordingly, given the notation for a and b given above, the distance between a and b is defined by:

$$d(a,b)^2 = \sum_{i=1}^{n} \left(\frac{V_i^A(a) - \bar{V_i^A}}{\sigma(V_i^A)} - \frac{V_i^B(b) - \bar{V_i^B}}{\sigma(V_i^B)}\right)^2$$

DBRL2: Distance-standardizing implementation of distance-based record linkage. The Euclidean distance was used. Therefore, the distance between a and b is defined by:

$$d(a,b)^2 = \sum_{i=1}^{n} \left(\frac{V_i^A(a) - V_i^B(b)}{\sigma(V_i^A - V_i^B)}\right)^2$$

DBRLM: Distance-based record linkage using the Mahalanobis distance. That is:

$$d(a,b)^2 = (a-b)'[Var(V^A) + Var(V^B) - 2Cov(V^A, V^B)]^{-1}(a-b)$$

where $Var(V^A)$ is the variance of attributes V^A, $Var(V^B)$ is the variance of attributes V^B and $Cov(V^A, V^B)$ is the covariance between attributes V^A and V^B.

The computation of $Cov(V^A, V^B)$ poses one difficulty: how records in A are lined up with records in B to compute the covariances. Two approaches can be considered:

- In a worst case scenario, it would be possible to know the correct links (a, b). Therefore, the covariance of attributes might be computed with the correct alignment between records.
- It is not possible to know a priori which are the correct matches between pairs of records. Therefore, any pair of records (a, b) are feasible. If any pair of records (a, b) are considered, the covariance is zero.

The re-identification using Mahalanobis distance with the first approach for computing the covariance will be denoted by DBRLM-COV. The second approach will be denoted by DBRLM-COV0.

KDBRL: Distance-based record linkage using a Kernel distance. That is, instead of computing distances between records (a, b) in the original n dimensional space, records are compared in a higher dimensional space H. Thus, let $\Phi(x)$ be the mapping of x into the higher space. Then, the distance between records a and b in H is defined as follows:

$$d(a,b)^2 = ||\Phi(a) - \Phi(b)||^2 = (\Phi(a) - \Phi(b))^2 =$$

$$= \Phi(a) \cdot \Phi(a) - 2\Phi(a) \cdot \Phi(b) + \Phi(b) \cdot \Phi(b) = K(a,a) - 2K(a,b) + K(b,b)$$

where K is a kernel function (*i.e.*, $K(a,b) = \Phi(a) \cdot \Phi(b)$).

We have considered polynomial kernels $K(x,y) = (1 + x \cdot y)^d$ for $d > 1$. With $d = 1$, the kernel record-linkage corresponds to the distance-based record linkage with the Euclidean distance.

Taking all this into account, the distance between a and b is defined as:

$$d(a,b)^2 = K(a,a) - 2K(a,b) + K(b,b)$$

with a kernel function K.

PRL: Probabilistic record linkage. The method is based on [10] and [11]. Our implementation follows [14].

3 The IPSO Synthetic Data Generators

Three variants of a procedure called Information Preserving Statistical Obfuscation (IPSO) are proposed in [2]. The basic form of IPSO will be called here Method A. Informally, suppose two sets of attributes X and Y, where the former are the confidential outcome attributes and the latter are quasi-identifier attributes. Then X are taken as independent and Y as dependent attributes. A multiple regression of Y on X is computed and fitted Y'_A attributes are computed. Finally, attributes X and Y'_A are released in place of X and Y.

In the above setting, conditional on the specific confidential attributes x_i, the quasi-identifier attributes Y_i are assumed to follow a multivariate normal distribution with covariance matrix $\Sigma = \{\sigma_{jk}\}$ and a mean vector $x_i B$, where B is the matrix of regression coefficients. Let $\hat{B}$ and $\hat{\Sigma}$ be the maximum likelihood estimates of B and Σ derived from the complete dataset (y, x). If a user fits a

Table 1. Re-identification experiments using dataset "Census" and methods IPSO-A, IPSO-B and IPSO-C

Quasi-identifier in external **A**	Quasi-identifier in released **B**
$v7, v12$	$v7_A^{S1}, v12_A^{S1}$
$v4, v7, v11, v12$	$v4_A^{S1}, v7_A^{S1}, v11_A^{S1}, v12_A^{S1}$
$v4, v7, v12, v13$	$v4_A^{S1}, v7_A^{S1}, v12_A^{S1}, v13_A^{S1}$
$v4, v7, v11, v12, v13$	$v4_A^{S1}, v7_A^{S1}, v11_A^{S1}, v12_A^{S1}, v13_A^{S1}$
$v1, v3, v4, v6, v7,\ v9, v11, v12, v13$	$v9_A^{S1}, v11_A^{S1}, v12_A^{S1}, v13_A^{S1},\ v1_A^{S1}, v3_A^{S1}, v4_A^{S1}, v6_A^{S1}, v7_A^{S1}$
$v7, v12$	$v7_A^{S2}, v12_A^{S2}$
$v4, v13$	$v4_A^{S2}, v13_A^{S2}$
$v7, v12, v13$	$v7_A^{S2}, v12_A^{S2}, v13_A^{S2}$
$v4, v7, v12, v13$	$v4_A^{S2}, v7_A^{S2}, v12_A^{S2}, v13_A^{S2}$

multiple regression model to (y'_A, x), she will get estimates $\hat{B}_A$ and $\hat{\Sigma}_A$ which, in general, are different from the estimates $\hat{B}$ and $\hat{\Sigma}$ obtained when fitting the model to the original data (y, x). IPSO Method B, IPSO-B, modifies y'_A into y'_B in such a way that the estimate $\hat{B}_B$ obtained by multiple linear regression from (y'_B, x) satisfies $\hat{B}_B = \hat{B}$.

A more ambitious goal is to come up with a data matrix y'_C such that, when a multivariate multiple regression model is fitted to (y'_C, x), *both* sufficient statistics $\hat{B}$ and $\hat{\Sigma}$ obtained on the original data (y, x) are preserved. This is done by the third IPSO method, IPSO-C.

4 The Test Datasets

We have used two reference datasets [1] used in the European project CASC:

1. The "Census" dataset contains 1080 records with 13 numerical attributes labeled $v1$ to $v13$. This dataset was used in CASC and in several other papers. [5,4,15,12,7,6,9].
2. The "EIA" dataset contains 4092 records with 15 attributes. The first five attributes are categorical and will not be used. We restrict attention to the last 10 numerical attributes, which will be labeled $v1$ to $v10$. This dataset was used in CASC, in [4,6,9] and partially in [12] (an undocumented subset of 1080 records from "EIA", called "Creta" dataset, was used in the latter paper).

5 Experiments

We have considered the datafiles "Census" and "EIA", with the same scenarios and the same re-identification experiments we used in [9]. In short, re-identification experiments are applied to pairs of external and released files using subsets of quasi-identifiers. In scenario $S1$ for the dataset "Census" there are nine quasi-identifiers; in scenario $S2$ for "Census" there are four quasi-identifiers.

Table 2. Re-identification experiments using dataset ”EIA” and methods IPSO-A, IPSO-B and IPSO-C

Quasi-identifier in external **A**	Quasi-identifier in released **B**
$v1$	$v1_A$
$v1, v7, v8$	$v1_A, v7_A, v8_A$
$v1, v2, v7, v8, v9$	$v1_A, v2_A, v7_A, v8_A, v9_A$
$v1$	$v1_B$
$v1, v7, v8$	$v1_B, v7_B, v8_B$
$v1, v2, v7, v8, v9$	$v1_B, v2_B, v7_B, v8_B, v9_B$
$v1$	$v1_C$
$v1, v7, v8$	$v1_C, v7_C, v8_C$
$v1, v2, v7, v8, v9$	$v1_C, v2_C, v7_C, v8_C, v9_C$

Table 3. Re-identification experiments using dataset ”Census” and method IPSO-A. Results in number of correct re-identifications over an overall number of 1080 records. DBRL1: attribute-standardizing implementation of distance-based record linkage (DBRL); DBRL2: distance-standardizing implementation of DBRL; DBRLM-COV and DBRLM-COV0: distance-based record linkage using Mahalanobis distance (covariances computed using the appropriate alingment or covariances set to zero); KDBRL: distance-based record linkage with kernel distance (polynomic kernel with d=2); PRL: probabilistic record linkage.

DBRL1	DBRL2	DBRLM-COV0	DBRLM-COV	KDBRL	PRL
145	133	135	123	146	133
91	75	126	60	89	82
95	87	137	66	94	103
98	87	129	62	97	86
23	40	123	67	24	97
104	92	93	84	100	92
59	65	63	57	61	65
94	85	89	68	91	86
109	104	106	44	106	103

For “EIA” there is a single scenario with five quasi-identifier attributes highly correlated with the rest of attributes. Released files (see [9,8] for details) were generated using the synthetic data generators IPSO-A, IPSO-B and IPSO-C. Table 1 lists the sets of quasi-identifiers considered for the “Census” data in the case of data generated using IPSO-A. Analogous sets of quasi-identifiers (vi_B^{S1} and vi_C^{S1} instead of vi_A^{S1}) were considered for the other IPSO-B and IPSO-C methods. Table 2 contains similar information corresponding to “EIA” datasets.

Note that in this paper only experiments with files sharing attributes have been considered.

Table 4. Re-identification experiments using dataset ”Census” and method IPSO-B. Results in number of correct re-identifications over an overall number of 1080 records. DBRL1: attribute-standardizing implementation of distance-based record linkage (DBRL); DBRL2: distance-standardizing implementation of DBRL; DBRLM-COV and DBRLM-COV0: distance-based record linkage using Mahalanobis distance (covariances computed using the appropriate alingment or covariances set to zero); KDBRL: distance-based record linkage with kernel distance (polynomic kernel with d=2); PRL: probabilistic record linkage.

DBRL1	DBRL2	DBRLM-COV0	DBRLM-COV	KDBRL	PRL
146	133	135	123	133	133
89	75	126	61	73	81
95	86	138	66	87	103
97	85	130	62	86	86
23	40	123	63	5	94
104	92	93	83	92	92
59	65	63	57	65	65
94	85	89	68	85	86
109	104	106	44	103	103

Table 5. Re-identification experiments using dataset ”Census” and method IPSO-C. Results in number of correct re-identifications over an overall number of 1080 records. DBRL1: attribute-standardizing implementation of distance-based record linkage (DBRL); DBRL2: distance-standardizing implementation of DBRL; DBRLM-COV and DBRLM-COV0: distance-based record linkage using Mahalanobis distance (covariances computed using the appropriate alingment or covariances set to zero); KDBRL: distance-based record linkage with kernel distance (polynomic kernel with d=2); PRL: probabilistic record linkage.

DBRL1	DBRL2	DBRLM-COV0	DBRLM-COV	KDBRL	PRL
34	34	34	34	33	34
37	37	42	19	39	32
24	24	24	11	23	23
39	39	44	17	40	36
24	24	50	11	25	43
47	47	47	44	49	48
19	19	20	20	19	18
40	40	34	34	41	37
35	35	41	41	32	33

The results of the experiments considered for the “Census” data for methods IPSO-A, IPSO-B and IPSO-C are given in Tables 3, 4 and 5. The results of the experiments using the file “EIA” are given in Table 6.

Table 6. Re-identification experiments using dataset ”EIA” and methods IPSO-A, IPSO-B and IPSO-C. Results in number of correct re-identifications over an overall number of 4092 records. DBRL1: attribute-standardizing implementation of distance-based record linkage (DBRL); DBRL2: distance-standardizing implementation of DBRL; DBRLM-COV and DBRLM-COV0: distance-based record linkage using Mahalanobis distance (covariances computed using the appropriate alingment or covariances set to zero); KDBRL: distance-based record linkage with kernel distance (polynomic kernel with d=2); PRL: probabilistic record linkage.

DBRL1	DBRL2	DBRLM-COV0	DBRLM-COV	KDBRL	PRL
14	9	9	9	14	8
16	15	18	9	16	16
65	121	3206	143	63	159
14	9	9	9	14	8
17	15	18	8	17	16
65	120	3194	135	62	159
11	11	11	11	11	10
6	6	14	8	6	5
53	53	773	46	54	93

6 Conclusions

Conclusions in [9] with respect to distance-based and probabilistic record linkage are also applicable here. In relation to the additional methods considered here we should point out that:

- Distance-based record linkage based on Mahalanobis distance achieves the highest number of re-identifications (3206 over 4092 records) in the case of the EIA datafile when the synthetic data generator is IPSO-A and all quasi-identifiers are considered. This corresponds to the re-identification of 78.3% of the records. Similarly, 3194 (78.05%) re-identifications are obtained for IPSO-B data. In the case of IPSO-C, the best performance is 773 re-identifications (which corresponds to 18.9% of the records).
- With respect to distance-based record linkage based on Mahalanobis distance, DBRLM-COV0 (*i.e.*, covariances between attributes V^A and V^B are set to zero) has a better performance than DBRLM-COV.
- The distance-based record linkage based on the kernel distance leads to results equivalents to the other distance-based methods. Only in one experiment does this method outperform the other ones. This experiment corresponds to “Census” data with synthetic data generated with IPSO-A (first experiment with two variables). In this case, 146 records are re-identified.

One possible explanation for the different behaviour of DBRLM-COV0 in ”Census” and ”EIA” is that quasi-identifiers in the latter dataset are more highly correlated.

In the experiments performed here, re-identification consists of finding the links between the original and the synthetic data. This corresponds to the assumption that the snooper knows a subset of the original data and tries to link such data with the synthetic data in order to disclose sensitive attributes. This re-identification is directed following the scheme in Figure 1. This re-identification scheme differs from the scheme considered in [9]. There, synthetic data was re-identified back to the original source data. The change in the scheme does not reveal any substantial differences among the methods already considered in [9]. The following results illustrate the minor differences:

- DBRL1 for "Census" data in scenario $S1$ on the data generated with IPSO-A leads to 144, 85, 104, 79 and 36 records re-identified when using the scheme in [9]. Instead, the current scheme leads to 145, 91, 95, 98 and 23, respectively.
- DBRL1 for "EIA" data on the data generated with IPSO-A, the previous scheme leads to 10, 23 and 65 re-identifications while the new one yields 14, 16 and 65 re-identifications, respectively.

Acknowledgments

We acknowledge partial support by the Government of Catalonia under grant 2005 SGR 00446, by the Spanish Ministry of Science and Education under project SEG2004-04352-C04-01/02 "PROPRIETAS" and by Cornell University under contracts no. 47632-10042 and 10043. Abowd acknowledges support from NSF-ITR grant SES-0427889 to Cornell University.

References

1. R. Brand, J. Domingo-Ferrer, and J. M. Mateo-Sanz. Reference data sets to test and compare sdc methods for protection of numerical microdata, 2002. European Project IST-2000-25069 CASC, http://neon.vb.cbs.nl/casc.
2. J. Burridge. Information preserving statistical obfuscation. *Statistics and Computing*, 13:321–327, 2003.
3. T. Dalenius. Finding a needle in a haystack - or identifying anonymous census records. *Journal of Official Statistics*, 2(3):329–336, 1986.
4. R. Dandekar, J. Domingo-Ferrer, and F. Sebé. Lhs-based hybrid microdata vs rank swapping and microaggregation for numeric microdata protection. In J. Domingo-Ferrer, editor, *Inference Control in Statistical Databases*, volume 2316 of *Lecture Notes in Computer Science*, pages 153–162, Berlin Heidelberg, 2002. Springer.
5. J. Domingo-Ferrer, J. M. Mateo-Sanz, and V. Torra. Comparing sdc methods for microdata on the basis of information loss and disclosure risk. In *Pre-proceedings of ETK-NTTS'2001 (vol. 2)*, pages 807–826, Luxemburg, 2001. Eurostat.
6. J. Domingo-Ferrer, F. Sebé, and A. Solanas. A polynomial-time approximation to optimal multivariate microaggregation. *Manuscript*, 2005.
7. J. Domingo-Ferrer and V. Torra. Ordinal, continuous and heterogenerous k-anonymity through microaggregation. *Data Mining and Knowledge Discovery*, 11(2):195–212, 2005.
8. J. Domingo-Ferrer, V. Torra, J. M. Mateo-Sanz, and F. Sebé. Research data center-based confidentiality research: Systematic measures of re-identification risk based on the probabilistic links of the partially synthetic data back to the original microdata, final report. Technical report, Rovira i Virgili University and IIIA-CSIC, 2005.

9. J. Domingo-Ferrer, V. Torra, J. M. Mateo-Sanz, and F. Sebé. Empirical disclosure risk assessment of the ipso synthetic data generators. In *Monographs in Official Statistics-Work Session On Statistical Data Confidentiality*, pages 227–238, Luxemburg, 2006. Eurostat.
10. I. P. Fellegi and A. B. Sunter. A theory for record linkage. *Journal of the American Statistical Association*, 64(328):1183–1210, 1969.
11. M. A. Jaro. Advances in record-linkage methodology as applied to matching the 1985 census of tampa, florida. *Journal of the American Statistical Association*, 84(406):414–420, 1989.
12. M. Laszlo and S. Mukherjee. Minimum spanning tree partitioning algorithm for microaggregation. *IEEE Transactions on Knowledge and Data Engineering*, 17(7):902–911, 2005.
13. P. Samarati. Protecting respondents' identities in microdata release. *IEEE Transactions on Knowledge and Data Engineering*, 13(6):1010–1027, 2001.
14. V. Torra and J. Domingo-Ferrer. Record linkage methods for multidatabase data mining. In V. Torra, editor, *Information Fusion in Data Mining*, pages 101–132, Berlin, 2003. Springer.
15. W. E. Yancey, W. E. Winkler, and R. H. Creecy. Disclosure risk assessment in perturbative microdata protection. In J. Domingo-Ferrer, editor, *Inference Control in Statistical Databases*, volume 2316 of *Lecture Notes in Computer Science*, pages 135–152, Berlin Heidelberg, 2002. Springer.

Improving Individual Risk Estimators

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. The release of survey microdata files requires a preliminary assessment of the disclosure risk of the data. Record-level risk measures can be useful for "local" protection (e.g. partially synthetic data [21], or local suppression [25]), and are also used in [22] and [16] to produce global risk measures [13] useful to assess data release. Whereas different proposals to estimating such risk measures are available in the literature, so far only a few attempts have been targeted to the evaluation of the statistical properties of these estimators. In this paper we pursue a simulation study that aims to evaluate the statistical properties of risk estimators. Besides presenting results about the Benedetti-Franconi individual risk estimator (see [11]), we also propose a strategy to produce improved risk estimates, and assess the latter by simulation.

The problem of estimating per record reidentification risk enjoys many similarities with that of small area estimation (see [19]): we propose to introduce external information, arising from a previous census, in risk estimation. To achieve this we consider a simple strategy, namely Structure Preserving Estimation (SPREE) of Purcell and Kish [18], and show by simulation that this procedure provides better estimates of the individual risk of reidentification disclosure, especially for records whose risk is high.

Keywords: Bayesian hierarchical models, disclosure, per record risk, SPREE, simulation study.

1 Introduction

The release of microdata files requires a preliminary assessment of the risk of disclosure of the data to be released. Sometimes a record-level measure of risk can be useful, that can be exploited to protect the data selectively; moreover (e.g. [16,22]) such a measure can be used to build an overall measure of risk for

J. Domingo-Ferrer and L. Franconi (Eds.): PSD 2006, LNCS 4302, pp. 243–256, 2006.

the whole data file. Whereas different proposals to estimating the risk of disclosure are available in the literature, so far only a few attempts have been targeted to the evaluation of the statistical properties of the estimators adopted. Giving at least an idea of the bias and MSE of the risk estimators is clearly crucial for risk assessment and subsequent data protection. In this paper we pursue a simulation study to evaluate the statistical properties of the Benedetti-Franconi individual (BF) risk estimator (see [11]). Such estimator is a function of the sample frequency of the cells in the contingency table built by cross tabulating the key variables that permit reidentification under a certain disclosure scenario (see [11] or [16] for details). We anticipate that the BF estimator exhibits appreciable underestimation of small population cells and negligible bias for large ones.

The problem of estimating record level measures of reidentification risk enjoys many similarities with that of small area estimation (a comprehensive account on the subject is in [19]). For this reason we believe that improvements in risk assessment can be achieved by exploiting these analogies. Especially when observing very low sample cell frequencies, etc., in order to obtain improved estimators it is often convenient to introduce external information. Given the type of key variables that typically characterize the disclosure scenario, we exploited the information conveyed by a previous census. As the census questionnaire usually collects this kind of variables, it is possible to consider the census contingency table built by cross classifying the key variables as external information. We consider a simple strategy, namely Structure Preserving Estimation (SPREE) (Purcell and Kish, 1980, [18]), to produce improved risk estimators.

Our simulation study is organized as follows: we simulate 1,000 synthetic samples 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 information that we borrow from the 1991 census refers to the structure of association in the contingency table based on the variables that we use to estimate the individual risk. The same contingency table is therefore considered at two different time occasions. 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.

The procedure analysed in the paper is shown to provide better estimates of the population cell frequencies that we try to infer from the sample in the process of risk estimation, and by consequence we expect it to lead to improved risk estimates. In Section 4 we propose different strategies for risk estimation and compare them by simulation.

2 The Reidentification Risk

The definition of disclosure that we adopt is based on the concept of reidentification (e.g. [4,9,7]) and is appropriate for the release of data arising from social surveys. A disclosure is a correct reidentification achieved by matching a target individual in a sample with an available list of units which contains individual identifiers, such as name and address (see [25]). The variables that can be used for reidentification are referred to as key variables. These are a major ingredient of the disclosure scenario defined by the Agency that is releasing the data. For social surveys, a scenario under which the key variables represent information available in the public registers collected under the current legal regulation is sometimes defined. In this case the key variables are selected among public variables that are available for the population and also present in the file to be released. These are usually categorical key variables (region of residence, sex, age, marital status, and so on), that can also be derived from brief social contact with records. It is assumed that potential intruders do not have any further information about records, for instance, whether or not they belong to the sample. If the sampling weights are released, under some circumstances these might give additional information for reidentification; for instance, very large enterprises are usually sampled with probability one. For social surveys on individuals, the latter is true if the sampling design weights are approximately constant at a geographical detail finer than the one defined by the corresponding key variables; however the effect of calibration and editing will act as a confounder. We also remark that sampling weights could be protected prior to data release. For the above reasons, we adopt a definition of risk that does not take this aspect into account.

In such a framework, the risk can be defined as a function of the cells of the contingency table built by cross-tabulating the key variables in the population. Records presenting combinations of key variables that are unusual or rare in the population have clearly a high disclosure risk, whereas rare or even unique combinations in the sample do not necessarily correspond to high risk individuals. The risk measure is thus defined at the record level; more precisely, it is a cell-specific measure, as it only depends on cells or combinations of the key variables introduced in the disclosure scenario.

To measure the disclosure risk under this framework we therefore focus on the K cells $\mathcal{C}_1, \ldots, \mathcal{C}_K$ of the contingency table above. We will consider the frequencies $\boldsymbol{F} = (F_1, \ldots, F_K)$ and $\boldsymbol{f} = (f_1, \ldots, f_K)$ in the population and sample table, respectively. In particular, the population counts $\boldsymbol{F}$ are usually unknown. In estimating most risk measures, inference about $\boldsymbol{F}$ is made, possibly only for the cells $\mathcal{C}_k$ with small sample frequency: for instance, [24] and [23] focus on sample uniques, while [3,8,10,20] also consider non unique cells in the sample.

The problem of risk estimation is closely related to that of estimating the elements of the vector $\boldsymbol{F}$ of population cell frequencies. In the next sections we consider different approaches to risk estimation and exploit the interplay between the two estimation problems.

3 The Benedetti-Franconi Risk Estimator

Benedetti and Franconi [2] introduced a framework to estimate the reidentification risk. A routine for computing the Benedetti-Franconi (BF) individual risk is currently implemented in the software μ-Argus, developed under the European Union project CASC on Computational Aspects of Statistical Confidentiality (see for instance [12]).

If one assumes, as described in Section 2, that individuals belonging to the same cell $\mathcal{C}_k$ are exchangeable for the intruder, then the probability of reidentification of individual i in cell $\mathcal{C}_k$ when F_k individuals of the population known to belong to it is $1/F_k$, $k = 1, \ldots, K$.

In order to infer the population frequency F_k of a given combination from its sample frequency f_k, a Bayesian approach is pursued, in that the posterior distribution of F_k given $\boldsymbol{f}$ is exploited. The risk is then defined as the expected value of $1/F_k$ under this distribution. For a general Bayesian formulation of reidentification, see [9,15,10].

As shown in [22] and [16], the model originally proposed by Benedetti and Franconi [2] corresponds to the following hierarchical model:

$$\begin{aligned} F_k|\pi_k &\sim \text{Poisson}(N\pi_k),\ F_k = 0, 1, \ldots, \\ f_k|F_k, \pi_k, p_k &\sim \text{binom}(F_k, p_k),\ f_k = 0, 1, \ldots, F_k\ , \end{aligned} \tag{1}$$

independently across cells. Furthermore, π_k in (1) above is assumed to follow an improper prior distribution, proportional to $1/\pi_k$, $k = 1, \ldots, K$.

The parameters p_k, each one representing the probability that a member of population cell $\mathcal{C}_k$ falls into the sample, are not further modelled. Following a kind of Empirical Bayesian approach, Benedetti and Franconi propose plugging-in the quantity

$$\hat{p}_k = \frac{f_k}{\hat{F}_k^{\mathrm{D}}} \tag{2}$$

in the final modelling step,

$$\hat{F}_k^{\mathrm{D}} = \sum_{i \in \mathcal{C}_k} w_i$$

representing the direct design unbiased estimator of the population counts F_k based on the sampling design weights w_i. Under model (1), the posterior distribution of F_k given $\boldsymbol{f}$ only depends on the corresponding sample cell frequency f_k (see [10]), and is of negative binomial type [22,16]:

$$\Pr\left(F_k = h | f_k = j\right) = \binom{h-1}{j-1} p_k^j \left(1 - p_k\right)^{h-j},\ \ h \geq j\ .$$

Consequently [2,16,11],

$$\begin{aligned} \hat{r}_k^{\mathrm{BF}} &= \mathrm{E}\left(\frac{1}{F_k}\middle| f_k\right) = \int_0^\infty \left\{\frac{\hat{p}_k \exp\left(-t\right)}{1 - \hat{q}_k \exp\left(-t\right)}\right\}^{f_k} dt \\ &= \frac{\hat{p}_k^{f_k}}{f_k}\ {}_2F_1(f_k, f_k; f_k + 1; \hat{q}_k), \end{aligned} \tag{3}$$

where $\hat{q}_k = 1 - \hat{p}_k$ and ${}_2F_1(a, b; c; z)$ is the Hypergeometric function (see [1]).

The estimator (3) only makes use of the sampling cell frequency f_k and the sampling design weights w_i of units belonging to cell $\mathcal{C}_k$. Given the marginal nature of model (1) on which it is based and the way the hyperparameters are treated, the structure of the contingency table that produces the cells on which the risk is estimated is not exploited. This is partially attenuated when calibrated sampling design weights [5] are used in (2), which is the case with most ISTAT surveys. Calibration (see [5]) ensures that estimates over certain domains return the corresponding known population totals. The variables on which calibration is normally performed are typically a subset of the key variables introduced in the disclosure scenario. Therefore partial account of the structure of association exhibited in the population by the key variables is made when calibration weights are used.

4 Alternative Estimators

As mentioned, the BF estimator (3) makes use of a calibration estimator, $\hat{F}_k^{\mathrm{D}}$, as a first-step estimate of the population cell frequency F_k. The calibration estimator may be very poor, especially for those classes whose population is very low. Being characterised by a potentially high risk of disclosure, these cells are clearly the ones of main interest in our framework. In [6] an experiment was conducted over 600 simulated samples to assess the performance of the individual risk of disclosure. However only the cells that are present in all samples were assessed; the highest risk cells, corresponding to small population frequencies, were therefore excluded from the study. In Section 5 we pursue a simulation experiment to assess the BF risk estimator. In particular we consider the behaviour of the BF risk estimator in small population cells.

Small area estimators have been extensively applied to reduce the MSE of direct estimators by means of external information and 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) proposed by Purcell and Kish [18]. For tabular data arising from cross-classification of categorical variables, Purcell and Kish [18] propose that the *association structure* obtained from an administrative or a census source can be exploited to improve the estimation of counts. The association structure derives from a frequency table known at a previous time $t_0 = t - L$ that completely describes the relationship among the variables that define the cross-tabulation. The given association structure is then updated on the basis of current information at time t on the (partial) association between the variables present in the *allocation structure*.

The *allocation structure* is usually represented by margins of the current frequency table; these are estimated from the survey data or, when auxiliary variables are available, obtained from administrative sources. Typically, counts on classes defined by sex and age can be obtained by administrative records, so that these are often used as auxiliary variables.

Aiming to improve the BF estimator especially for low counts, we propose two alternative risk estimators, both based on the SPREE of F_k. These are described in Section 4.2; in the next section we first introduce the SPREE.

4.1 The Structure Preserving Estimator

Any multi-way table can be reduced to three-way by properly re-defining the classification. Let us denote by d the geographical or administrative domain (or partition), by h the classification given by the auxiliary variables (sex and age in the above example) and by i the classification given by the survey-variables. Let X_{dhi} be the association structure, i.e. the known table at previous time $t_0 = t - L$. Finally, define by F_{dhi} the current table to be estimated and by m the allocation structure, i.e. the updated margins.

Note that while in small area estimation problems F_{dhi} is usually only a means to allow estimation of parameters of interest, that consist of margins $X_{d.i}$, in disclosure estimation F_{dhi} itself is the ultimate inferential goal.

The SPREE method consists in adjusting the X_{dhi} to agree with the updated information in m, while preserving the relationship 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 [18], explicit solutions exist only in trivial cases. In the general case, Iterative Proportional Fitting (IPF), which consists in iteratively adjusting the marginal constraints until convergence, is applied to obtain an approximate solution, denoted by $\hat{F}^{\text{SPREE}}$, to the above problem.

Depending on the information available, different specifications of m can be given. Here we consider the specification used in our application (See Section 5), namely $m = (\{\hat{F}_{.hi}\}, \{F_{dh.}\})$, where $\hat{F}_{.hi}$ are direct estimates and $F_{dh.}$come from administrative registers.

In our case the IPF has the following structure: at the first step the starting values are set equal to values in the association structure

$$ {}^{0}\hat{F}^{\text{SPREE}}_{dhi} = X_{dhi} \ \ . $$

At step k, cell counts are adjusted to the marginal constraints in two stages:

$$ {}^{(1)k}\hat{F}^{\text{SPREE}}_{dhi} = \frac{{}^{k-1}\hat{F}^{\text{SPREE}}_{dhi}}{{}^{k-1}\hat{F}^{\text{SPREE}}_{.hi}} \times \hat{F}_{.hi} \ \ , $$

and

$$ {}^{k}\hat{F}^{SPREE}_{dhi} = \frac{{}^{(1)k}\hat{F}^{SPREE}_{dhi}}{{}^{(1)}\hat{F}^{SPREE}_{dh.}} \times F_{dh.} \ \ ; $$

step k is repeated until convergence.

It can be shown (see [18]) that this methodology preserves 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 $\hat{F}^{\text{SPREE}}$ depends on the extent to which this equality holds for the data. For further details on SPREE see [18] and [26].

4.2 Two SPREE-Based Risk Estimators

Having observed that the estimand is $r_k = 1/F_k$ for nonempty sample cells, our first proposal simply estimates r_k by

$$\hat{r}_k^{\text{SPREE}} = \frac{1}{\hat{F}_k^{\text{SPREE}}} \ . \tag{4}$$

The naive estimator (4) is assessed by simulation in Section 6.

For our second proposal we refer to model (1), however replacing $\hat{F}_k^{\text{D}}$ by the small area estimator $\hat{F}_k^{\text{SPREE}}$ in (2) to obtain a first-step estimator of the cell probability p_k. For the first-step we propose using

$$\hat{\hat{p}}_k = f_k / \hat{F}_k^{\text{SPREE}} \ ,$$

thus obtaining the risk estimator

$$\hat{r}_k^{\text{BF SPREE}} = \frac{\hat{\hat{p}}_k^{f_k}}{f_k} \ {}_2F_1(f_k, f_k; f_k + 1; \hat{\hat{q}}_k) \ . \tag{5}$$

We expect that both (4) and (5) represent an improvement over the "standard" BF estimator (3). We also want to compare the naive estimator (4) with the model-based estimator (5).

5 Simulation Plan

To evaluate the statistical properties of the three estimators (3), (4) and (5), a simulation study was performed, in which samples were selected from a known real population.

In order to mimic as closely as possible a real situation, we applied a realistic sample strategy, namely Labour Force Survey (LFS); LFS is a very large survey whose sampling scheme is a standard design commonly applied to the main Italian social surveys. A total of 1,000 samples has been selected from the 2001 Italian Census data, comprising 6 Italian regions (Val d'Aosta, Piemonte, Toscana, Umbria, Campania, Molise). The population amounts to over 15 millions, whereas the effective sample size in terms of individuals is over 80,000. Choice of the above mentioned regions was motivated by several reasons: 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 larger risks of disclosure) and finally the substantial variation of their sampling rates. Note that sampling rates do vary highly among regions (ranging approximately from 0.004 to 0.03) since sample size is planned to guarantee a given target level in the precision of the estimates disseminated with the survey results. For Labour Force Survey, in particular, sample sizes are set to ensure stable regional quarterly estimates of the unemployment rates and to guarantee pre-fixed precision for yearly estimates at province level.

The LFS is based on a complex sample design with stratification of municipalities which are primary sampling units (PSU). In each sample municipality a systematic sample of households (secondary sampling units or SSU) is selected; each member of sampled households is included in the LFS sample. The stratification of PSU is carried out in each province (administrative areas inside regions) according to their dimension in terms of residents; municipalities whose size is bigger than a benchmark[1] represent one member-strata and are classified as Self-Representing Areas (SRA); these PSU are selected with probability one. For the other strata, containing Not Self-Representing Areas (NSRA), one municipality is selected with probability proportional to its size by means of the randomized systematic procedure first introduced by Madow [14].

The final weights, that are released to allow estimation of LFS quantities of interest, are obtained by applying to the basic weights (inverse selection probabilities) a calibration process that controls on known totals of sex and ages (see [5]). The main totals involved in the calibration process are defined by: sex by 14 age-classes at the regional level, sex by 5 age-classes at the province level, sex by 5 age-classes for the bigger municipalities. To reduce the simulation burden, we have calibrated only on sex by age at regional level.

The variables that we selected as key variables are region of residence, sex, age (in 20 classes), marital status (in 4 classes), education (in 5 classes). According to the Italian legislation, these variables contain information available from external registers; besides easy to get from brief social contact with the record.

Note that planning of sample size, which is not based on the characteristic of the individuals in the sample, is such that the sample dimension for the 4 800 cells of the classification on the key-variables under analysis may be very small or even empty. As the cell of the cross-tabulation does not represent a planned domain for LFS, we expect that especially the cells with smaller 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 application of the risk estimators based on SPREE, we needed also an association structure at a previous time. Complete information on it could be derived only from the previous census, relative to 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.

In the terminology of Section 4.1, the allocation structure has been defined as $m = (\{\hat{F}_{.hi}\}, \{F_{dh.}\})$, $F_{dh.}$ being the 2001 census counts of the marginal cross-table defined by: sex by age (classes indicated with h) by region (classes indicated with d). In practice these counts would come from updated administrative source. $\hat{F}_{.hi}$, instead, represents the calibration estimate of marginal cross-table defined by: education by marital status (classes denoted with i) by sex by age (classes denoted with h).

[1] Function of the sampling rate and the minimum number of interview to be performed in each PSU.

The availability of the target population from which to extract our samples allows us to assess the performance of the estimators, as clearly the estimand is known and equals $1/F_k$ for each k. Performance of risk estimates has been based on bias and relative root mean square error over the samples where the cell has been observed, as the risk is not defined (and not of interest) when sample cell is empty. A conditional analysis of the above measures has also been performed for given cell sample size ($f_k = 1, 2$).

For the smaller 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. As remarked in Section 7, 1,000 samples are likely too few for a definite assertion of the performances for the smaller cells, but can be useful to outline the expected pattern.

6 Results

We expected the BF estimator $\hat{r}_k^{\mathrm{BF}}$ to behave poorly for small population frequencies, that are the risky ones. Indeed $\hat{r}_k^{\mathrm{BF}}$ is based on a direct estimator of population cell size that is calibrated on much larger domains (see Section 5) than the ones we are using. We observed that the estimator depends mainly on the observed cell sample size f_k, so that risky cells cannot be effectively distinguished from the safe ones. As compared to the BF, the SPREE introduces a much larger variation of the estimates across cells, which makes the alternative estimators appealing to discriminate between safe and risky records. Fig. 1 reports an overall graphical assessment of the conditional behaviour of the three estimators over sample unique cells for all samples. To avoid plotting all the replications of sample uniques across the simulations, for each cell we plotted a summary of the estimates over our 1,000 samples. We show the minimum and maximum (grey dots) and the median (black dots) of the estimates over all eligible samples. Fig. 1 shows that both $\hat{r}_k^{\mathrm{BF\,SPREE}}$ and $\hat{r}_k^{\mathrm{SPREE}}$ are successful in differentially detecting risky and safe sample uniques, although $\hat{r}_k^{\mathrm{BF\,SPREE}}$ shows overestimation of risk for the sample unique records that correspond to large population cells. As already remarked, the BF estimator has not enough variation to allow for discrimination of sample uniques.

Fig. 2 shows that the BF estimator is also unconditionally largely biased for small cells in the population. In this case, the alternative estimates are also negatively biased, but the distribution is skewed toward zero. The bias decreases with F_k, as large cells in the population are likely to produce large cells in the sample. When we focus on very small population cells (up to 10-15 records), the BF estimator tends to underestimate the true risk (See table 1 in the Appendix for $F_k = 2$ and 5); the alternative estimators only have negative bias for population uniques, although they exhibit a very large variation. For moderate population cell size, both the BF and the BF SPREE estimators have positive bias, which is negligible for the SPREE estimator. All the estimators have negligible bias as the population cell size gets larger. The bias for all the estimators

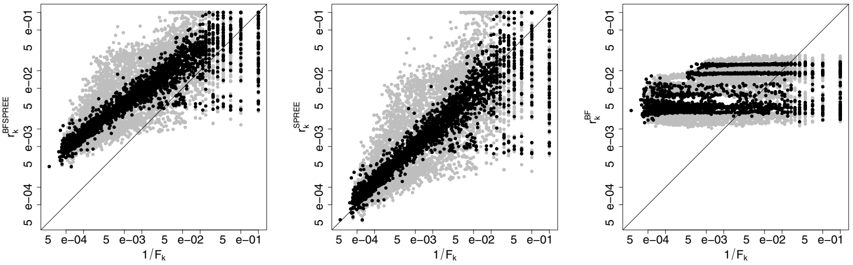


Fig. 1. Assessment of the three estimators for sample unique cells over all the simulated samples. Per cell minimum and maximum (grey dots) and median (black dots) of the estimates over all samples are plotted.

shows dependence, especially noticeable for the BF, on the sampling rates. As mentioned in Section 5, the sampling fractions vary across regions, being higher for Molise and Val d'Aosta than for the other regions.

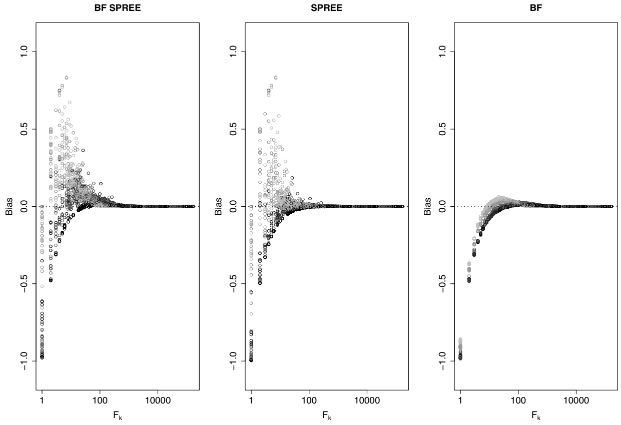


Fig. 2. Bias of the three estimators vs population cell frequency. Darker grey levels indicate regions with smaller sampling fractions.

We next evaluated the relative root MSE of the estimators,

$$\mathrm{RRMSE}\left(\hat{r}_k\right) = \frac{\sqrt{\mathrm{E}\left(\hat{r}_k - r_k\right)^2}}{r_k},$$

for which the same dependency on the sampling fraction can be observed (see Fig. 3). We also notice a clear decreasing pattern of the RRMSE of the proposed estimators with increasing population cell frequencies.

The unconditional evaluation of the RRMSE tends to favour the SPREE estimator which is particularly well behaved for moderate to large population cells.

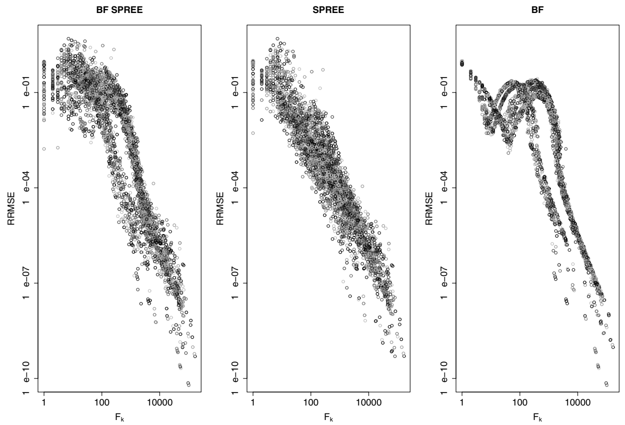


Fig. 3. Plot of root relative MSE for the three estimators. Darker grey levels indicate regions with smaller sampling fractions.

The same conclusions are suggested by the conditional analysis of the RRMSE for the sample unique cells. The figures analysed so far contain very small as well as larger cells; we then conditioned on both f_k and F_k: for such cells the alternative estimators $\hat{r}_k^{\text{BF SPREE}}$ and $\hat{r}_k^{\text{SPREE}}$ are comparable with respect to RRMSE.

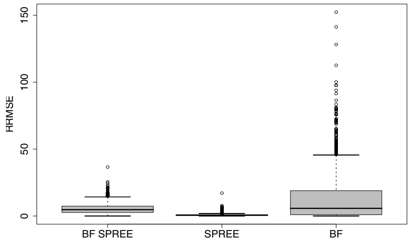


Fig. 4. Conditional assessment of RRMSE of the three estimators for sample uniques

7 Comments

In this paper we presented a comparative analysis of the BF estimator (3) and two alternative proposals (4.1) and (5), both based on the SPREE methodology. As far as the two SPREE based estimators are concerned, no one emerges as clearly superior to the other. The naive estimator, that is obtained as the reciprocal of $\hat{F}_k^{\text{SPREE}}$, should not be appropriate for small cells, as it results in overestimating the risk. We saw however that for very small population cells, namely the extremely risky ones, all the estimators, although to different extents, underestimate the true risk. We also remark that since the risk is a bounded parameter, use of the MSE to assess estimators' performance is perhaps not the best choice, other loss functions being perhaps more appropriate.

Both (4) and (5) require an iterative estimation procedure. In our experience the computational effort associated with the SPREE estimator is not an issue, as a straightforward iterative proportional fitting algorithm is used. Once an estimate of the population cell frequencies has been obtained, the risk can be easily computed in a single step via an approximation that has shown to be satisfactory even with small frequencies [17]. The feasibility for the person who is in charge of the data release to have access to census data might be an issue. 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. First, the population table from which the association structure is borrowed must be organized in an appropriate way. Secondly, the margins of the above table 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 to be released, some treatments, such as recoding, are usually needed.

Such a process is nontrivial and might also be computationally demanding, depending on the size of the population. An advantage is that the association structure between the key variables in the population enjoys stability over time. This implies that the same association structure can be considered at subsequent releases, the only change being the update of the margins of the table. In our simulation plan we purposely chose a ten years lag and observed that this does not affect the estimates to a large extent, the only necessary adjustment being the update in the margins.

A major difficulty with the SPREE is that it does not satisfy the natural requirement that estimates $\hat{F}_k^{\mathrm{SPREE}} > f_k$, so that sometimes even frequencies less than one might be obtained. This undesirable behaviour, not shared by the BF estimator, turns into estimated risks outside the natural $[0, 1]$ interval, and clearly affects the overall performance of the estimators to a large extent. In the experiment presented here, we simply adjusted our estimates, but we believe that the presented estimators can be improved in several respects. We will show results of modifying the proposed estimators in a subsequent paper.

Acknowledgments. The empirical study was conducted under a joint research project 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 the Programme Committee for comments that greatly improved the paper.

References

1. M. Abramowitz and I. A. Stegun. *Handbook of Mathematical Functions.* Dover, New York, 1965.
2. R. Benedetti and L. Franconi. Statistical and technological solutions for controlled data dissemination. In *Pre-proceedings of New Techniques and Technologies for Statistics*, volume 1, pages 225–232, Sorrento, June 4-6 1998.
3. M. Carlson. Assessing microdata disclosure risk using the Poisson-inverse Gaussian distribution. *Statistics in Transition*, 5:901–925, 2002.
4. G. Chen and S. Keller-McNulty. Estimation of identification disclosure risk in microdata. *Journal of Official Statistics*, 14:79–95, 1998.
5. J. C. Deville and C. E. Särndal. Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87:367–382, 1992.
6. L. Di Consiglio, L. Franconi, and G. Seri. Assessing individual risk of disclosure: an experiment. In *Proceedings of the Joint ECE/Eurostat Work Session on Statistical Data Confidentiality*, Luxembourg, April 7-9 2003.
7. G. T. Duncan and D. Lambert. Disclosure-limited data dissemination (with comments). *Journal of the American Statistical Association*, 81:10–27, 1986.
8. E. A. H. Elamir and C. J. Skinner. Modeling the re-identification risk per record in microdata. In *54th Session of the International Statistical Institute*, Berlin, August 13-20 2003.
9. S. E. Fienberg and U. E. Makov. Confidentiality, uniqueness, and disclosure limitation for categorical data. *Journal of Official Statistics*, 14:385–397, 1998.
10. J. J. Forster. Bayesian methods for disclosure risk assessment. In *Proceedings of the Joint ECE/Eurostat Work Session on Statistical Data Confidentiality, Geneva 9-11 November 2005*, pages 99–108. Luxembourg, 2005.
11. L. Franconi and S. Polettini. Individual risk estimation in μ-Argus: A review. In J. Domingo-Ferrer and V. Torra, editors, *Privacy in Statistical Databases 2004*, volume 2316 of *Lecture Notes in Computer Science*, pages 262–272. Springer, Berlin, 2004.
12. A. Hundepool. The CASC project. In Josep Domingo-Ferrer, editor, *Inference Control in Statistical Databases*, volume 2316 of *Lecture Notes in Computer Science*, pages 172–180. Springer, 2002.
13. D. Lambert. Measures of disclosure risk and harm. *Journal of Official Statistics*, 9:313–331, 1993.
14. W.G. Madow. On the theory of systematic sampling ii. *The Annals of Mathematical Statistics*, 20:333–354, 1949.
15. Y. Omori. Measuring identification disclosure risk for categorical microdata by posterior population uniqueness. In *Proceedings of the Conference on Statistical Data Protection*, pages 59–76, Lisbon, March, 25-27, 1998 1999. Eurostat, Luxembourg.
16. S. Polettini. Some remarks on the individual risk methodology. In *Proceedings of the Joint ECE/Eurostat Work Session on Statistical Data Confidentiality*, Luxembourg, April 7-9 2003.
17. S. Polettini. Revision of "Guidelines for the protection of social microdata using individual risk methodology: Application within μ-Argus version 3.2", by S. Polettini and G. seri. CASC-Computational Aspects of Statistical Confidentiality Deliverable No: 1.2-D3, 2004. Avaliable at `http://neon.vb.cbs.nl/casc/deliv/CASC_1.2D3_guidelines_new.pdf`.
18. N. J. Purcell and L. Kish. Postcensal estimates for local areas (small domains). *International Statistical Review*, 48:3–18, 1980.

19. J. N. K. Rao. *Small area estimation*. John Wiley & Sons, Hoboken, New Jersey, 2003.
20. J. P. Reiter. Estimating risks of identification disclosure for microdata. *Journal of the American Statistical Association*, 100:1103–1113, 2005.
21. J. P. Reiter. Releasing multiply-imputed, synthetic public use microdata: An illustration and empirical study. *Journal of the Royal Statistical Society, Series A*, 168, 2005.
22. Y. Rinott. On models for statistical disclosure risk estimation. In *Proceedings of the Joint ECE/Eurostat Work Session on Statistical Data Confidentiality, Luxembourg, 7-9 April 2003*, Luxembourg, 2003.
23. C. J. Skinner and M. J. Elliot. A measure of disclosure risk for microdata. *Journal of the Royal Statistical Society, Series B*, 64:855–867, 2002.
24. C. J. Skinner and D. J. Holmes. Estimating the re-identification risk per record in microdata. *Journal of Official Statistics*, 14:361–372, 1998.
25. L. Willenborg and T. de Waal. *Elements of Statistical Disclosure Control*. Springer, New York, 2001.
26. L. Zhang and R. L. Chambers. Small area estimates for cross-classifications. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 66(2):479–496, 2004.

Appendix

Table 1. Summaries for the bias of the three proposed estimators for $F_k = 2$ and 5

		Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
$F_k = 2$	$\hat{r}_k^{\mathrm{BF\,SPREE}}$	−0.48	−0.26	0.19	0.07	0.36	0.50
	$\hat{r}_k^{\mathrm{SPREE}}$	−0.50	−0.40	0.02	−0.01	0.30	0.50
	$\hat{r}_k^{\mathrm{BF}}$	−0.49	−0.47	−0.41	−0.42	−0.37	−0.37
	$\mathcal{C}_k : n_s > 0$	3	9	44	38.92	67	87
$F_k = 5$	$\hat{r}_k^{\mathrm{BF\,SPREE}}$	−0.18	0.03	0.19	0.22	0.39	0.78
	$\hat{r}_k^{\mathrm{SPREE}}$	−0.20	−0.11	0.05	0.11	0.25	0.78
	$\hat{r}_k^{\mathrm{BF}}$	−0.19	−0.17	−0.11	−0.12	−0.08	−0.06
	$\mathcal{C}_k : n_s > 0$	7	29	108	105.70	174	200

Single-Database Private Information Retrieval Schemes : Overview, Performance Study, and Usage with Statistical Databases

Carlos Aguilar Melchor* and Yves Deswarte *

LAAS-CNRS, Université de Toulouse
{carlos.aguilar, yves.deswarte}@laas.fr

Abstract. This paper presents an overview of the current single-database private information retrieval (PIR) schemes and proposes to explore the usage of these protocols with statistical databases. The vicinity of this research field with the one of Oblivious Transfer, and the different performance measures used for the last few years have resulted in re-discoveries and contradictory comparisons of performance in different publications. The contribution of this paper is twofold. First, we present the different schemes through the innovations they have brought to this field of research, which gives a global view of the evolution since the first of these schemes was presented by Kushilevitz and Ostrovsky in 1997. We know of no other survey of the current PIR protocols. We also compare the most representative of these schemes with a single set of communication performance measures. When compared to the usage of global communication cost as a single measure, we assert that this set simplifies the evaluation of the cost of using PIR and reveals the best adapted scheme to each situation. We conclude this overview and performance study by introducing some important issues resulting from PIR usage with statistical databases and highlighting some directions for further research.

1 Introduction

Privacy is a major concern in the cyberspace. In statistical databases, much work has been done to protect private data from statistical inference, but to the best of the authors' knowledge, no attention has been given to the protection of the privacy of the users requesting the statistical data. Still, limiting the information about the users that the database administrators can obtain from the requests can be a major issue. For example, statistics collected by a medical research facility can reveal a lot of critical information on which are its privileged research axes, and the leakage of this information can lead to unfair competition.

To protect his privacy, a user accessing a non-statistical database may want to retrieve an element without revealing which element he is interested in. A trivial solution is for the user to download the entire database and retrieve locally the element he wants to obtain. Private Information Retrieval (PIR for short) schemes aim to provide the

* This work was partially supported by the European Commission - IST PRIME Integrated Project (IST-507591). http://prime-project.eu.org

J. Domingo-Ferrer and L. Franconi (Eds.): PSD 2006, LNCS 4302, pp. 257–265, 2006.

same confidentiality while reducing the communication cost with respect to the trivial solution. If users retrieve blocks of bits from the database, we talk of Private Block Retrieval (PBR). Both PIR and PBR were introduced by Chor, Goldreich, Kushilevitz, and Sudan in 1995 [1]. In their paper, they proposed a set of schemes to implement PIR through replicated databases, which provide users with information-theoretic security as long as the database replicas do not collude against the users.

A user of a statistical-database is not supposed to request single elements of the database. We will not deal in this paper with how to transform the existing PIR schemes into schemes allowing users to make statistical requests. It is for example trivial to obtain a mean value with a PIR scheme based on an homomorphic encryption scheme (see section 2), but evaluating which statistics can be realized with which scheme, and how, is beyond the scope of this paper. Security issues raised from the usage of PIR schemes in statistical databases will be commented in the conclusion, as well as directions for further research in this field.

In this paper, we will focus on PIR schemes that do not need the database to be replicated, which are usually called single-database PIR schemes. Users' privacy in these schemes is ensured only against computationally-bounded attackers. It is in fact proved that there exists no information-theoretically secure single-database PIR scheme with sub-linear communication cost [1]. The first of these schemes was presented in 1997 by Kushilevitz and Ostrovsky, and since then improved schemes have been proposed by different authors [2,3,4,5,6,7].

All of these schemes follow a similar approach, but it is difficult to understand which are the innovations brought by each of them, and the impact that the different innovations have on communication performance. We present in this paper the fundamental approach that all of these schemes follow, and indicate why each of them has meant a step forward and how this approach can be pursued to develop future schemes.

Single-bit and block retrieval notions are often mixed together in single-database PIR research. The seminal papers on PIR evaluated the schemes' performance through the communication cost to obtain a single bit. For block retrieval, the query size and the expansion factor on the information sent by the database to the users would be far more comprehensive measures than the global communication cost generally used. Some authors just add remarks on what are the *communication rate*, or the *evolution of the relative cost when retrieving large blocks*.

Such an approach has a negative impact on clearness. For example, the titles of the papers [5,6,7] indicate respectively logarithmic, log-squared, and constant evaluation of their communication performance but each of them uses a different measure. Another major example is the one of the quest for the single-database PIR with logarithmic communication cost. To retrieve a single bit in a n bit database without privacy concerns a user must send a $log(n)$-bit query (the index of the bit) and the database has to send back one single bit. This has a total communication cost of $(log(n) + 1)$ bits and therefore whether it is possible to privately retrieve a bit with a communication cost of $O(log(n))$ bits would be an interesting result. This approach is clearly not adapted for block retrieval in which case the fundamental issue would be whether it is possible to retrieve blocks of bits with queries of $O(log(n))$ bits and a constant expansion factor for the database replies[1]. Even if PIR schemes are more likely to be used for block retrieval, their asymptotic performance is generally evaluated by the proximity to the $O(log(n))$ limit of their total communication cost which is not only unpractical, but also theoretically less interesting than comparing them to the $O(log(n))$ and constant limit for block retrieval.

For these reasons we propose a set of measures to evaluate PIR schemes which take into account block retrieval, and is independent of the application which the scheme is used for. We give some examples of applications to illustrate how this set of measures gives a clear notion of which would be the communication costs when using a given PIR scheme instead of a non-private retrieval.

In Section 2 we present the overview and analysis of the existing single-database PIR schemes, and in Section 3 we introduce the set of measures and compare their performances.

2 Analysis of the Existing Single-Database PIR Schemes

In [8], Kushilevitz and Ostrovsky created the first single-database PIR scheme, by using quadratic residues. What exactly are quadratic residues is not as important as their properties:

- the user can efficiently generate numbers which are quadratic residues (QRs) and numbers which are quadratic non residues (QNRs),
- the user can efficiently test if a number is a QR or a QNR,
- the user can send sets of such numbers to a database which will be unable to distinguish QRs from QNRs
- there is an operation *OP*, computable by such a database, that from a set of QRs and QNRs gives a QNR if and only if the number of QNRs in the initial set is odd.

The idea behind this PIR scheme is for the query to be constituted of one QR number for each bit of the database except for the bit to be retrieved and a QNR number for that bit. The database computes the operation *OP* over the set of numbers associated with the bits in the database's string set to one (and ignores the others), and sends the result to the user. If the bit the user is interested in is set to one the database will have selected a QNR among the numbers and the result will be a QNR. If the bit the user is interested in is set to zero, the database will have selected only QRs and the result of the operation will be a QR number. Figure 1 resumes this idea.

With such a scheme the user has to send n numbers and the database replies a single number. The communication cost is thus $O(n)$. To reduce this cost, the basic scheme presented by the authors integrates a load balancing technique presented in the seminal paper about PIR [1]. The principle is to see the n-bit database as a matrix of s lines and t columns (with $s \times t = n$). The user sends t numbers and the database sends s numbers, one for every line in the matrix. As Figure 2 shows, the user retrieves a full column of data, containing the bit he is interested in.

[1] For instance, a scheme with $log(n)$-bits long queries and a reply expansion factor of $log(n)$ would satisfy the $O(log(n))$ requirement for single-bit retrievals but not constant expansion one for block retrievals.

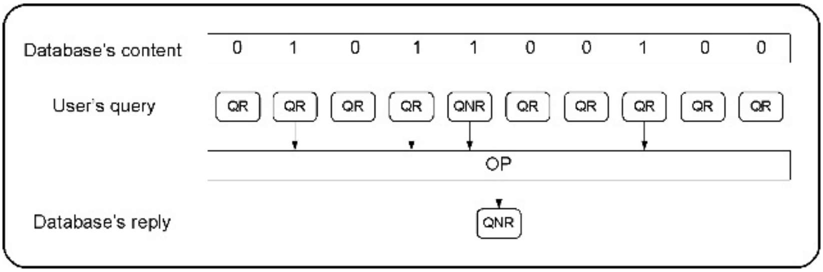


Fig. 1. QRs and PIR schemes

The first use of quadratic residues is to make the database give the users an information without knowing which. The second recursive use is made for the user to extract the expected result from the database output. This technique enables to reduce the communication cost to an expansion factor of t numbers for the query and s numbers for the reply, i.e. a communication cost of $O(n^{1/2})$.

This approach is effective for two reasons. First, a user does not need to retrieve all of the data sent by the server when using the load balancing technique, but only the number associated with the bit he is interested in. Second, PIR schemes are meant to provide sub-linear communication and therefore the output generated by the database is smaller than the database itself, and therefore using twice the scheme will result on a communication improvement. The maximum number of recursions will be given by the initial dimension of the representation of the database. If we represent the database as an $s \times t$ rectangle only two levels of recursion will be possible, if more generally the database is seen as a L-dimension hyper-rectangle, L levels of recursion are possible, to reduce the communication to $O(n^{1/L})$.

The article by Kushilevitz and Ostrovsky led to two other works. The first is Eran Mann's master thesis [3] that gives a theoretical framework and generalizes the quadratic residues scheme to any family of trapdoor predicates with properties similar to those achieved by the quadratic residues (he calls them homomorphic trapdoor predicates). The second work is a paper by Julien P. Stern [2], which made a major outbreak. Stern proposes exactly the same scheme than Kushilevitz and Ostrovsky, except that, instead of using a trapdoor predicate that can only encode one bit of information (for example being a QR or a QNR), he proposes to use homomorphic encryption algorithms that have all the properties needed, but can encode many bits of information in every number resulting from the *OP* operation. When users try to retrieve blocks, this is of course very interesting, since every number sent back by the database can encode many bits contained in the database. However, even if the user is not interested in receiving a block of information, the possibility to encode many bits in each number is very interesting. The reason for this is pretty simple: the database reply can be a single number, instead of s numbers when using load balancing, or $O(L \times n^1/L)$ when using the maximum recursion.

One year later, Cachin, Micali, and Stadler presented a new scheme [4] based on a new trapdoor predicate that they called the ϕ-assumption. Whereas this may seem as a step backwards after Stern's work on homomorphic encryption scheme, it is not. The main reason is that this assumption has a very interesting property: a user can create

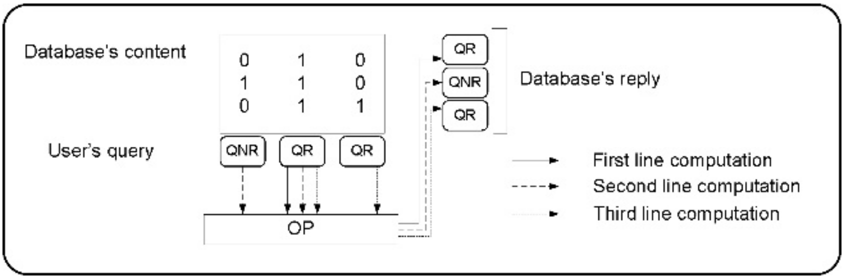


Fig. 2. Kushilevitz and Ostrovsky basic scheme

a compact generator, out of which the database can obtain the numbers forming the query. This causes a spectacular drop in the query size and even if the system is not practically implementable, when database size increases this approach beats asymptotically all the previous schemes. The basic idea is to create first a number generator, and afterwards create a trapdoor predicate such that the number associated with the index interesting the user has special properties. This approach in fact has led recently to a very interesting variation described below.

In 2004 there was a rediscovery of Stern's proposal [5], and a proposition by Lipmaa [6] which is basically Stern's construction with the recently discovered length-flexible homomorphic encryption scheme of Damgärd and Jurik [9]. In his paper Lipmaa twists Stern's construction, taking profit of the length-flexible cryptosystem to provide PIR schemes that are both practical and asymptotically interesting.

Finally, last year, Gentry and Ramzan presented a Scheme [7], which like Lipmaa's scheme is practical and presents an asymptotical improvement, even if for many applications Lipmaa's construction is better, as shown in the next section. In their paper, the authors present a construction that generalizes the proposal of Cachin et al., and their scheme can be implemented using a slight variation of the ϕ-assumption. Aside from the generalization, two major modifications are done with respect to the initial scheme. The first modification is pretty much the same as Stern did with respect to Kushilevitz and Ostrovsky's scheme: modifying the trapdoor predicate to encode more than a single bit. The second modification is very simple and consists on using the same numbers (which are associated with the database bits) for all the queries. This is almost trivial but was not proposed by Cachin et al., and it allows to make very small queries that just describe the structure in which the desired numbers will have special properties. However the cost reduction applies only to all the queries except the first one, since the numbers must be exchanged at least once. If users send only one query (or a small number of queries) to a given server,the communication cost per query may be unaffordable.

3 Performance Comparison

In PIR protocols, database contents are usually described as n-bit strings. With such a representation it is not possible to say whether a database contains n one-bit elements or

n/l l-bit elements. For example, a database of one million user profiles of 10 Kbits each cannot be distinguished from another one with one thousand songs of 10 Mbits each. To obtain a more intuitive representation, we prefer to describe the database as a set of n l-bit elements. If $l = 1$ we obtain of course the classic n-bit string representation.

The set of measures we will use to compare the schemes is: setup cost, query size, reply size, and maximum number of encodable bits by PIR reply that we will call the chunk size. When a scheme is implemented with a chunk size greater than $\alpha \times l$, α being an integer constant, a trivial improvement can be done: the database can be seen as composed of n/α elements of size $\alpha \times l$. This does not increase the size of the database reply, but lowers either the setup cost or the queries size. As this improvement is common to all schemes and is database dependent, we will not include it on the results, letting the reader evaluate it for her/his own application.

Table 1. Performance comparison

Algorithm	KO	Stern	CMS	Lipmaa	GR
Setup cost	k	$3 \times k$	0	$3 \times k$	$n^{1/L} \times k$
Query size	$L \times k \times n^{\frac{1}{L}}$	$L \times (s+1) \times k \times n^{\frac{1}{L}}$	$K^4 + 2 \times K^5$	$L \times \left(s + \frac{L+1}{2}\right) \times k \times \left(n^{\frac{1}{L}} - 1\right)$	$2 \times L \times k$
Reply size	k^L	$s \times \left(\frac{s+1}{s}\right)^L \times k$	K^5	$(s+L) \times k$	$k \times 5^{L-1}$
Chunk size	1	$s \times k$	1	$s \times k$	$k/5$

The use of the recursive construction proposed by Kushilevitz and Ostrovsky will be represented by a parameter noted L, $L = 1$ meaning that no recursion is done. The integer k represents a factorization-type security parameter, which for practical applications should not be lower than 1024. We have taken a different notation for Cachin et al.'s security parameter, K, as the authors fixed $K > log^2(n)$, even if their only non-factorizable integer had K^5 bits. This gives a much stronger constraint than $k > log^3(n)$, i.e., the usual asymptotic estimation against factorization, and therefore, as the security assumptions are different, we use different notations for the security parameters. Finally, s represents an integer parameter that can be fixed by the user. The computational cost of the schemes having this parameters is at least in $O((sk)^2)$ and therefore s must be kept close to one, in order to reduce computational cost.

The performance results presented in Table 1 for the scheme proposed by Kushilevitz and Ostrovsky (that we have noted KO) are not exactly the same as the one presented in their paper. These authors stayed with some load balancing, instead of pushing the recursive scheme to its maximum level. This strategy has been abandoned on current schemes and therefore to give a better comparison we have provided the results for the maximum recursive scheme rather than for the scheme with load balancing. Stern's and Lipmaa's schemes can be implemented with various encryption algorithms. The results presented in the table represent an implementation with the Damgärd-Jurik cryptosystem, which is the most effective and versatile homomorphic cryptosystem to date. Gentry and Ramzan (GR) did not propose the usage of the recursive construction of Kushilevitz and Ostrovsky. Maybe this is the best example of why we think the tools that have been used to improve existing schemes are unclear and need to be explicitly presented. There is no reason why Gentry and Ramzan's scheme should not use the recursive construction and its versatility is greatly improved with its usage. For Cachin et al. (CMS), we have not included L in the performance results, since it was designed as a theoretic scheme. The reason it has been included in Table 1 is to make visible the impact on asymptotic performance resulting from the innovations they introduced.

Indeed, the results illustrate the impact of the innovations presented in the previous Section. The drastic reduction of the reply size in recursive schemes that can be observed between the first and the second columns is the result of Stern's introduction of chunk sizes greater than unity. Query size in columns three and five does not depend on n, which is the result of Cachin et al.'s approach of generating the trapdoor predicates *after* defining the numbers forming the queries, which are associated with the database entries. Lipmaa's usage of length-flexible cryptosystems lowers from geometric to linear the increase of Stern's replies as L grows. Finally, the possibility of block retrieval, and the replacement of Cachin et al.'s number generators by a fixed set of numbers chosen on the setup step by Gentry and Ramzan, gives the first PIR scheme with a communication cost independent of n and efficient in practice[2].

The total communication cost for single-bit retrievals and retrievals of blocks smaller than the chunk size will be *Query size* + *Reply size*. Retrievals of blocks larger than the chunk size will result in a database reply expansion factor of *Reply size/Chunk size*. These figures give some basic information on how the schemes can apply. Retrieval of large elements should be done with Lipmaa's scheme, which has always the best expansion factor. When retrieving small elements Gentry and Ramzan's scheme should be used. But the setup cost induced by this scheme can be unaffordable for large databases if used with only few queries, or even impossible to deal with, for example for a user with a hand-held device. In this case Stern's scheme will often be a better choice than Lipmaa's since its queries are much smaller for large values of L. To illustrate this we give three examples:

- an online video-club with ten thousand movies of one hundred Gigabits,
- a global stock exchange database with one hundred thousand entries, each of two kilobits,
- a governmental database on citizens with two hundred million entries of five kilobits.

To limit computational cost we will consider that $s = 2$ for Lipmaa's and Stern's schemes. The size of the elements in the first example render inapplicable Gentry and Ramzan's scheme. Users should use Lipmaa's scheme with $L = 1$ to minimize their communication costs, sending 30 Mbits requests and receiving the movies with an expansion factor of 1.5. On the second example, Gentry and Ramzan's scheme is clearly the best choice. It can be used with $L = 1$, with which queries and downloads will be respectively of two and ten kilobits for each entry. Without recursion, the setup scheme needs however that the users exchange and stock one hundred megabits of data, and

[2] In an asymptotic evaluation we must suppose $k > O(log^3(n))$ [7]. Furthermore this scheme is based on the existence of enough prime numbers lower than $2^{k/5}$, however for any practical parameters there is no need to increase k above the factorization limit we have fixed. Indeed, if $k = 1024$, the results presented here are valid for $n < 10^{58}$.

therefore it can be used with $L = 2$ in which case the users will just have to stock 320 kilobits of data. With this level of recursion, the queries and downloads will be respectively of four and fifty kilobits. The third example is a case in which Gentry and Ramzan's scheme is unusable without recursion. The setup cost would be of 200 Gigabits. With $L = 2$ this scheme becomes interesting, since users have to stock 14 Mbits of data, send 4 kilobit queries and retrieve the information from the database in 125 kilobit downloads. Some users may prefer to use Stern's scheme with queries of 170 kilobits and downloads of 150 kilobits, avoiding in this way the 14 Mbit exchange. Lipmaa's scheme cannot provide such a good performance, since requests are larger than 512 kilobits for any value of L.

4 Conclusion

Trapdoor predicates with homomorphic properties allowed Kushilevitz and Ostrovsky to design the first single database PIR scheme. Two major improvements followed, the possibility to encode many bits in a single answer [2], and the possibility to create trapdoor predicates over given sets of numbers [4]. The use of length-flexible homomorphic cryptosystems [6] has reduced the size of the database replies for a given level of recursion. On the other side the schemes which are based on creating trapdoor predicates do not seem to have used all the possibilities explored on other papers. Versatility is improved in Gentry and Ramzan's schemes when they are used recursively following the construction of Kushilevitz and Ostrovsky. Other improvements may be possible, such as efficient prime number generators to further reduce the setup cost.

The usage of PIR schemes with statistical databases raises many research issues. Inference protection techniques often need the database to retrieve some information on the users queries. But, it is possible for a user to prove that a query respects a given pattern without revealing the query itself [2]. These *zero-knowledge proofs* are much used in cryptography, and whether they can be used to provide users with the possibility to obtain complex statistics while giving enough information to the database to ensure that no statistical inference can be done, is a challenging and most interesting issue. Some inference protection techniques like fixed perturbation [10] do not require the database to obtain any information on the user queries and will therefore be more interesting than query-based perturbation techniques [11] if used together with PIR protocols. A new performance parameter need therefore to be introduced (the amount of information needed by the database) and its evaluation for the different existing techniques is also an important issue to deal with in our context. We hope this paper will motivate research in these directions.

References

1. Chor, B., Goldreich, O., Kushilevitz, E., Sudan, M.: Private Information Retrieval. In: FOCS: IEEE Symposium on Foundations of Computer Science (FOCS). (1995) 41–50
2. Stern, J.P.: A New Efficient All-Or-Nothing Disclosure of Secrets Protocol. In Ohta, K., Pei, D., eds.: ASIACRYPT. Volume 1514 of Lecture Notes in Computer Science., Springer (1998) 357–371

3. Mann, E.: Private Access to Distributed Information, Technion Master's Thesis, Israel (2004)
4. Cachin, C., Micali, S., Stadler, M.: Computationally Private Information Retrieval with Polylogarithmic Communication. In: EUROCRYPT: Advances in Cryptology: Proceedings of EUROCRYPT. (1999) 402–414
5. Chang, Y.C.: Single Database Private Information Retrieval with Logarithmic Communication. In: ACISP: Information Security and Privacy: Australasian Conference. (2004) 50–61
6. Lipmaa, H.: An Oblivious Transfer Protocol with Log-Squared Communication. In Zhou, J., Lopez, J., Deng, R.H., Bao, F., eds.: ISC. Volume 3650 of Lecture Notes in Computer Science., Springer (2005) 314–328
7. Gentry, C., Ramzan, Z.: Single-Database Private Information Retrieval with Constant Communication Rate. In: ICALP: Annual International Colloquium on Automata, Languages and Programming. (2005) 803–815
8. Kushilevitz, E., Ostrovsky, R.: Replication Is Not Needed: Single Database, Computationally-Private Information Retrieval (extended abstract). In: FOCS: IEEE Symposium on Foundations of Computer Science (FOCS). (1997) 364–373
9. Damgrd, I., Jurik, M.: A Length-Flexible Threshold Cryptosystem with Applications. In: ACISP 2003. (2003) 350–364
10. Traub, J.F., Yemini, Y., Wozniakowski, H.: The statistical security of a statistical database. ACM Trans. Database Syst. **9**(4) (1984) 672–679
11. Beck, L.L.: A security machanism for statistical database. ACM Trans. Database Syst. **5**(3) (1980) 316–338

Privacy-Preserving Data Set Union

Alberto Maria Segre[1], Andrew Wildenberg[2], Veronica Vieland[3],
and Ying Zhang[4]

[1] Department of Computer Science, The University of Iowa, Iowa City,
IA 52246 (USA)
segre@cs.uiowa.edu

[2] Department of Computer Science, Cornell College, Mount Vernon,
IA 52314 (USA)
awildenberg@cornellcollege.edu

[3] Columbus Children's Research Institute, Ohio State University,
Columbus, OH 43205 (USA)
vielandv@pediatrics.ohio-state.edu

[4] Program in Applied Mathematical and Computational Sciences,
The University of Iowa, Iowa City, IA 52246 (USA)
yizha@math.uiowa.edu

Abstract. This paper describes a cryptographic protocol for merging two or more data sets without divulging those identifying records; technically, the protocol computes a *blind set-theoretic union*. Applications for this protocol arise, for example, in data analysis for biomedical application areas, where identifying fields (*e.g.,* patient names) are protected by governmental privacy regulations or by institutional research board policies.

Keywords: privacy-presenting data mining, blind database union.

1 Introduction

The analysis of data collected from multiple sources presents a variety of special challenges. Since more data generally yield better and/or more meaningful results, it is often desirable to apply statistical analyses to the union of multiple data sets. Such is the case, for example, when clinical data on a particular disease and its treatment are collected independently in the context of multiple research studies.

When privacy regulations, such as HIPAA [3], prohibit the sharing of identifying patient information, it is no longer possible to openly calculate the union of the data sets. Since the replicated records cannot be identified, the usual procedure is to simply concatenate the data sets to produce an approximation of their set-theoretic union. If there are no records present in more than one data set, then the approximation is exact; but if any records are shared, the result is usually a violation of the assumptions underlying whatever statistical analysis is to be performed.

J. Domingo-Ferrer and L. Franconi (Eds.): PSD 2006, LNCS 4302, pp. 266–276, 2006.

Our motivating application is to apply genetic linkage analysis tools to genotypic data collected from multiple clinical sites. *Genetic linkage analysis* describes a set of statistical genetics techniques used to locate on the genome the genes responsible for an inherited trait or disease (see, for example, [11]). A characteristic of such studies is unsystematic recruitment of families with multiple members affected with the disease under investigation; and as a result, it is not uncommon for motivated families to enroll themselves in multiple studies, that is, to be represented redundantly in data sets collected at different sites. When the data are then combined for analysis across sites, the presence of duplicated individuals appearing in more than one subset of the data violates a key assumption of the statistical analysis – that each family be considered only once – and can distort the results, even if there are only a small number of duplicates. This can be particularly problematic in studying rare genetic disorders, for which data sets tend to be small.

What is needed is a mechanism that identifies the overlapping records so that they won't be used twice. The difficulty is that identifying overlapping records entails comparing identifying information, which, by law, cannot be shared. In this paper, we present cryptographic protocols that allow two parties to negotiate the true set-theoretic union of their privately-held data sets without revealing restricted identifying information.

2 Background and Related Work

Let a *data set* consist of multiple *records*, where each record corresponds to a real-world entity we wish to model. Each record consists of a number of *attributes* or *fields*, where each attribute's *value* consists of a single typed datum; for example, "birthdate," "social security number," or "diagnosis" might be attributes in a patient information database, and "04/01/1947," "484-11-1991," and "asthma" might be the respective values of these attributes for a given record, which in turn represents a real-world person. Assume that Alice and Bob each possess data sets, A and B, with identical format, consisting of $|A|$ and $|B|$ records, respectively. Assume also that the attributes divide into two disjoint attribute subsets, denoted I and D, where I consists of the protected "identifying" attributes and D consists of the unprotected "data" attributes. In the terminology of relational databases, we expect that no combination of "data" attributes could serve as a candidate key, that is, could uniquely determine a real-world entity modeled by a data set record. But more to the point, the distinction between which attributes are protected "identifying" attributes and which are unprotected "data" attributes is really a matter of policy. These are fixed by the regulatory context (*e.g.,* HIPAA) in which the parties are operating.

More formally, Alice owns data set $A = \{I_a, D_a\}_i$ for $0 \leq i < |A|$, Bob owns data set $B = \{I_b, D_b\}_j$ for $0 \leq j < |B|$, and we wish to compute $D_{ab} = \Pi_D(A \cup_I B)$, where $\cup_I$ is the set-theoretic union operation that considers two records equal if and only if their identifiers are equal, and Π is a multiset relational algebra project operation (here Π_D is used to filter out identifying fields leaving only the data fields from each record in the resulting union). Note that we fully expect that the data fields, at least, will be subject to noise or measurement error.

An obvious and convenient solution to this problem is to take advantage of a mutually trusted third party, Ugo. Since Alice and Bob both trust Ugo, it is a simple matter for them to send him A and B, so that he can compute $A \cup_I B$. In fact, the existence of a trusted third party is not unusual in modern multisite clinical studies, where the data analysis site may be disparate from the data collection sites. In these situations, the data analysis site can be assigned to act as the trusted third party from the outset. Indeed, using a trusted third party should be the preferred solution, by virtue of its simplicity, in situations where it is permissible to do so. Unfortunately, the more normal situation is that no *a priori* agreement to share data exists, and so legal and policy issues may actually preclude selecting and using any sort of trusted third party *a posteriori.* Thus, in this paper, we consider the more interesting question of whether the union can be negotiated after the fact, without violating privacy concerns, and without resorting to a trusted third party.

There is a significant amount of related work in this area, which generally goes by the name *privacy-preserving data mining.* This work divides roughly into two categories; approaches that rely on sharing perturbed versions of each participant's data, and those that rely on cryptographic techniques. Here, we briefly summarize these approaches and explain why they are inadequate for our problem.

Distortion approaches assume that participants can disclose appropriately perturbed versions of their data without violating whatever privacy restrictions are in force. The core idea is that appropriately randomized records will be hard to identify: the trick is to ensure that the computation on the collected distorted data yields the same outcome – or at least close enough to the same outcome – as the computation when applied to the set-theoretic union of the original data [6,7]. There are two shortcomings with this approach that make it unsuitable for our problem. First, successful randomization strategies depend on intimate knowledge of the analysis algorithms which will be applied to the distorted data. In practice, these algorithms have typically been statistical algorithms such as sum, average, and max/min [1], or *e.g.,* , decision tree [13] construction algorithms [2], neither of which approach the complexity of genetic linkage-analysis algorithms [9,4]. Second, and more important, distortion strategies work by camouflaging the individual and reasoning about statistical distributions: no attempt is made to remove duplicated data points or account for any bias such replicated data points – distorted or not – may introduce.

In contrast, cryptographic techniques for the privacy-preserving database union have their basis in work introduced over 20 years ago on *secure two-party computation* [14]; Yao's work was later extended to multiple parties [5], hence the more usual name *secure multiparty computation*, or SMC. The idea is to devise a protocol that governs how two or more parties can exchange information in order to allow one (or both) of the parties to compute a function based on inputs held privately by each without revealing their own inputs. The general solution is based on circuit evaluation: the specified computational problem is first represented as a combinatorial circuit, then the participating parties run a short protocol for every gate in the circuit. The protocols generated depend on the size of the circuit, which in turn depends on the size of the input domain and on the complexity of the specified computation when expressed as a combinatorial circuit. In general, the solution to SMC is too inefficient to be useful in real applications, with the notable exception of some very simple cases, such as comparing the magnitude of two integers, where the computation remains tractable.

To get around the prohibitive costs associated with SMC, one approach is to combine SMC calculations with an inherently less secure but more cost-effective wrapper. This approach trades off divulging more information than the full SMC implementation for more efficiency; a good example is the ID3 decision tree construction protocol of [10]. By divulging intermediate results as they are computed, participants can construct the decision tree in tandem, at much lower cost by repeatedly using an SMC primitive to compute the information gain at each node, rather than trying to use SMC to compute the entire tree in one calculation. This approach works particularly well for ID3 tree construction, since the intermediate results divulged are explicitly part of the final solution anyway. Nonetheless, the protocol is cumbersome, and is not readily generalized to other types of data mining algorithms, and, once again, since the data mining algorithms employed are not as sensitive to replicated records as our linkage analysis algorithms, most solutions simply ignore the problem of replicated records and opt to work on the concatenation of data sets rather than their set-theoretic union.

One notable exception is the work of Kantarcioglu and Clifton on distributed data mining of association rules [8]. Since their work addresses a problem very similar to ours, it is not surprising that they explicitly consider the removal of duplicate records. Kantarcioglu and Clifton rely on a *commutative cipher* to blind records as they are exchanged among three or more participants in order compute the true set-theoretic union.[1] Ultimately, however, their solution still falls short in two ways.

First, at the end of the protocol, all records are revealed to the participants; the protocol is designed only to protect the identity of the original owner of each record, rather than the identifying attributes of the individual records (this also explains why their protocol is not suitable for use by only two parties: clearly, any record Alice does not recognize must have belonged to Bob). We

[1] A commutative cipher is an encryption algorithm having the property that a message encoded twice using two different keys must also be decoded twice, once with each key, but with decoding operations occurring in either order. Several, but not all, encryption algorithms in common use have this property (such as, for example, the RSA public-key encryption algorithm). For technical reasons having to do, at least in part, with RSA's vulnerability to chosen plaintext attacks, Kantarcioglu and Clifton rely on Pohlig-Hellman encryption [12], but the choice of cipher is not important to understanding their protocol, or, for that matter, the protocol introduced in this paper.

might try to get around this problem by using, *e.g.*, a *one-way hash function*[2] to obscure the identifying attributes prior to the union operation. But this doesn't really work as intended, because once the identifying attributes' one-way hash postimages are revealed, it is an easy matter for one participant to mount a *dictionary attack* on the other participant's data sets (*i.e.*, to "generate and test" candidate records to see if they are present). A second problem is that their protocol requires that matching records match completely, and does not account for noise or measurement error in at least some of the fields.

This paper presents a general solution to the two-party privacy-preserving data set union problem. Our protocol allows Alice to compute $D_{ab} = \Pi_D(A \cup_I B)$ with help from Bob, while Bob only learns $|A|$ and $|U_{ab}|$, the size of the resulting union $U_{ab} = A \cup_I B$. Neither Alice nor Bob learn which of their own records are replicated in the other participant's data set, nor are their own records' identifying attributes ever disclosed to the other party. Our protocol is robust to noise and measurement error in data fields, and, in cases where the "data" attributes of replicated records (*i.e.*, records in multiple data sets having identical "identifying" attributes) diverge, our protocol ensures that none of the divergent data values ever "leak" to other participants. The same protocol is easily extended to the multiparty case.

3 A Two-Party Privacy-Preserving Database Union Protocol

Our two-party protocol relies on several underlying assumptions. First, we assume that communication between parties is secure; that is, that no one but the originator and intended recipient can read any message sent as part of the protocol. This assumption is easily met, either by traditional means (*e.g.*, a trusted courier service) or by electronic/cryptographic means (*e.g.*, secure sockets, Open SSH, etc.). Second, we assume that the parties have been properly authenticated, that is, no one can pass themselves off as somebody else. This second assumption is typically met in a network environment by using a password-based authentication mechanism, and in a traditional environment by using a token-based authentication mechanism (*e.g.*, having the courier check the recipient's ID card). Third, we assume that all participants are basically honest and cooperative, yet still curious, an assumption commonly referred to as the *semihonest model.* Thus all participants will follow the rules of the protocol, but are still expected to try to discover what we are intent on keeping secret.

[2] A *one-way function* is a mathematical function that is easy to calculate in one direction (*i.e.*, calculating the output from the input) but infeasible to calculate in the other (*i.e.*, calculating the input from the output). Cryptographic systems make extensive use of a special kind of one-way function, called a *one-way hash function*, where a *hash function* maps its input, or *preimage* (which may be of arbitrary length) onto an output, or *postimage* (which is of fixed length). A hash function is also one-way if it is computationally hard to derive a preimage from the corresponding postimage.

Like Kantarcioglu and Clifton, our protocol uses a commutative cipher, but, in addition, it uses a keyed commutative one-way hash function. We'll let $K_a(M)$ denote encryption of message M with Alice's key K_a, and let $\overline{K}_a$ represent the decryption operator (so that $\overline{K}_a K_a(M) = M$). Further, let $H_a(M)$ denote the postimage of Alice's keyed commutative one-way hash function H_a applied to message M. Similarly defining K_b and H_b for Bob's cipher and hash, respectively, we can describe our protocol as follows.

1. Alice provides Bob with copies of her records, where the identifier is hashed using Alice's one-way keyed hash function and a random hash key H_a of her own choosing, and the data fields are encrypted using a commutative cipher with a random encryption key K_a of Alice's own choosing.

   $$\text{Alice} \rightarrow \text{Bob}: \quad \{H_a(I_a), K_a(D_a)\}_i \quad 0 \leq i < |A|$$

   The hash and cipher keys H_a and K_a are Alice's secrets, and should never be revealed.
2. Bob retains a copy of Alice's hashed identifiers and their associated encrypted data fields "in escrow" for later use. He then rehashes copies of Alice's hashed identifiers using the same commutative one-way hash function with a new random hash key H_b of his choosing and shuffles the result, returning the now doubly-hashed identifiers to Alice in some random order.

   $$\text{Bob} \rightarrow \text{Alice}: \quad \{H_b H_a(I_a)\}_i \quad 0 \leq i < |A|$$

   The hash key H_b is Bob's secret, and should never be revealed.
3. Bob next provides Alice with copies of his records, where the identifiers are hashed using the same commutative one-way hash key H_b of the previous step and the data fields are encrypted using a commutative cipher with random key K_b of Bob's own choosing.

   $$\text{Bob} \rightarrow \text{Alice}: \quad \{H_b(I_b), K_b(D_b)\}_j \quad 0 \leq j < |B|$$

   The encryption key K_b is also Bob's secret, and should never be revealed.
4. Alice uses her one-way hash function key H_a to rehash Bob's hashed identifiers, and her encryption key K_a to encrypt the associated data fields. At this point, Alice knows both

   $$\begin{array}{ll} \text{Alice: } \{H_b H_a(I_a)\}_i & 0 \leq i < |A| \\ \{H_a H_b(I_b), K_a K_b(D_b)\}_j & 0 \leq j < |B| \end{array}$$

   and recall that, because we are using a commutative cipher and a commutative one-way hash function, $K_b K_a(x) = K_a K_b(x)$ and $H_b H_a(x) = H_a H_b(x)$.
5. Alice now computes the union of the doubly-hashed identifiers and their associated data, if any, and then fills out the missing data fields with random bit strings, R. Each R must be identical in size to the encrypted data fields, so that all the records in the union have like format and size. She then shuffles the result, and returns it to Bob in some random order.

$$\text{Alice} \rightarrow \text{Bob}: \quad \{H_b H_a(I_{ab}), \{K_a K_b(D_b) \vee R\}\}_l \quad 0 \leq l < |U_{ab}|$$

where $I_{ab} = I_a \cup I_b$ and each doubly-hashed identifier $H_b H_a(I_{ab})$ is paired with either its doubly-encrypted data field $K_a K_b(D_b)$, if available, or else some random pattern R.

6. Bob decrypts, using $\overline{K}_b$, the data fields, producing:

$$\text{Bob: } \left\{H_b H_a(I_{ab}), \left\{K_a(D_b) \vee \overline{K}_b(R)\right\}\right\}_l \; 0 \leq l < |U_{ab}|$$

Since the random fillers R are just random bit sequences, $\overline{K}_b(R)$ is also a random bit sequence, and remains indistinguishable, to Bob, from $\overline{K}_b K_b K_a(D_b)$ $= K_a(D_b)$.

7. Next, Bob applies his hash H_b to the identifiers in Alice's original records (which he had previously, in step 2, retained "in escrow"), and then merges the resulting records with those just received from Alice in the previous step, overwriting any existing data fields in the latter.

$$\text{Bob: } \{H_b H_a(I_{ab}), K_a(D_{ab})\}_l \; 0 \leq l < |U_{ab}|$$

where D_{ab} denotes the data attributes associated each record in I_{ab} with D_a taking precedence over D_b for records in the intersection of the two data sets. Note that the overwritten fields are either just random bit strings or encrypted, and therefore unrecognizable, versions of Bob's own data for records that appear in both data sets.

8. Bob discards the doubly-hashed identifiers, then shuffles and returns the remaining now singly-encrypted data fields to Alice:

$$\text{Bob} \rightarrow \text{Alice}: \quad \{K_a(D_{ab})\}_l \quad 0 \leq l < |U_{ab}|$$

9. Alice decrypts, using $\overline{K}_a$, the data fields received from Bob, to produce the set-theoretic union of the two original data sets:

$$\text{Alice: } \{D_{ab}\}_l \; 0 \leq l < |U_{ab}|$$

As we shall soon see, Alice should *not* share this result with Bob in any form; the result is for Alice and Alice alone.

Having established that Alice is indeed able to compute $\Pi_D(A \cup_I B)$, we next consider whether the protocol permits either Bob or Alice to acquire additional information which should instead have been protected.

4 Discussion

Recall that, while we are primarily interested in safeguarding identifying information (*i.e.*, I_a and I_b), we must also safeguard other aspects of the union, such as the number and/or identity of any records in the intersection of A and B. Clearly, both Alice and Bob learn $|U_{ab}|$, the size of the union. Alice learns

$|B|$, the size of Bob's data set, in step 3, so it is a simple matter to compute $|A \cap B| = |A| + |B| - |U_{ab}|$ once she computes the union in step 5. Similarly, Bob can also compute the size of the intersection in step 5 when Alice sends the shuffled union back to him for decryption (he learns $|A|$ in step 1).

It is possible, with some effort, to obscure the exact value of $|A|$ from Bob and $|B|$ from Alice by having both parties introduce extraneous records (selected from a predetermined illegal data attribute distribution) which are then filtered out by Alice at the end of the protocol. But such inflationary masking still reveals loose bounds on $|A|$ and $|B|$ and, moreover, is not a critical element of our protocol: in our application, the data set sizes *per se* are never confidential information. In the end, our data analysis will lead to publication in the medical/scientific literature, where the sizes of the data sets are generally revealed.

Leaving aside the size of the intersection, the whole point of the protocol is not to reveal any I_a to Bob or any I_b to Alice. Alice obtains $H_b(I_b)$ in step 3, while Bob obtains $H_a(I_a)$ in step 1. Of course, these values are only as secure as the chosen keyed one-way hash function; since one-way hash functions cannot easily be inverted (indeed, that's the definition of a one-way hash function), neither Alice nor Bob can reconstruct an identifier from its postimage. Furthermore, as long as Alice cannot determine Bob's hash key H_b and as long as Bob can not determine Alice's hash key H_a, it is also impossible for either party to engage in a dictionary attack to reveal the other party's identifiers.

Note that even if both Alice and Bob learn how many of their original records were replicated in each others' data set, they can not know *which* of their records were actually replicated[3]. From Alice's perspective, this is because the end product of the algorithm is $\Pi_D(A) + \Pi_D(B \setminus_I A)$ where $+$ is the concatenation operator and $\setminus_I$ is the set difference operator that considers two records equal if and only if their identifiers are equal. In other words, all of Alice's original data fields are present in the result, and the only time she could have matched her own identifiers to Bob's identifiers (step 4) they were blinded by H_b. From Bob's perspective, while he was ultimately responsible for overwriting his own replicated data attributes while restoring Alice's data from escrow (step 7), he was not able to recognize them or their identifiers because they were blinded by H_a. Note further that it is clearly in Bob's interest to use Alice's data from escrow and not deviate from the protocol described, since doing so could only be to Alice's, and not Bob's, advantage.

The protocol also precludes the sort of dictionary attack alluded to in Section 2, since neither party has access to the other party's hash function key. Even if Alice adds additional "probe" records to her data set *a priori*, by overwriting his

[3] More precisely, $\mathrm{Prob}(\{I_b, D_b\} \in A \cap B) = \frac{|A \cap B|}{|B|} = \frac{|A|+|B|-|U_{ab}|}{|B|}$, which goes to 1 when $|U_{ab}| = |A|$ and goes to 0 when $|U_{ab}| = |A| + |B|$ (a symmetric formulation holds for $\mathrm{Prob}(\{I_a, D_a\} \in A \cap B)$). So in the special cases where all or none of one participant's records are replicated in the other participant's data set, the extent of the overlap is clear once $|U_{ab}|$, $|A|$, and $|B|$ are known. In other cases, random guessing about whether a particular record is in the intersection of the data sets is the best anyone can hope to do.

own data in step 7, Bob ensures that Alice will only ever see the "probe" data attributes she provides for records in the intersection. And since Bob does not return the doubly hashed identifiers to Alice, there is no means for Alice to probe the intersection after the fact; all probe records need to be added before the protocol starts (for Alice and Bob to collude here makes no sense, since if they're willing to collude they may as well share data openly). There is still one relatively weak form of dictionary attack available to Alice: by submitting only "probe" records (along with, optionally, some randomly-generated "distractor" records) as A, Alice can determine whether all of her probe records are also in Bob's database once she learns $|B|$ and $|U_{ab}|$. Of course, this will only work once, since Bob would have to be stupid not to wonder why Alice is repeatedly asking to engage in computing the union of her data sets with his: also, it violates our previously assumed semihonest model since Alice's actions violate the rules of the protocol.

It is important to note that Alice should never share her computed union U_{ab} with Bob. This is because the D_{ab} attributes in U_{ab} may not necessarily be the same as the D_{ba} attributes in U_{ba}, the union computed if Bob and Alice replay the protocol with roles reversed. To see why this is so, consider what happens when Alice and Bob share a record with identifiers $I_a = I_b$, but with $D_a \neq D_b$. Since the union computed by Alice contains D_a, and not D_b, Bob would immediately recognize the missing D_b and deduce that that particular record is replicated in both data sets. If, on the other hand, the protocol is replayed with roles reversed, Bob would see his own data attributes for any record in the intersection, and not those originally belonging to Alice.

5 Conclusion

We have presented a solution to the two-party privacy-preserving database union problem. Our solution allows an initiating participant to compute the true set-theoretic union (*i.e.*, without duplicated subjects) of a collection of data sets without obtaining identifying information about records belonging to other participants or divulging identifying information about one's own records. Our solution operates under the standard semihonest model, which assumes cooperating, yet still curious, participants, and precludes a participant from learning exactly which of his or her own records are present in the other participant's data set. Although in this paper we have assumed a horizontal partitioning of the data (*i.e.*, all parties have data sets with an identical collection of attributes), our protocol is equally suitable for vertically partitioned data or even mixed data, where some attributes are present in only some of the participants' data sets (in such cases, the escrow process ensures that values for unshared attributes are not acquired in the case of a record in the intersection of the data sets).

Moreover, unlike the existing work in secure multiparty computation, which is notoriously inefficient in practice, our protocol is quite efficient. Since each participant encrypts and decrypts a record at most once with each key in their possession, and since each participant hashes each identifier at most once, the cost of the protocol is $O(n)$ where $n = |A| + |B|$.

In short, our protocol is cheap to execute, easy to implement, and does not require a trusted third party. The protocol is easily extended to multiple parties with some minor modifications. Our protocol is therefore suitable for use in the analysis of data obtained from multisite clinical studies, where prior arrangements for sharing data have not been made, and, therefore, the sharing of identifying record information is precluded by law. This problem is also encountered in practice when data mining in a commercial setting, where data set owners are willing to cooperate to a certain extent, yet are keen to protect identifying values of their own records.

Acknowledgments

The authors are grateful to Dimo Dimotrov, Robert Hansen, Eunjin Jung, Ramon Lawrence, Meredith Patterson, and Alessio Signorini for their helpful comments. Support for this research was provided in part by the National Science Foundation through grants ITR/ACI0218491 (AMS) and CCLI/A&I0511391 (APW), Cornell College (APW), the National Alliance for Autism Research through the Autism Genome Project grant (VJV), and the National Institutes of Health through grant R01/NS042165 (VJV).

References

1. N. Adam and J. Wortman. Security-control methods for statistical databases: A comparative study. *Association for Computing Machinery Computing Surveys*, 21(4):515–556, December 1989.
2. R. Agrawal and R. Srikant. Privacy-preserving data mining. In *Proc. of the ACM SIGMOD Conf. on Management of Data*, pages 439–450. ACM Press, May 2000.
3. G.J. Annas. HIPAA regulations – a new era of medical-record privacy? *The New England Journal of Medicine*, 348(13):1486–1490, April 10 2003.
4. R.C. Elston and J. Stewart. General model for the genetic analysis of pedigree data. *Human Heredity*, 21:523–542, 1971.
5. O. Goldreich, S. Micali, and A. Wigderson. How to play any mental game: A completeness theorem for protocols with honest majority. In *Annual ACM Symposium on Theory of Computing*, 1987.
6. A. Hundepool, A. vanDeWetering, R. Ramaswamy, P.P. deWolf, S. Giessing, M. Fischietti, J.J. Salazar, J. Castro, and P. Lowthian. The τ-argus user's manual, version 3.1. http://neon.vb.cbs.nl/CENEX/Software/TauManualV31.pdf, November 2004.
7. A. Hundepool, A. vanDeWetering, R. Ramaswamy, L. Franconi, S. Polettini, A. Capobianchi, P.P. deWolf, J. Domingo, V. Torra, R. Brand, and S. Giessing. The μ-argus user's manual, version 4.0. http://neon.vb.cbs.nl/CASC/deliv/MUmanual4.0.pdf, November 2004.
8. M. Kantarcioglu and C. Clifton. Privacy-preserving distributed mining of association rules on horizontally partitioned data. *IEEE Transactions on Knowledge and Data Engineering*, 16(9):1026–1037, September 2004.
9. E. Lander and P. Green. Construction of multilocus genetic linkage maps in humans. *Proceedings of the National Academy of Sciences*, 84:2363–2367, April 1987.

10. Y. Lindell and B. Pinkas. Privacy preserving data mining. *Journal of Cryptology*, 15(3):177–206, June 2002.
11. J. Ott. *Analysis of Human Genetic Linkage*. Johns Hopkins University Press, 1999.
12. S.C. Pohlig and M.E. Hellman. An improved algorithm for computing logarithm of $GF(p)$ and its cryptographic significance. *IEEE Transactions on Information Theory*, IT24:106–110, 1978.
13. J.R. Quinlan. Induction of decision trees. *Machine Learning*, 1(1):81–106, 1986.
14. A.C. Yao. Protocols for secure computations. In M.S. Carberry, editor, *Annual Symposium on Foundations of Computer Science*, pages 160–164, 1982.

“Secure” Log-Linear and Logistic Regression Analysis of Distributed Databases

Stephen E. Fienberg[1,2], William J. Fulp[1], Aleksandra B. Slavkovic[3], and Tracey A. Wrobel[3]

[1] Department of Statistics, Carnegie Mellon University
[2] Cylab and Machine Learning Department, Carnegie Mellon University
[3] Department of Statistics, Pennsylvania State University

Abstract. The machine learning community has focused on confidentiality problems associated with statistical analyses that “integrate” data stored in multiple, distributed databases where there are barriers to simply integrating the databases. This paper discusses various techniques which can be used to perform statistical analysis for categorical data, especially in the form of log-linear analysis and logistic regression over partitioned databases, while limiting confidentiality concerns. We show how ideas from the current literature that focus on “secure” summations and secure regression analysis can be adapted or generalized to the categorical data setting.

1 Introduction

There are many scientific or business settings which require statistical analysis that “integrate” data stored in multiple, distributed databases. Unfortunately, there can be barriers to simply integrating the databases. In many cases, the owners of the distributed databases are bound by confidentiality to their data subjects, and cannot allow outsiders access. This paper discusses various techniques which can be used to perform statistical analysis for categorical data, especially in the form of log-linear analysis and logistic regression over partitioned databases, while limiting confidentiality concerns. The technique used depends on how the database is partitioned, either horizontally (with the same variables but different cases) or vertically (with the same cases but different variables), and also whether log-linear or logistic regression analysis is the goal. This paper will focus primarily on horizontally partitioned databases, and especially on the fully categorical data situation in which case the minimal sufficient statistics are marginal totals and logistic regression is effectively equivalent to log-linear model analysis (e.g., see [1,3,7]).

Much of the literature on privacy-preserving data mining and secure computation has focused on regression problems. A subset of the technical issues relevant to those problems are of interest in this paper. In the vertically partitioned case the concern remains the same, that is specifying a full model based on all of the variables. But in the the horizontally partitioned case there is a new element, whether any single owner actually has enough data to get maximum

J. Domingo-Ferrer and L. Franconi (Eds.): PSD 2006, LNCS 4302, pp. 277–290, 2006.

likelihood estimates (MLEs)! For regression problems this is primarily an issue of identification and we usually require that the sample size n is greater that the number of variables p, although as n increases we get greater accuracy for our regression coefficients and our inferences. But for categorical data problems we will often need to deal with a different form of degeneracy due to sparse data—that associated with patterns of zero counts which yield MLEs on the boundary of the parameter space and thus "do not exist" (for details on existence see especially [6,9,23]). Thus a very important reason for entering into arrangements to do secure computation is that pooled sufficient statistics and tables may well produce existence when no single party has sufficient data to assure the same.

While this paper will focus primarily on log-linear modeling and logistic regression for horizontally partitioned databases, there has been a lot of recent work on broader literature related to partitioned databases. The National Institute of Statistical Sciences (NISS) has produced much work for securely combining a horizontally partitioned database and on performing linear regression analysis on a horizontally partitioned database without actually integrating the data (e.g., see [15,16,21]). Theory regarding performing linear regression on vertically partitioned databases has also been devoloped (e.g., see [14]). There has also been work exploring some broader issues of the privacy impact of data mining methods and their work is related to the literature on secure multi-party computation (e.g., see [27]). Specifically, Kantarcioglu and Clifton [13] discuss mining of association rules on horizontally partitioned database, while the work of Vaidya et al. [25] relates to mining for association rules on vertically partitioned database.

The paper is organized in the following manner. In the next section we present a formulation of the general problem. Then, in Section 3, we turn to the problem of secure computation for log-linear models (and logit models) over horizontally partitioned databases and we relate some of the ideas to the literature on disclosure limitation for single databases involving such data. In Section 4 we present a technique for dealing with logistic regression over horizontally partitioned databases and we contrast it with the approach from section 3 in the case of categorical predictors. We conclude with a discussion of distributed database techniques and other ongoing work.

2 Problem Formulation

Consider a "global" database that is partitioned among a number of parties or "owners." These owners could be thought of as companies or people who have distinct parts of the global database. In a statistical context, these owners are referred to as agencies. These agencies may want to perform log-linear or logistic analysis on the global database, but are unable or unwilling to combine the databases for confidentiality or other proprietary reasons. The goal is to share the statistical analysis as if the global database existed, without actually creating it in a form that any of the owners can identify and utilize.

2.1 Partitioned Database Types

There are two types of partitioned databases discussed in this paper, horizontally and vertically partitioned databases. We are going to assume there are K agencies with $K \geq 2$, but note that a case with $K = 2$ is often trivial for security purposes. Horizontally partitioned data is the case such that agencies share the same fields but not the same individuals, or subjects. Assume the data consist of vectors $\mathbf{X}$ and $\mathbf{Y}$, such that:

$$\mathbf{X'} = [\mathbf{X^{(1)}}, \mathbf{X^{(2)}}, \cdots, \mathbf{X^{(k)}}] \text{ and } \mathbf{Y'} = [\mathbf{Y^{(1)}}, \mathbf{Y^{(2)}}, \cdots, \mathbf{Y^{(k)}}], \tag{1}$$

and $\mathbf{X^{(k)}}$ is the matrix of independent variables, $\mathbf{Y^{(k)}}$ is the vector of responses, and $n^{(k)}$ is the number of individuals, all that belong to agency k, $k = 1, \ldots, K$. Let $N = \sum_{k=1}^{K} n^{(k)}$. Each $\mathbf{X^{(k)}}$ is an $n^{(k)} \times p$ matrix and we will assume that the first column of each $\mathbf{X^{(k)}}$ matrix is a column of 1's. We will refer to $\mathbf{X}$ and $\mathbf{Y}$ as the "global" predictor matrix and the "global" response vector respectively ([22]). For horizontally partitioned databases it is assumed that agencies all have the same variables, and that no agencies share observations. Also, the attributes need to be in the same order.

In vertically partitioned data, agencies all have the same subjects, but different attributes. Assume the data looks like the following:

$$[\mathbf{YX}] = \left[\mathbf{Y}\, \mathbf{X^{(1)}} \ldots \mathbf{X^{(k-1)}}\right], \tag{2}$$

where $\mathbf{X^{(k)}}$ is the matrix of a distinct number of independent variables on all N subjects, $\mathbf{Y}$ is the vector of responses, and $p^{(k)}$ is the number of variables for agency k, $k = 1, \ldots, K$. Note that each $\mathbf{X^{(k)}}$ is an $N \times p^{(k)}$ matrix and we will assume that the first column of the $\mathbf{X^{(1)}}$ matrix is a column of 1's. For vertically partitioned database it is assumed that agencies all have the same observations, and that no agencies share variables. In order to match up a vertically partitioned database, all agencies must have a global identifier, such as social security number. We are currently working on the problem of vertically partitioned data in the categorical data setting but do not report on any results here.

There is a third possible kind of partitioning which goes well beyond the two special cases and corresponds more closely to real-world settings, namely horizontally and vertically overlapping data, perhaps with measurement error. Kohnen et al. [19] treat a special case of this in the form of vertically partitioned, partially overlapping as an incomplete data regression problem, and use the EM algorithm to estimate values of the "missing" data.

3 Secure Computation for Horizontally Partitioned Categorical Databases

Karr et al. [16] outline an approach that allows for secure maximum likelihood estimation for a density belonging to an exponential family. This technique can be used for log-linear model analysis in fully categorical data situations, where the minimal sufficient statistics are sets of marginal totals. This secure maximum likelihood technique uses a process called secure summation, which we describe first and then point out how this fits with the exponential family formulation. We then discuss the implementation for log-linear models as well as a possible way to simply combine the tables securely.

Secure Summation. Consider agencies which all have a single number, and would like to know the sum of all their numbers. However, the agencies do not want to reveal their individual number to any other agency. Secure summation is a process where the sum of all the agencies can be securely computed. The basic idea is that one agency adds a random number R to their number v_1 and then reports to the next agency in line $R + v_1$. The second agency adds their number v_2 to the number received and sends $R+v_1+v_2$ to the third agency. The pattern continues until agency k has computed $R + v_1 + \ldots + v_k$ and gives the number to agency 1. Agency 1 then subtracts R from the total, and shares the number with all of the other agencies. As long as multiple agencies are not colluding, secure summation is a very secure process. For a more detailed description of this process, consult Karr et al. [16]. There are other techniques that have been suggested to eliminate collusion but we do not consider them here.

3.1 Secure Maximum Likelihood Estimation for Exponential Families

Consider a global database $\{x_i\}$ modeled as independent samples from an unknown density $f(\theta, \cdot)$ belonging to an exponential family:

$$\log f(\theta, x) = \sum_{\ell=1}^{L} c_\ell(x) d_\ell(\theta). \tag{3}$$

Here the $\{d_\ell(\theta)\}$ are known as canonical parameters and the $\{c_\ell(x)\}$ are the corresponding minimal sufficient statistics (MSSs). Then under the assumption of independence of L rows, the global log-likelihood function is

$$\log L(\theta, x) = \sum_{\ell=1}^{L} d_\ell(\theta) \left[\sum_{k=1}^{K} \sum_{x_i \in D_k} c_\ell(x_i) \right], \tag{4}$$

where D_k is the database of owner k.

If the database owners can agree in advance on the model (3), e.g., the log-linear model with no second order interaction, they can use secure summation to compute each of the L terms in (4). Then each agency can maximize the likelihood function however they choose. There remains serious potential confidentiality problems once $L \geq 2$ since the MSSs are not independent of one another and they jointly contain information about the full table. We thus need to check the extent to which this information is sufficient to seriously compromise the confidentiality of any individual in the database — i.e., if one party can identify with sufficiently high probability an individual in another party's database. In what follows we exploit the fact that log-linear models have a discrete exponential family structure.

3.2 Secure Maximum Likelihood for Log-Linear Models

The secure maximal likelihood technique can be used for fitting a log-linear model. Consider a three-dimensional model coming from simple multinomial sampling. We are therefore assuming that the total sample size n is fixed. In this situation, the p.d.f. for the multinomial distribution of $\{n_{ijk}\}$ is

$$\frac{n!}{\prod\limits_{i,j,k} n_{ijk}!} \prod_{i,j,k} \left(\frac{m_{ijk}}{n}\right)^{n_{ijk}}, \tag{5}$$

where $\{m_{ijk}\}$ are the expected cell counts. The log-likelihood of the multinomial is readily obtained from the p.d.f. (5) as

$$\text{constant} + \sum_{i,j,k} n_{ijk} \log(m_{ijk}) - n \log(n). \tag{6}$$

Since the first and third term do not depend on the expected cell counts m_{ijk}, we need only to consider the remaining middle term, the kernel of this function. The saturated log-linear model for the expected cell count m_{ijk} is

$$\log(m_{ijk}) = u + u_{1(i)} + u_{2(j)} + u_{3(k)} + u_{12(ij)} + u_{13(ik)} + u_{23(jk)} + u_{123(ijk)}. \tag{7}$$

Substituting for m_{ijk} from (7) into (6), we obtain for the kernel

$$\begin{aligned}\sum_{i,j,k} n_{ijk} \log(m_{ijk}) = {} & Nu + \sum_i n_{i++}u_{1(i)} + \sum_j n_{+j+}u_{2(j)} + \sum_k n_{++k}u_{3(k)} \\ & + \sum_{i,j} n_{ij+}u_{12(ij)} + \sum_{i,k} n_{i+k}u_{13(ik)} \\ & + \sum_{jk} n_{+jk}u_{23(jk)} + \sum_{i,j,k} n_{ijk}u_{123(ijk)}. \end{aligned} \tag{8}$$

Since the multinomial distribution belongs to the class of discrete exponential family densities, the minimal sufficient statistics (MSSs) are the observed count n-terms adjacent to the unknown parameters, the u-terms. If we consider an unsaturated model the n_{ijk} terms fall out of expression (8), and those terms that remain give the MSSs. These marginal tables can then be used to estimate the cell expectations $\{\hat{m}_{ijk}\}$ under the model. In fact, it is in general multi-way tables that the MSSs correspond to the highest order u-terms in the model and the likelihood equations are found by setting them equal to their expectations (e.g., see [1,3,8,12]). Further, since the multinomial distribution is in the exponential family, working with log-linear models allows us to use the general secure maximum likelihood equation (4). Similarly, if the sampling model was "Poisson" or product-multinomial, the MSSs are essentially the same once we add in any margin fixed by the sampling scheme, and so the same secure computation idea works. In the product-multinomial situation, the log-linear model can be re-expressed as a logit model and this provides a way for dealing with the secure logistic regression computation problem in the fully categorical data case.

Now consider a horizontally partitioned categorical database. Since (4) is satisfied for log-linear models, it is possible to use secure summation to find the global sufficient statistics, which are marginals that correspond to the highest order u-terms in the model. The agencies will use multiple secure summation processes to create the global marginal statistics. The first agency adds a random number to each marginal value agreed to be summed, and then passes the values to the next agency. The second agency adds their numbers to the marginals, and passes them along. Once the first agency receives these it removes the random values and shares the marginals with all the agencies. If only necessary marginals for a specific model are computed through secure summation, the downside of this process is limited model comparison. If we wish to assess the fit of the model, then we can compare it to a larger log-linear model with additional u-terms. Thus we need to compute additional marginal tables in order to estimate the expected values under the larger model. The two models could be compared to see whether the more parsimonious provides an adequate fit to the data.

As we noted above, the MSSs, i.e., the marginal tables, carry information about the full table. This can come in the form of bounds for cell counts, or actual distributions over possible tables, for example see [5,8,10,11,24]. Computing and thus revealing additional combined marginal totals increases the information known about the individual cell in the overall combined table, possibly to an unacceptable level. Thus to protect individual level confidentiality in this setting we need to go beyond secure computation to incorporate methods from the more traditional disclosure limitation literature. There is also related literature on association rule mining, e.g., see [13,27], but it either focuses on the release of a single marginal or the form of the rule without the relevant data which turns out to be marginal totals [11]. Since using the association rule requires data to allow one to make predictions, releasing just the rule is rarely "useful."

3.3 Secure Contingency Table Analysis

Depending on the level of confidentiality, agencies may be willing to create a global contingency table, as long as the sources of data elements remain protected. Once a global contingency table is created, statistical analysis can be performed normally on the full database. A secure contingency table of counts or sums can be created using multiple secure summations. The general process is as described earlier in the paper, but instead of the first agency creating just one random number, the agency will create a random number for each cell in the table. Then the secure summation pattern applied to every cell in the table continues until the first agency gets the table back, removes all of the random cell values, and reports the full contingency table to the other agencies.

Often a categorical database is too large and sparse for this secure summation process to be efficient enough to use. If that is the case, then a secure data integration process can be used to get a list of cells which have non-zero cell counts, c.f., see discussion in [4] on issues with large sparse contingency tables. This general process is summarized later in this paper. The only adjustment for secure contingency table analysis is that the "data" being inserted into the growing database is really a list of non-zero cell counts. Once a list of non-zero cell counts is created, the multiple secure summation process can be used to get the complete table. This way, the agencies only need to use secure summation for a possibly very small subset of cells in a given table.

The secure contingency table process is only effective if the data elements themselves do not reveal from which party they come. This problem of the data revealing their source is one faced by other methodologies on secure data integration, e.g., see [17,27].

Secure Data Integration. Secure data integration is the process of securely combining observations of horizontally distributed databases into one data set. The basic secure data integration process consists of agencies incrementally contributing data into a growing database until the full database is complete. The goal of SDI is to combine these databases in a way so that the agencies will not be able to tell which agency a particular observation came from, except of course for the agency which originally had that observation. Karr et al. [16] lay out the secure data integration process in a reasonably complete fashion.

The growing database is passed from agency to agency in a round robin order, but in an order unknown to the agencies. Therefore, a trusted third party must be used, but the data can be encrypted so that only agencies can read the data. As the growing database is passed around, the agencies input a random number of observations into the database. This pattern continues until all the observations are into the growing database. Using this secure data integration process, a database can be securely combined.

4 Logistic Regression over Horizontally Partitioned Data

In this setting, logistic regression over a horizontally partitioned database is desired. We first explain that logistic regression can be considered as a specific form of the log-linear modeling. Later in the section we explain a specific technique for performing logistic regression over a horizontally partitioned database, which is not related to log-linear modeling.

4.1 Logistic Regression from Log-Linear Modeling

It is possible to use the approach above for log-linear models to do secure logistic analysis if all the explanatory variables are categorical. Consider simple logistic regression case with a single binary response variable. We can represent the data in contingency table form. The linear logistic regression model for this problem is essentially identical to the *logit* model found by differencing the log expectations for the two levels of the response variable, c.f. [3,7].

We illustrate for a logistic regression model with two binary explanatory variables (variables 1 and 2) and a binary response variable (variable 3). The data form a $2 \times 2 \times 2$ table. We work with the no second-order interaction model and construct the logit, i.e.,

$$\begin{aligned}
\text{logit}_{ij} &= \log\left(\frac{m_{ij1}}{m_{ij2}}\right) = \log(m_{ij1}) - \log(m_{ij2}) \\
&= \left[u + u_{1(i)} + u_{2(j)} + u_{3(1)} + u_{12(ij)} + u_{13(i1)} + u_{23(j1)}\right] \\
&\quad - \left[u + u_{1(i)} + u_{2(j)} + u_{3(2)} + u_{12(ij)} + u_{13(i2)} + u_{23(j2)}\right] \\
&= (u_{3(1)} - u_{3(2)}) + (u_{13(i1)} - u_{13(i2)}) + (u_{23(j1)} - u_{23(j2)}). \qquad (9)
\end{aligned}$$

Since we may place zero-sum constraints over the k index, the logit model in equation (9) simplifies to

$$\log\left(\frac{m_{ij1}}{m_{ij2}}\right) = 2u_{3(1)} + 2u_{13(i1)} + 2u_{23(j1)} = \beta_0 + \beta_1 x_1 + \beta_2 x_2, \qquad (10)$$

where $x_1 = 1$ for $\log(m_{1j1}/m_{1j2})$ and $x_1 = 0$ for $\log(m_{2j1}/m_{2j2})$ and similarly for x_2. Therefore, performing logistic regression over a horizontally partitioned database can be acheived through the techniques discussed in Section 3.

4.2 Secure Logistic Regression Approach

We now turn to a more general approach for logistic regression over a horizontally partitioned databases using ideas from secure regression (e.g. see [15],[16],[22]). In ordinary linear regression, the estimate of the vector of coefficients is

$$\hat{\beta} = (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{Y}. \qquad (11)$$

To find the global $\hat{\beta}$ vector, agency k calculates their own $((\mathbf{X}^{(k)})^t\mathbf{X}^{(k)})$ and $(\mathbf{X}^{(k)})^t\mathbf{Y}^{(k)}$ matrices. The sum of these respective matrices are the global $\mathbf{X}^t\mathbf{X}$ and $\mathbf{X}^t\mathbf{Y}$ matrices. Since the direct sharing of these matrices results in a full disclosure, the agencies need to employ some other method such as secure summation described earlier in the paper. In this secure summation process, the first agency adds a random matrix to its data matrix. The remaining agencies add their raw data to the updated matrix until in the last step the first agency subtracts off their added random values and shares the global matrices. Reiter [22] discusses some possibilities of a disclosure with this method.

We are suggesting to use the developed secure matrix sharing techniques and apply them to the logistic regression setting. We wish to fit a logistic regression

$$\log(\frac{\pi}{1-\pi}) = \mathbf{X}\beta \qquad (12)$$

model to the global data, $\mathbf{X}$ and $\mathbf{Y}$. In logistic regression, the vector of coefficients, or β, is of interest, but since the estimate of β cannot be found in closed form, we use Newton-Raphson or a related iterative method. At each iteration of Newton-Raphson, we calculate the new estimate of $\hat{\beta}$ by

$$\hat{\beta}^{(s+1)} = \hat{\beta}^{(s)} + (\mathbf{X}^t\mathbf{W}^{(s)}\mathbf{X})^{-1}\mathbf{X}^t(\mathbf{Y} - \mu^{(s)}) \tag{13}$$

where $\mathbf{W}^{(s)} = Diag(n_j\pi_j^{(s)}(1 - \pi_j^{(s)}))$, $\mu^{(s)} = n_j\pi_j^{(s)}$ and $\pi_j^{(s)}$ is the probability of a "success" for the j^{th} observation in the iteration s, $j = 1, \cdots, N$. The algorithm stops when the estimate converges. Note that we require an initial estimate of $\hat{\beta}$ (e.g., see [1] for more details).

Now we can apply the secure summation approach to our logistic regression analysis. We can choose an initial estimate for the Newton-Raphson procedure in two ways: (*i*) the parties can discuss and share an initial estimate of the coefficients, or (*ii*) we can compute initial estimates using ordinary linear regression of the responses and predictors using secure regression computations. In order to update β, we need the parts shown in (13). We can break the last term on the right-hand side up into two parts: the $(\mathbf{X}^t\mathbf{W}^{(s)}\mathbf{X})^{-1}$ matrix and the $\mathbf{X}^t(\mathbf{Y}-\mu^{(s)})$ matrix. At each iteration of Newton-Raphson, we update the π vector, and thus update the $\mathbf{W}$ matrix and the vector μ. We can easily show that

$$\begin{aligned}\mathbf{X}^t\mathbf{W}^{(s)}\mathbf{X} = {} & (\mathbf{X}^{(1)})^t(\mathbf{W}^{(1)})^{(s)}\mathbf{X}^{(1)} + (\mathbf{X}^{(2)})^t(\mathbf{W}^{(2)})^{(s)}\mathbf{X}^{(2)} \\ & + \cdots + (\mathbf{X}^{(1)})^t(\mathbf{W}^{(k)})^{(s)}\mathbf{X}^{(k)}\end{aligned} \tag{14}$$

and

$$\begin{aligned}\mathbf{X}^t(\mathbf{Y} - \mu^{(s)}) = {} & \mathbf{X}^{(1)}(\mathbf{Y}^{(1)} - (\mu^{(1)})^{(s)}) + \mathbf{X}^{(2)}(\mathbf{Y}^{(2)} - (\mu^{(2)})^{(s)}) \\ & + \cdots + \mathbf{X}^{(k)}(\mathbf{Y}^{(k)} - (\mu^{(k)})^{(s)})\end{aligned} \tag{15}$$

where $(\mu^{(k)})^{(s)}$ is the vector of $n_l^{(k)}\hat{\pi}_l^{(k)}$ values and $(\mathbf{W}^{(k)})^{(s)} = Diag(n_l^{(k)}\hat{\pi}_l^{(k)}(1-\hat{\pi}_l^{(k)})$ for agency k, $k = 1, \cdots, K$, $l = 1, \cdots, n^{(k)}$ and for iteration, s. This means that for one iteration of Newton-Raphson, we can find the new estimate of β by using secure summation as suggested by Reiter [22].

One major drawback of this method is that we have to perform secure matrix sharing for every iteration of the algorithm; every time it runs, we have to share the old $\hat{\beta}$ vector with all of the agencies so they may calculate their individual pieces. When all variables are categorical, this method involves more computation than using the log-linear model approach to logistic regression, where only the relevant marginal totals must be shared among the agencies. In the more general setting, we also have no simple way to check on potential disclosure of individual level data and thus we are providing security only for the parties and not necessarily for the individuals in their databases, e.g., see discussion in [22] for the linear regression secure computation problem.

Diagnostics. Finding the coefficients of a regression equation is not sufficient; we need to know whether the model has a reasonable fit to the data. One way to assess the fit is to use various forms of model diagnostics such as residuals, but this can potentially increase the risk of disclosure. As with the log-linear model approach we can compare log-likelihood functions of the larger model and the more parsimonious model. The log-likelihood for the logistic regression is:

$$\sum_{j=1}^{N} y_j\{\log(\pi_j) + (1-y_j)\log(1-\pi_j)\}. \tag{16}$$

We can rewrite the equation in terms of the K agencies and use secure summation to find this value

$$\sum_{k=1}^{K}\sum_{j=1}^{n^{(k)}}\{y_j^{(k)}\log(\pi_j^{(k)}) + (1-y_j^{(k)})\log(1-\pi_j^{(k)})\}, \tag{17}$$

as well Pearson's χ^2 statistic or the deviance:

$$X^2 = \sum_{k=1}^{K}\sum_{j=1}^{n^{(k)}}\left(\frac{y_j^{(k)} - n_j^{(k)}\pi_j^{(k)}}{\sqrt{n_j^{(k)}\pi_j^{(k)}(1-\pi_j^{(k)})}}\right)^2 \tag{18}$$

$$G^2 = 2\sum_{k=1}^{K}\sum_{j=1}^{n^{(k)}}\left\{y_j^{(k)}\log\left(\frac{y_j^{(k)}}{\hat{\mu}_j^{(k)}}\right) + (n_j^{(k)} - y_j^{(k)})\log\left(\frac{n_j^{(k)} - y_j^{(k)}}{n_j^{(k)} - \hat{\mu}_j^{(k)}}\right)\right\}. \tag{19}$$

If the change in the likelihood is large with respect to a chi-square statistic with (d.f.) degrees of freedom, we can reject the null hypothesis and conclude that the simpler model provides a better fit to the data.

4.3 Comparison of "Secure" Log-Linear Regression Methods

To demonstrate the difference in computation between the log-linear method for logistic regression and the secure logistic regression method, we will go through a simple example. The example is *not* intended to show how secure the processes are, but *only* to demonstrate the difference between computation in the two methods. Any use of secure summation between just two agencies is useless, because both agencies can simply subtract their number from the final result to find the other agency's data.

The data in Table 1 come from a randomized clinical trial on the effectiveness of an analgesic drug for patients in two different centers and with two different statuses reported on in [18], c.f. Fienberg and Slavkovic [11]. Treatment has 2 levels: Active=1 and Placebo=2. The original response had 3 levels: Poor=1, Moderate=2, and Excellent=3, but for the purposes of this example we combine the last two levels so the response variable is binary: Poor=1 and Not Poor=2.

The data from the first center correspond to Agency 1 and those from the second center to Agency 2 (see Table 1). Consider the possibility that the two centers would like to do statistical analysis over their combined data (see Table 2), but are unwilling to share their individual cell values.

Table 1. Clinical trial data by Agency

Agency 1 Data				Agency 2 Data			
		Response				Response	
Status	Treatment	1	2	Status	Treatment	1	2
1	1	3	25	1	1	12	12
1	2	11	22	1	2	11	10
2	1	3	26	2	1	3	13
2	2	6	18	2	2	6	12

Table 2. Combined clinical trial data over the clinical center

		Response	
Status	Treatment	1	2
1	1	15	37
1	2	22	32
2	1	6	39
2	2	12	30

Log-Linear Approach for Logistic Regression. We first consider logistic regression from log-linear modeling. We fit the log-linear model with no second order interaction which corresponds to the logistic regression model with no interaction (c.f. Section 4.1, and equation (10)). Note the i index relates to variable Status, the j to Treatment, and the k to Response. The two agencies first use secure summation to compute the 12 marginal totals, i.e., MSSs, n_{ij+}, n_{i+k}, and n_{+jk}. For example, to find n_{11+}, Agency 1 adds some random number to its n_{11+} value of 28, and sends the number to Agency 2. Agency 2 adds their n_{11+} value of 24 and sends the updated value to Agency 1, who subtracts the random number and reveals the total n_{11+} value of 52 (see Table 3 for the relevant marginals.)

Table 3. Relevant marginal values computed through secure summation

ind val	n_{ij+}	n_{i+k}	n_{+jk}
11	52	37	21
12	54	69	76
21	45	18	34
22	42	69	62

Next, we fit the desired log-linear model in Splus via *loglin* function that uses *iterative proportion fitting* (IPF); it converged in 3 iterations. Table 4 reports 4 relevant log odds values.

Secure Logistic Regression Approach. In the secure logistic regression approach, we consider the data in a database form instead of a contingency table. We use the Newton-Raphson algorithm to fit the logistic regression model presented in Equation (12). We used 0s for the initial $\hat{\beta}^{(0)}$ values. Since we know the response variable must be 0 or 1, we would not expect the $\hat{\beta}$ values to be very far from the $(-1, 1)$ interval. The algorithm converged in 4 iterations, and Table 4 reports 4 relevant log odds values.

Table 4. The estimated log odd ratios from the two different models

Log-Linear Model	Logistic Regression
$\log\left\{\frac{\hat{m}_{111}}{\hat{m}_{112}}\right\} = -0.989228$	$\log\left\{\frac{\hat{\pi}_{11}}{1-\hat{\pi}_{11}}\right\} = -0.989230$
$\log\left\{\frac{\hat{m}_{121}}{\hat{m}_{122}}\right\} = -0.305730$	$\log\left\{\frac{\hat{\pi}_{12}}{1-\hat{\pi}_{12}}\right\} = -0.305717$
$\log\left\{\frac{\hat{m}_{211}}{\hat{m}_{212}}\right\} = -1.707879$	$\log\left\{\frac{\hat{\pi}_{21}}{1-\hat{\pi}_{21}}\right\} = -1.707895$
$\log\left\{\frac{\hat{m}_{221}}{\hat{m}_{222}}\right\} = -1.024381$	$\log\left\{\frac{\hat{\pi}_{22}}{1-\hat{\pi}_{22}}\right\} = -1.024382$

Comparison of the Two Approaches. The results for the two approaches as reported in Table 4 agree as expected, but there is a significant computational difference. In the log-linear approach to logistic regression the agencies only need to perform one round of secure summation during this entire process to compute the relevant marginal values. After the relevant marginals have been revealed, the agencies can perform the analysis with them, and do not need to share any information again, thus reducing computations.

The secure logistic regression approach is computationally more intensive than the log-linear method since the agencies need to do secure summation at each iteration of the Newton-Raphson algorithm. Also, in real life settings, the data are likely to be more complex, meaning more iterations needed. This would make the secure logistic regression approach relatively even slower.

5 Conclusion

We have outlined a pair of approaches to carry out “valid” statistical analysis for log-linear model logistic regression of horizontally partitioned databases that does not require actually integrating the data. This allows parties (e.g., statistical agencies) to perform analyses on the global database while not revealing to one another details of the global database beyond those used for the joint computation. For the fully categorical data case we noted that log-linear models provided an alternative approach to logistic regression and one which also allowed us to respect the confidentiality of the data subjects. We also outlined a possible way to securely create a contingency table for horizontally partitioned categorical databases.

We are still developing ideas for logistic regression and log-linear models for strictly vertically partitioned databases and we would like to move towards problems involving partially overlapping data bases with measurement error.

Acknowledgments

The research reported here was supported in part by NSF grants EIA9876619 and IIS0131884 to the National Institute of Statistical Sciences, by NSF Grant SES-0532407 to the Department of statistics, Pennsylvania state University, and by Army contract DAAD19-02-1-3-0389 to CyLab at Carnegie Mellon University.

References

1. Agresti, A. (2002). *Categorical Data Analysis, Second Edition.* Wiley, New York.
2. Bertino, E., Igor Nai Fovino, I.N., and Provenza, L.P. (2005). "A framework for evaluating privacy preserving data mining algorithms," *Data Mining and Knowledge Discovery*, 11, 121–154.
3. Bishop, Y.M.M., Fienberg, S.E., and Holland, P.W. (1975). *Discrete Multivariate Analysis: Therory and Practice.* MIT Press, Cambridge, MA.
4. Dobra, A., Fienberg, S.E., Rinaldo, A., and Zhou, Yi. (2006). "Confidentiality Protection and Utility for Contingency Table Data: Algorithms and Links to Statistical Theory," Unpublished manuscript.
5. Dobra, A., Fienberg, S.E., and Trottini, M. (2003). "Assessing the risk of disclosure of confidential categorical data," in J. Bernardo et al., eds., *Bayesian Statistics 7*, Oxford University Press, 125–144.
6. Eriksson, N., Fienberg, S.E., and Rinaldo, A., and Sullivant, S. (2006). "Polyhedral conditions for the non-existence of the MLE for hierarchical log-linear models," *Journal of Symbolic Computation,* 41, 222–233.
7. Fienberg, S.E. (1980). *The Analysis of Cross-Classified Categorical Data.* MIT Press, Cambridge, MA.
8. Fienberg, S.E. (2004). "Datamining and Disclosure Limitation for Categorical Statistical Databases," *Proceedings of Workshop on Privacy and Security Aspects of Data Mining, Fourth IEEE International Conference on Data Mining* (ICDM), Brighton, UK.
9. Fienberg and Rinaldo (2006). "Three centuries of categorical data analysis: log-linear models and maximum likelihood estimation," *Journal of Statistical Planning and Inference*, to appear.
10. Fienberg, S.E. and Slavkovic, A.B. (2004a). "Making the release of confidential data from multi-way tables count," *Chance*, 17(3), 5-10.
11. Fienberg, S.E. and Slavkovic, A.B. (2005) Preserving the Confidentiality of Categorical Statistical Data Bases When Releasing Information for Association Rules, *Data Mining and Knowledge Discovery Journal.* 11(2), 155-180.
12. Haberman, S. J. (1974). *The Analysis of Frequency Data,* University of Chicago Press, Chicago, Illinois.
13. Kantarcioglu, M. Clifton, C. 2004. "Privacy preserving data mining of association rules on horizontally partitioned data," *Transactions on Knowledge and Data Engineering*, 16, 1026–1037.
14. Karr, A. F., Lin, X., Reiter, J. P., and Sanil, A. P. (2004). "Privacy preserving analysis of vertically partitioned data using secure matrix products," *J. Official Statist*, Submitted for publication. Available on-line at www.niss.org/dgii/technicalreports.html.
15. Karr, A. F., Lin, X., Sanil, A. P., and Reiter, J. P. (2005a). "Secure regressions on distributed databases," *Journal of Computational and Graphical Statistics*, 14, 263– 279.

16. Karr, A.F., Fulp, W.J., Vera, F., Young, S.S. (2005b). “Secure, Privacy-Preserving Analysis of Distributed Databases.” Available on-line at www.niss.org/dgii/techreports.html.
17. Karr, A. F., Lin, X., Sanil, A. P., and Reiter, J. P. (2006). “Secure statistical analysis of distributed databases,” In *Statistical Methods in Counterterrorism: Game Theory, Modeling, Syndromic Surveillance, and Biometric Authentication.* Edited by A. Wilson, G. Wilson, and D. Olwell. Springer, New York.
18. Koch, G., Amara, J., Atkinson, S. and Stanish, W. 1983. Overview of categorical analysis methods. SAS-SUGI, 8:785-795.
19. Kohnen, C.N., Reiter, J.P., Karr, A.F., Lin, X., Sanil, A.P. (2005). “Secure regression for vertically partitioned, partially overlapping data,” Available on-line at www.niss.org/dgii/techreports.html.
20. Reiter, J.P. (2003). Model diagnostics for remote access regression servers. *Statistics and Computing*, 13, 371–380.
21. Reiter, J.P. (2004). Secure regression on distributed databases. Unpublished manuscript.
22. Reiter, J.P. and Kohnen, C. (2005). “Categorical data regression diagnostics for remote access servers. *Journal of Statistical Computation and Simulation*, 75, 889–903.
23. Rinaldo, A. (2005). *Maximum Likelihood Estimation for Log-linear Models.* Ph.D. Dissertation, Department of Statistics, Carnegie Mellon University.
24. Slavkovic, A.B. (2004) “Statistical disclosure limitation with released marginal and conditionals for contingency tables,” *Proceedings of Workshop on Privacy and Security Aspects of Data Mining ICDM 04.* IEEE Computer Society Press, 13-20.
25. Vaidya, J. and Clifton, C. (2002). “Privacy preserving association rule mining in vertically partitioned data,” *The Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Edmonton, Alberta, Canada.
26. Vaidya, J.; Clifton, C. (2004). “Privacy-preserving data mining: Why, how, and when,” *IEEE Security and Privacy*, 2 No. 6, 19–27.
27. Vaidya, J., Clifton, C., Zhu, M. (2006). *Privacy Preserving Data Mining.* Springer Verlag, New York.

Measuring the Impact of Data Protection Techniques on Data Utility: Evidence from the Survey of Consumer Finances

Arthur Kennickell[1] and Julia Lane[2]

[1] Federal Reserve Board
arthur.kennickell@frb.gov
[2] NORC/University of Chicago
Lane-julia@Norc.uchicago.edu

Abstract. Despite the fact that much empirical economic research is based on public-use data files, the debate on the impact of disclosure protection on data quality has largely been conducted among statisticians and computer scientists. Remarkably, economists have shown very little interest in this subject, which has potentially profound implications for research. Without input from such subject-matter experts, statistical agencies may make decisions that unnecessarily obstruct analysis. This paper examines the impact of the application of disclosure protection techniques on a survey that is heavily used by both economists and policy-makers: the Survey of Consumer Finances. It evaluates the ability of different approaches to convey information about changes in data utility to subject matter experts.

1 Introduction

Data collectors face a complex problem. Usually substantial sums of money often public money are expended to collect data for research and policy purposes. There is an assumed obligation to make those data as fully and freely available as possible. Moreover, data collectors often create elaborate structures to create high quality data. However, ethical and often legal considerations force the collectors to take some set of actions to limit the ability of data users to identify respondents. During a time of rapidly improving technology for data linkage, like the present, public data sets are potentially increasingly vulnerable to intrusions. The most natural response of the most benign data collector would be to alter the data and to do so progressively over time in ways that seem likely to limit the possibilities of intrusion. In the absence of guidance by subject-matter experts, there is no reason to think that such changes would be in any way optimal for analytical purposes.

Despite the fact that much empirical economic research is based on public-use data files, the debate on the impact of disclosure protection on data quality has largely been conducted among statisticians and computer scientists. Remarkably, economists have shown very little interest in this subject, which has potentially profound implications for research. Without input from such subject-matter experts, statistical agencies may make decisions that unnecessarily obstruct analysis. The impact can

J. Domingo-Ferrer and L. Franconi (Eds.): PSD 2006, LNCS 4302, pp. 291 – 303, 2006.

range from simply reducing the precision of parameter estimates to biasing results or, in the worst case, closing down entire areas of research.

The practical consequences of such unguided data alterations are often quite substantial. For example, if data changes driven by disclosure protection are broadened over time, the true precision (as opposed to the precision computed from straightforward use of altered data) of parameter estimates is reduced. Thus, economists might incorrectly con-clude that an economic phenomenon like race or sex discrimination was no longer an issue, even though the result is purely as an artifact of disclosure limitation techniques. Similarly, biased coefficients could lead to incorrect evaluation of the benefits and costs of different policies. Even if distortions that are employed preserve the first moments of a distribution, the second, third and fourth moments of a distribution can be distorted. Moreover, some techniques that may relatively harmless in a static context, can be very harmful in a dynamic context. Despite the potential consequences, few, if any, statistical agencies inform researchers about the potential consequences of disclosure protection techniques on the quality of their analysis.

This paper examines the impact of the application of disclosure protection techniques on a survey that is heavily used by both economists and policy-makers: the Survey of Consumer Finances. It discusses different approaches to convey information about changes in data utility to subject matter experts. We begin by reviewing the current literature on definitions and measures of data utility.

2 Data Utility

2.1 Definitions

Developing a definition of data utility for disclosure-protected microdata is relatively straightforward conceptually, but much more difficult to implement in a meaningful way. The emerging consensus appears to be based around the utility of the data for inference. Duncan et al. [4], for example, describe data utility as “a measure of the value of information to a legitimate data user”. Karr et al. [7] define data quality, which is the precursor to data utility, as “the capability of data to be used effectively, economically and rapidly to inform and evaluate decisions. Necessarily, DQ is multi–dimensional, going beyond record level accuracy to include such factors as accessibility, relevance, timeliness, metadata, documentation, user capabilities and expectations, cost and context-specific domain knowledge”. Karr et al. [8] then define data utility as the ability to preserve the same inferences from released microdata as for the protected data. Statistical agencies define the concept slightly less formally, although the basic concept is the same. For example the OMB definition of utility is the “usefulness of the information for the intended audience’s anticipated purposes.”[1] Similarly, Haworth et al. [6], writing for the European statistical system, define utility as "the totality of features or characteristics of a product or service that bear on its ability to satisfy stated or implied needs of customers".

[1] The Office of Management and Budget (OMB) Guidelines for IQ (Office of Man-agement and Budget, 2002), cited in Karr et al. [8].

Implementing this consensus is more difficult. As Duncan et al. [4] point out, early measures of information loss (the opposite of data utility) for tabular data were quite primitive, and included the percentage of suppressed cells, the total number or number of categories suppressed. Domingo-Ferrer and Torra [3], attempted to develop measures on the principle that user analyses (e.g. regressions, means, etc.) on released data and on the original data should yield the same or at least similar results. A similar approach has been taken by Winkler [12], who defines a dataset as analytically valid if the following is approximately preserved (some conditions apply only to continuous variables): Means and covariances on a small set of subdomains; Marginal values for a few tabulations of the data. Winkler goes further in stating that a microdata file is analytically interesting if six variables on important subdomains are provided that can be validly analyzed.

2.2 Data Quality Metrics

Not surprisingly, given the conceptual discussion above, the different metrics that have been developed in the literature attempt to measure the amount of information loss associated with the use of the data. A few of the metrics are reviewed here, using the notation of the original authors.

Duncan et al. [4] focus in on the user's key parameters of interest, θ, and use the reciprocal of a Mean Square Error as their measure of utility:

$$U = \left[E(\hat{\Theta}_{x'} - \hat{\Theta}_x)^2\right]^{-1}$$

Where θ is the set of parameters of interest to the user, and the sub-scripts x and x' referring to the masked and unmasked data respec-tively. This approach has a number of advantages. First, the measure has a direct analogue with a measure of risk. Second, it penalizes large differences more than small. In addition, the metric is one that is famil-iar to most statisticians, and it has intuitive appeal in that large numbers reflect high levels of utility, smaller number reflect lower measures. Finally, the metric is measured over the outcomes of interest to users – namely the set of parameters of interest. However, it has a number of disadvantages as well. The most obvious is that it is not scale invariant, so that although it is straightforward to make comparisons across different types of disclosure techniques on the same set of analytical exercises, it is not straightforward to compare across different specifications. In addition, there is no natural interpretation of the order of magnitude of the measure.

Domingo/Torra [3] take a more catholic approach in listing a vari-ety of summary statistics of the information in the released dataset (denoted by a prime) and the original dataset, such as the variance covari-ance matrices V (on X) and V' (on X'), the correlation matrices R and R', correlation matrices RF and RF' between the original variables and the principal components factors obtained through principal components analysis, the commonality between each of the original variables and the first principal component C and C' [2] and the factor score coefficient matrices

[2] Commonality is the percent of each variable that is explained by the principal component.

Table 1.

	Mean square error	Mean abs. error	Mean variation
$X-X'$	$\frac{\sum_{j=1}^{p}\sum_{i=1}^{n}(x_{ij}-x'_{ij})^2}{np}$	$\frac{\sum_{j=1}^{p}\sum_{i=1}^{n}\lvert x_{ij}-x'_{ij}\rvert}{np}$	$\frac{\sum_{j=1}^{p}\sum_{i=1}^{n}\frac{\lvert x_{ij}-x'_{ij}\rvert}{\lvert x_{ij}\rvert}}{np}$
$V-V'$	$\frac{\sum_{j=1}^{p}\sum_{1\le i\le j}(v_{ij}-v'_{ij})^2}{\frac{p(p+1)}{2}}$	$\frac{\sum_{j=1}^{p}\sum_{1\le i\le j}\lvert v_{ij}-v'_{ij}\rvert}{\frac{p(p+1)}{2}}$	$\frac{\sum_{j=1}^{p}\sum_{1\le i\le j}\frac{\lvert v_{ij}-v'_{ij}\rvert}{\lvert v_{ij}\rvert}}{\frac{p(p+1)}{2}}$
$R-R'$	$\frac{\sum_{j=1}^{p}\sum_{1\le i\le j}(r_{ij}-r'_{ij})^2}{\frac{p(p-1)}{2}}$	$\frac{\sum_{j=1}^{p}\sum_{1\le i\le j}\lvert r_{ij}-r'_{ij}\rvert}{\frac{p(p-1)}{2}}$	$\frac{\sum_{j=1}^{p}\sum_{1\le i\le j}\frac{\lvert r_{ij}-r'_{ij}\rvert}{\lvert r_{ij}\rvert}}{\frac{p(p-1)}{2}}$
$RF-RF'$	$\frac{\sum_{j=1}^{p}w_j\sum_{i=1}^{p}(rf_{ij}-rf'_{ij})^2}{p^2}$	$\frac{\sum_{j=1}^{p}w_j\sum_{i=1}^{p}\lvert rf_{ij}-rf'_{ij}\rvert}{p^2}$	$\frac{\sum_{j=1}^{p}w_j\sum_{i=1}^{p}\frac{\lvert rf_{ij}-rf'_{ij}\rvert}{\lvert rf_{ij}\rvert}}{p^2}$
$C-C'$	$\frac{\sum_{i=1}^{p}(c_i-c'_i)^2}{p}$	$\frac{\sum_{i=1}^{p}\lvert c_i-c'_i\rvert}{p}$	$\frac{\sum_{i=1}^{p}\frac{\lvert c_i-c'_i\rvert}{\lvert c_i\rvert}}{p}$
$F-F'$	$\frac{\sum_{j=1}^{p}w_j\sum_{i=1}^{p}(f_{ij}-f'_{ij})^2}{p^2}$	$\frac{\sum_{j=1}^{p}w_j\sum_{i=1}^{p}\lvert f_{ij}-f'_{ij}\rvert}{p^2}$	$\frac{\sum_{j=1}^{p}w_j\sum_{i=1}^{p}\frac{\lvert f_{ij}-f'_{ij}\rvert}{\lvert f_{ij}\rvert}}{p^2}$

Source: [5]

F and F'.[3] The summary statistics are listed in Table 1, and include the mean square error, the mean absolute error, and the mean variation of each of these measures.

[3] Matrix *F* contains the factors that should multiply each variable in *X* to obtain its projection on each principal component. *F'* is the corresponding matrix for *X'*.

These approaches have a different type of appeal. The advantage is that they summarize the differences between the disclosure-proofed and original input data, rather than on a set of parameters that may be very different for different groups of users. The metrics on which at least some of them are measured, like the correlations, are scale invariant. They are also all based on approaches that are familiar to statisticians. However, a major disadvantage is that the information that is included is likely to be too much to permit users to discriminate across disclosure protection approaches. For example, some datasets, like the National Longitudinal Surveys of Youth, or the Survey of Consumer Finances, have literally thousands of variables and while some are much more important than others, the metrics weight each input variable equally.

An alternative approach, which has not been suggested in the literature, but is certainly intuitively appealing, is to simply report the percent difference in key input variables and in parameter estimates. This has the twin advantages of being scale invariant and easily understood; the disadvantage is that percent differences are not standard statistical measures with well defined properties.

In any event, none of these summary statistics has been widely adopted, leaving researchers in the dark about the impact of disclosure protec-tions on the quality of their analysis. For example, the most recent version of μ-Argus, the microdata protection package produced by the CASC project, devotes only one paragraph to measuring the impact of disclosure protection techniques on data quality:

> "In case of applying local suppressions only, μ–ARGUS simply counts the number of local suppressions. The more suppressions the higher the information loss. In case of automatic global re-coding μ–ARGUS uses an information loss measure that uses the following parameters: a valuation of the importance of an identify-ing variable (according to the data protector), as well as a valuation of each of the possible predefined codings for each identifying vari-able."

P 43, μ-Argus Manual 4.0, December 2004

Similarly, the Census Bureau's review of disclosure protection protocols, while providing an exhaustive list of ways to protect microdata, does not provide the impact on data utility Zayatz [13].

3 Description of Survey of Consumer Finances and Typical Uses of Data

The SCF has been conducted every three years by the FRB with the cooperation of the Statistics of Income Division (SOI) of the Internal Revenue Service since 1983. NORC has performed the data collection since 1992. This computer-assisted-personal interviewing (CAPI) survey collects data from a nationally representative sample of American households using a dual-frame sample design. One part is a multi-stage area-probability sample selected from the NORC National Frame. The other part, which is selected using statistical records derived from tax returns, is stratified to over-sample wealthy households [4]. The data are used to examine cross-sectional variation as well as to evaluate trends over time [5].

The survey gathers detailed data on households' balance sheets---their assets and liabilities---as well as collecting information on income, work, pensions, use of financial institutions, demographic characteristics and attitudes. Most of this information is commonly viewed as highly confidential by respondents. Thus, efforts to assure respondents of the measures taken to protect the confidentiality of their data play a central role in persuading them to participate in the survey and to provide reliable information. The pledge given to respondents becomes, at the very least, a moral obligation for the data collectors to take every effort to fulfill it. Furthermore, the data are collected under the framework of the Confidential Information Protection and Statistical Efficiency Act (CIPSEA) of 2002. Under this act, when respondents are told that their data are being collected "for statistical purposes only," as respondents in the SCF are told, there is also a strong legal obligation to ensure the protection of the confidentiality of the data collected. For the SCF, there is an additional obligation imposed by the use tax-derived data in the sample design. As a part of the agreement with SOI that makes the data available, the survey is obliged to develop and implement a plan for the release of micro data that passes a review by SOI staff.

The public version of the SCF, which is described in more detail below, is the only version of the data available outside the core project group at the FRB. Although it is possible for researchers within the Federal Reserve and at other institutions to request special estimates from the internal version of the data, the great majority of policy research and longer-term research is done with the public version of the data. Data users in many areas—taxation, saving, retirement, personal finance, more general finance, financial market regulation, and other areas—depend on the reliability of estimates obtained from the public data set. Thus, it is imperative that the actions taken to limit disclosure do not induce serious distortions of estimates obtained from this data set.

The necessity of alterations to the SCF data for purposes of disclosure limitation also stands in contrast to the strong push in the survey to produce high-quality data. Large amounts of resources are devoted to training and monitoring interviewers for purposes of quality control. For example, Athey and Kennickell describe a new procedure undertaken for the 2004 SCF to deal quickly with data quality issues during the field period of the survey [1]. The survey also uses great care in data processing and documentation to ensure that the data are handled and described in a way that that would ultimately be most useful for research. For example, the survey documents the original content of every variable; it employs multiple imputation to provide a

[4] This tax-based sample serves two purposes. First, it allows the survey to obtain sufficient numbers of people in different wealth groups to support the estimation required of the survey. Second, it allows for control for nonresponse, which the data indicate is highly correlated with wealth. This sample excludes people identified by *Forbes* as being among the wealthiest 400 people in the U.S. This restriction recognizes the very low probability that anyone in that group could be persuaded to participate in the SCF. This *Forbes* group accounted for approximately 2 percent of total household net worth in 2004.

[5] For a description of the data, see Athey and Kenneckill [1]. For a review of the SCF methodology and references to other supporting research, see [9].

measurable basis for the amount of missing information, and it bases the imputation on a broad set of covariates to support a wide variety of multi-variate analyses of the data.

4 Description of Disclosure Limitation Approaches

4.1 Generally Used Approaches

A number of different disclosure limitation techniques are used by statistical agencies: a good summary is provided by the Federal Committee on Statistical Confidentiality's Confidentiality and Data Access Committee[6].

The list of options is quite long. Some options can be categorized as the direct reduction of information -- variable deletion, recoding variables into larger categories, rounding continuous variables using top and bottom coding, using local suppression and enlarging geographic areas Zayatz [13].

Another set of options can be described as the perturbation of information: the microdata set is distorted prior to its publication. In this way, unique combinations of scores in the original data set may disappear and new unique combinations may appear in the perturbed data set; such confusion is beneficial for preserving statistical confidentiality. Examples of these include noise addition, data swapping, blanking and imputation, micro-aggregation, PRAM (post randomization Method of Perturbation) and the use of multiple imputation/modeling to generate synthetic data[7]

4.2 Approach Used in Survey of Consumer Finances (Including Changes over Time)

A number of different techniques are applied for purposes of disclosure limitation in the SCF.[8] The most basic change made to the data set for public release is that some cases are deleted. If an observation is deleted if it has net worth greater than the level of the least wealthy person identified in the Forbes list of the wealthiest 400 people in the U.S.; there were three such cases in the 2004 SCF. The view supporting this alteration is that too much information is available that could be matched with the SCF to identify extremely wealthy individuals.

Some variables available in the internal version of the data are not released at all. Geographic information is generally recognized as being one of the most useful things to know in deducing the identity of a survey respondent. Absence of such information poses a particular problem for researchers who wish to exploit variation in institutional and other structures across states to identify important elements factors in statistical models of economic behavior. Variables related to the sample design, the administration of the interview and a variety of other variables noted in detail in the SCF codebook are also suppressed.

Some categorical and other discrete variables are coarsened in the SCF public data set. For example, the detailed 4-digit occupation codes determined from verbatim responses from the respondents are reduced to one of six codes. For family members

[6] http://www.fcsm.gov/committees/cdac/index.html

[7] Zayatz, ibid, μ-Argus Manual 4.0, December 2004.

[8] For more details on the procedures applied to the SCF data to protect the identity of respondents, see Kennickell[10].

other than the household "head" and that person's spouse or partner, their ages are reduced to an indicator of whether they are aged 18 or older. For a number of other discrete variables, categories with small numbers of responses are combined with similar categories. Again, all such changes are documented in detail in the survey codebook.

Dollar variables in the SCF are all subjected to a type of rounding and the degree of rounding varies with the magnitude of the figure rounded. For example, values of $1 million or more are rounded to the nearest $10,000 and values between $10,000 and $1 million are rounded to the nearest $1,000. To minimize systematic distortions, the data are rounded up or down with probability proportional to the value modulo the rounding value. That is, a value of $1,222,221 would be rounded to $1.23 million with probability 2,221/10,000 and to $1.22 million with probability 7,879/1000. A number of other variables are also rounded. For example, the size of a farm or ranch is rounded to the nearest five acres, the proportion of pension assets held in stocks is rounded to the nearest five percent, and the last year that the household filed for bankruptcy (if it is has ever done so) is rounded to the nearest three years, an interval selected as appropriate for research purposes.

Top-coding and bottom-coding are used very sparingly. A decision to truncate the data in this way is usually made because the set of people affected is very small and very far removed from the rest of the distribution of households. For example, the number of checking accounts is topcoded at 10 and the age of the respondent is topcoded at 95. Negative values of certain income components and total income are bottom-coded at $-9.

The only other disclosure limitation procedure applied that has at least the potential for causing significant distortion of the data is a type of data simulation. This technique is applied to a set of cases selected systematically on the basis of their unusual values in terms of a set of characteristics and a random set of cases selected to assist in masking the primary set of cases. In the 2004 SCF, fewer than 350 cases were selected for this treatment. For the cases selected, the multiple imputation model developed for the SCF is used to simulate the values of all dollar variables; the values of all other variables are taken either as they were originally reported or as they were imputed in the final iteration of the iterative imputation routine. Even though the multiple imputation routines used for the simulations add a random error from the distribution of the unexplained variance of the variable simulated, because the sample size is relatively small one might still expect the cases selected for their unusual values to exhibit some regression toward the mean, and thus induce a serious distortion of the right tails of a number of distributions. Two factors help to mitigate this potential problem. First, the imputation model inputs tend to sustain some of the unusual qualities of cases. The imputation framework proceeds sequentially over variables, using as inputs covariances estimated using the final iteration of the imputed data and conditioning variables for the cases whose dollar values are to be simulated. All of the non-dollar-denominated conditional variables are taken from the final imputed data. The dollar values are initially taken from that data set as well, but once a value is simulated, the simulated value is used in later models in the sequence. Second, bounds are imposed on the outcomes of the simulations. These ranges are set as a baseline percent plus a randomized addition. The details of this process cannot be revealed, but the ranges are designed to provide a tight enough range to ensure that values cannot become too much larger or smaller, but also to allow sufficient range for the true values to be effectively disguised [9].

To further complicate the task of a potential data intruder, other unspecified changes are made to the data. The number of such changes is relatively small and the changes are almost all of a sort that would be highly unlikely to affect any analysis that took account of the inherent sampling variability in the data.

Unlike the case of changes made to the data through coding, editing, and imputation, changes as a result of disclosure reduction procedures are not documented in the shadow variables available for every case and every variable. For example, a shadow variable for a simulated variable would be indistinguishable from that for an unaltered variable that had originally been imputed using range information.

The procedures described here have been in place since the 1989 SCF. However, changes have been made in a variety of the details of the application of the procedures. The main changes have been in the set of variables suppressed and the degree of coarsening applied to categorical and discrete variables. Care has been taken at every such step to ensure as much backward continuity of measurement as possible.

Finally, data users have been encouraged to give feedback when the disclosure limitation procedures have interfered with research. The overwhelmingly most common complaint has been the lack of geographic information noted above. Users might also be concerned about the distorting effects of the disclosure limitation procedures, but they would be unable to make a judgment about these effect with the data available to them. Among other things, this paper is intended provide such an evaluation.

5 Description of Impact on SCF Analysis

In this section we analyse the impact of the disclosure protection approach on the utility of some of the most commonly used SCF variables: income, individual net worth and the debt to income ratio, as well as the conditioning variable, age. We begin by applying the Duncan approach to comparing summary statistics derived from disclosure protected and original measures of net worth and debt to income; both overall and by income and age categories. We then describe the same differences in terms of percent change. We do the same exercise to summarize the impact of disclosure proofing on the results of a common regression. Finally, we summarize a subset of the Domingo/Torra statistics.

Table 2 presents the first set of measures for the mean and the median summary statistics, with the statistic calculated from the original data presented in the first column. The first interesting result is that the percent change in the statistic as a result of calculating the data from disclosure protected data, is quite small – less than 2% in all cases. The effects are also shown quite vividly in Figures 1 and 2. The second result is that the Duncan measure does capture the differences in consequences on data utility quite well: bigger numbers (reflecting higher utility) are consistently found where the percent errors are smaller. However, a major problem is that the Duncan measure is difficult to interpret. The measure for net worth is very small,

[9] Detailed examination of the simulation results for the SCF suggests that the process does not cause serious univariate distortion of the data. See [10].

Table 2. Measures of Data Quality Based on Sample Statistics

Variable	Net Worth						Debt To Income					
Statistic	Mean			Median			Mean			Median		
	Orig	%diff	Duncan	Orig	%diff	Duncan	Orig	%diff	Duncan	Orig	%diff	Duncan
All Incomes	448230	0.05%	.00002	93098	0.1%	0.0001	0.2011	-0.03%	307787011	0.1236	0%	NA
Income Quintiles												
0-20	72620	1.53%	.00001	7496	1.01%	0.00017	0.3115	0.36%	816,027.4	0.0000	0.00%	NA
20-40	122037	-1.55%	.0000	34348	-1.02%	0.00001	0.1664	-1.00%	361,589.2	0.0783	-1.55%	677,404
40-60	193820	-0.62%	.0000	71605	-0.10%	0.00018	0.1930	0.05%	9,245,5621	0.1508	0.73%	824,946
60-80	342800	0.75%	.0000	159950	0.03%	0.00040	0.1854	0.31%	303,5122	0.1796	0.60%	855,753
80-90	485006	-0.69%	.0000	311146	-0.63%	0.00000	0.1737	-0.11%	25,507,601	0.1728	0.10%	35,856,431
90-100	2534413	0.19%	.0000	924127	0.43%	0.00000	0.1245	-0.51%	2,433,795	0.1117	-1.17%	586,292

reflecting the large scale of the variable; the measure on the debt to income ratio is very large, reflecting the variable's much smaller relative scale. As a result, making cross variable comparisons is difficult, as is making a determination of whether the loss in utility is "big" or "small".

We repeated the exercise for a standard regression analysis, and report the results in Table 3. A major concern with the application of the type of techniques used in disclosure proofing the SCF is that parameter estimates will be biased down, standard errors will be biased up, and the consequences will be that null hypotheses will wrongly fail to be rejected. A visual inspection of the parameter estimates derived from both the original and the disclosure proofed data suggests that these fears are substantially unfounded: both the parameter estimates and the standard errors are substantially unchanged after the application of the disclosure protection techniques. This is confirmed by examining the percent standard errors, which are reported in the next column. However, the Duncan measures are not particularly useful in conveying the information to current and prospective users of the public use data.

Finally, we calculated a subset of the Domingo/Torra metrics, but chose the one based on correlations matrices in view of the scale issues discussed above. We chose a data matrix of four variables: financial assets, non-financial assets, debt and income. The MSE of the correlation matrix was effectively 0; the MAE was .05, while the MV was .13. This confirms that the effect of the disclosure protection on the quality of the input matrix was relatively minor.

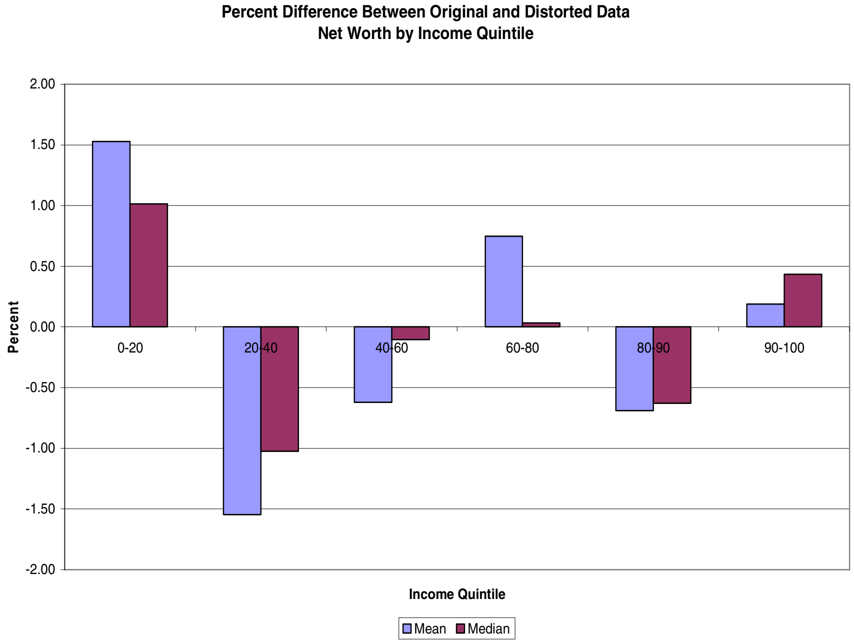


Fig. 1.

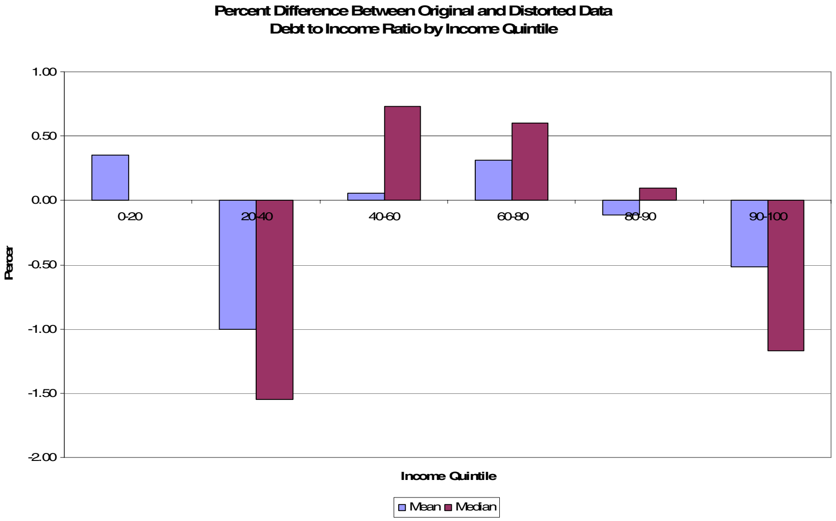


Fig. 2.

Table 3. Results of Standard Regression

	Original	**Distorted**	**Percent**	
	Data	**Data**	**Difference**	**Duncan**
	-11.9974	-12.0347	-0.31	718
Intercept	I.4336)	(.4369)	-0.75	94,500
	0.1771	0.1770	0.04	192,901,234
Age	(.0166)	(.0168)	-1.56	14,907,350
Age	-0.0931	-0.0929	0.21	26,570,305
(squared)	(.0154)	(.0157)	-1.44	20,108,990
	1.5196	1.5226	-0.20	107,076
Income	(.0266)	(.0269)	-1.33	8,025,102

Dependent Variable, Log of net worth; Standard errors in parentheses

6 Summary and Conclusion

The creation of public use datasets has been an important factor in advancing empirical social science research. National statistical institutes have rightly expended substantial energy to protecting the confidentiality of the respondents by using a variety of disclosure protection techniques. Recently, more attention has been paid to creating metrics that capture the impact of those techniques on data quality. This paper has demonstrated that those metrics, while possibly useful in summarizing the impact to the agencies themselves, are of limited use in conveying the information to researchers. Simpler measures, such as the percentage change in parameters from commonly used analytical work, might be more appropriate. The work of Karr et al.

(2006) which suggests the use of the Kullback-Liebler metric is appropriate, could also be usefully investigated [8]. In further research, we intend to examine the impact of different types of protection techniques, such as topcoding and rounding, on data quality using these different metrics and using common estimation techniques.

References

1. Athey, L. and Kennickell, A. “Managing Data Quality on the 2004 Survey of Consumer Finances” Proceedings of the American Statistical Association (2005)pp. 3767-3774
2. Bucks, B., Kennickell, A. and Moore, K. “Recent Changes in U.S. Family Finances: Evidence from the 2001 and 2004 Survey of Consumer Finances,” Federal Reserve Bulletin, (2006), pp. A1-A38.
3. Domingo-Ferrer, J. and Torra, V. “Disclosure protection methods and information loss for microdata", in Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies, eds. P. Doyle, J. Lane, J. Theeuwes and L. Zayatz, Amsterdam: North-Holland, (2001) pp. 91-110.
4. Duncan, G., Fienberg, S., Krishnan, R., Padman R. and Roehrig, S., Disclosure Limitation Methods and Information Loss for Tabular Data in Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies, eds. P. Doyle, J. Lane, J. Theeuwes and L. Zayatz, Amsterdam: North-Holland, (2001) pp. 135-166
5. Fries G., Johnson, B. and Woodburn, R., "Analyzing the Disclosure Review Procedures for the 1995 Survey of Consumer Finances," http://www.federalreserve.gov/pubs/oss/oss2/method.html) (1997)
6. Haworth, M., Bergdahl, M., Booleman, M., Jones, T., and Magaleno, M., “LEG chapter on Quality Framework,” Proceedings of Q2001, Stockholm, Sweden, May 2001, CD-ROM (2001).
7. Karr, A., Sanil, A. and Banks, D. Data Quality: A Statistical Perspective NISS Technical Report Number 151, (2005)
8. Karr, A., Kohnen, C., Oganian, A., Reiter, J. and Sanil, A. “A Framework for Evaluating the Utility of Data Altered to Protect Confidentiality,” NISS Technical Report Number 153 (2005),
9. Kennickell, A., “Wealth Measurement in the Survey of Consumer Finances: Methodology and Directions for Future Research,” working paper, http://www.federalreserve.gov/pubs/oss/oss2/method.html. (2000)
10. Kennickell, A. (November 1997), "Multiple Imputation and Disclosure Protection: The Case of the 1995 SCF", http://www.federalreserve.gov/pubs/oss/oss2/method.html.
11. Kennickell, A., “Multiple Imputation in the Survey of Consumer Finances,” Proceedings of the Section on Survey Methods Research, Annual Meetings of the American Statistical Association, Dallas, (1998) pp 3767-3774.
12. Winkler, W., “Methods and Analyses for Determining Quality,” Keynote address at the 2005 ACM SIGMOD Workshop on Information Quality in Information Systems (2005) (http://iqis.irisa.fr/UTH).
13. Zayatz, L., "Disclosure Avoidance Practices and Research at the U.S. Census Bureau: An Update", Research Report Series (Statistics #2005-06), Statistical Research Division, U.S. Census Bureau, Washington, D.C. (2005)

Protecting the Confidentiality of Survey Tabular Data by Adding Noise to the Underlying Microdata: Application to the Commodity Flow Survey*

Paul Massell, Laura Zayatz, and Jeremy Funk

Statistical Research Division
U.S. Census Bureau
Washington, D.C. 20233
{Paul.B.Massell, Laura.Zayatz, Jeremy.M.Funk}@census.gov

Abstract. The Commodity Flow Survey (CFS) produces data on the movement of goods in the United States. The data from the CFS are used by analysts for transportation modeling, planning and decision-making. Cell suppression has been used over the years to protect responding companies' values in CFS data. Data users, especially transportation modelers, would like to have access to data tables that do not have missing data due to suppression. To meet this need, we are testing the application of a noise protection method (Evans et al [3]) that involves adding noise to the underlying CFS microdata prior to tabulation to protect sensitive cells in CFS tables released to the public. Initial findings of this research have been positive. This paper describes detailed analyses that may be performed to evaluate the effectiveness of the noise protection.

Keywords: Disclosure Avoidance, Noise Protection, Tabular Data, Microdata.

1 Background: History of Disclosure Avoidance Methods Used for Commodity Flow Survey (CFS) Tables

Title 13 of the U.S. Code requires that the Census Bureau maintain the confidentiality of information provided by respondents. The Census Bureau has conducted the CFS in 1993, 1997, and 2002. Specifically, CFS data provide information on shipments originating from manufacturing, mining, wholesale, auxiliary warehouses, and selected retail establishments in the 50 states and the District of Columbia. While it is an establishment survey, its focus is on the characteristics of shipments, not on the establishments or the companies who transport the goods. The survey's goals are to estimate the characteristics associated with the origin and ultimate destination of shipments, the distances traveled by goods, the commodities shipped, the modes of transportation used to transport the goods, and the volume of the shipments measured by weight and value.

[*] This report is released to inform interested parties of ongoing research and to encourage discussion of work in progress. The views expressed on statistical issues are those of the authors and not necessarily those of the U.S. Census Bureau.

J. Domingo-Ferrer and L. Franconi (Eds.): PSD 2006, LNCS 4302, pp. 304 – 317, 2006.

In 1997 and 2002, many cells in the most detailed tables were suppressed. The main reason for suppression was a cell exceeding a data quality threshold. However, complementary cell suppression disclosure avoidance methods have forced additional suppressions as well. Users have voiced opposition to this latter source of cell suppression. In the 1993 CFS, and to a lesser extent in 1997, cell suppression affected the usability of the data by eliminating certain table cell information in a non-random fashion. For certain commodity flow models, this type of missing data renders the model useless. This provided motivation for exploring the use of a protection method that allowed publication of a larger number of cell values, even if that involved perturbing those values.

2 The EZS Noise Method: General Properties, Evaluation, and Calibration

2.1 General Properties

The particular noise addition method that we chose for testing protection of CFS tables is one we call the EZS noise method. It was developed in the late 1990's by Tim Evans, Laura Zayatz, and John Slanta, all of whom were, at that time, mathematical statisticians at the U.S. Census Bureau (Evans, Zayatz, Slanta, [3]). This method has the following appealing properties.

Adding Noise to Microdata. The EZS method adds noise to the microdata underlying the tables rather than directly to table cell values themselves. Assuming for now that in all tables the cells represent values of the same magnitude variable, a single pure noise factor (i.e., a multiplier) is generated for each record. The perturbed value for the magnitude variable is computed with a simple formula that involves the single noise factor and the weights for the record. Tables are generated from the microdata using the perturbed values from each record in the same way that tables are generated from non-perturbed weighted values when noise is not used. Therefore the code for table creation does not have to be modified, and the tables generated are automatically additive and receive disclosure protection in a consistent way. Thus the issue of protecting linked tables, which is often difficult to achieve with other protection methods, is automatic for this method, at least for tables generated from a single microdata set.

An Approximate Value Can Be Released for Every Cell. Noise methods generally have the property that an approximate value is produced for each cell. Often these approximate, or perturbed values, are close enough to the true values so that they are still useful to data users for various statistical purposes. By contrast, in tables protected by cell suppression, the suppressed cells appear to yield no information, although with some effort an adept data user can exploit the additivity of the table in order to associate an uncertainty interval with each suppressed cell. However, the uncertainty interval for a suppressed cell is typically much larger than the implicit uncertainty interval for a perturbed cell. That is, a perturbed value conveys more information than a suppressed cell about the true cell value. Since most tabular statistical procedures are designed to work with complete tables, protection methods that produce complete tables of perturbed values are generally more useful to statistical modelers than methods that produce tables with suppressed cells.

Protection at the Company Level. The U.S. Census Bureau is required to protect economic data at the company level. Typically, companies consist of two or more establishments at different locations. In the CFS, each company responds to the survey with data for each of its establishments. Cell values often include data from only a single establishment of a company since cell values are often defined partly by geography. Of course, Title 13 of the U.S. Code requires that establishment values must be protected from disclosure, but that is not enough. In addition, the sum of values for any set of establishments for a given company must also be protected. This includes the company value, which is the sum of the values for all its establishments. Protection at the company level requires complicated code when implemented for cell suppression; however, for the EZS noise method company level protection is easy to implement.

Ease of Implementation. The EZS algorithm involves only a small number of formulas, each of which is quite simple to code. A statistical office would probably choose to write a short program (only a page or two of SAS code) that generates noise for the relevant magnitude data variable in each microdata record. Run time for this noise program is short (e.g, a few minutes for microdata that consisted of about two million CFS records). Then the office runs the original table generating program against this 'noisy' microdata file for each table of interest.

2.2 Evaluation

Any method for protecting statistical tables should be analyzed statistically and/or mathematically to evaluate its effectiveness. First, one needs to define a measure that is applied to each cell that determines whether a cell value can be released (published) as is or requires modification to protect it. We call cells that require protection "sensitive cells." For economic magnitude data at the U.S. Census Bureau, sensitivity is usually based on the p% rule (WP22, [7]). Since the CFS is a sample survey, in contrast to a complete census, we need to use the generalized form of the p% rule that accounts for the protection provided by weights (e.g., sampling weights or non-response weights) that are unknown to table users. Also, CFS cell values are rounded, and rounding provides additional protection. Thus we use a version of the p% rule that accounts for protection due to rounding and weighting.

Perturbation of Microdata for Protecting Table Cells. Since the EZS method directly perturbs the microdata underlying the tables to be generated, one could say that the effect of noise addition in EZS on cell values is indirect. We say this because a pure noise factor is generated for each microdata record, and is then used along with weights to produce a perturbed value for that record. Finally the perturbed values for all the records associated with a given cell are summed to produce a perturbed table cell value. Since the pure noise factor is generated by a random number generator, all the quantities derived from the pure noise factors, including the cell value, are unpredictable. The absolute value of the amount of perturbation that is required for a cell, based on the protection rule being used, we call the "nominal" amount. In our applications here, it is based on the same extended p% rule that determines sensitivity. The absolute value of the difference between the perturbed cell value and the true (unperturbed) cell value; i.e., the absolute value of the actual perturbation, is usually greater than the nominal value. That is, most cells receive sufficient protection as measured by the p% rule. However, often a small percentage of cells receive perturbation that is less than the nominal amount. The notion of alpha error (defined below) is designed to measure the prevalence of these cases. Note however, that when a cell consists of data from a single company, the EZS algorithm ensures that the cell is perturbed by at least its nominal amount. This ensures that a company's unweighted contributed value in a "one-company" cell will never be published without a significant modification.

Alpha Error: Measuring Under-Perturbation. "Indirect perturbation" may have a drawback since it is based on (random) noise generation. It may not provide the desired (or 'nominal') amount of perturbation for some cells. That is, some cells may be perturbed less than the p% rule "requires." Of course, to be an acceptable method, it must provide the 'nominal' amount of perturbation for almost all sensitive cells. We may tolerate under-protection for a small percentage of sensitive cells. Even for this small set of cells, it is desirable that most of the cells receive a significant fraction of their nominal perturbation. In this context, we use the term "alpha error" to denote the amount of under-protection of sensitive cells. A possible formula for this will be given below. It involves the ratio (denoted 'frac') of actual perturbation to nominal perturbation when that ratio is less than 1. The percentage of sensitive cells with a value of frac less than 1 provides a global measure of under-protection and the specific values of frac provide a local measure. We currently define the alpha error as the sum over sensitive cells with frac less than 1 of (1-frac) divided by the total number of sensitive cells (#SEN). This quantity always lies in the interval [0,1]. If all sensitive cells are fully protected alpha = 0. If all sensitive cells receive no protection, alpha = 1. The acceptable upper limit for this quantity is an agency policy decision. Thus, a formula for this alpha error is:

(1/#(SEN)) times the sum over sensitive cells i of max(0, 1 – frac (i)) .

This formula is not applicable if there are no sensitive cells, and in this case alpha can be interpreted as 0 since there is no under-protection occurring. Alpha is also zero if all sensitive cells receive at least their nominal perturbation. Both of these cases occur among the five tables discussed in section 3. The acceptable level for alpha error is a decision for the group at an agency tasked with setting policy for disclosure avoidance.

Beta Error: Measuring Over-Perturbation. In addition to measuring the degree of under-protection, it is helpful to measure the amount of over-perturbation as well. Over-perturbation does not increase disclosure risk but it is an important data quality issue that must be considered. Data that are greatly over-perturbed are likely to be of substantially lower quality than the original data, and will therefore be of less value to data users. If a particular application of the EZS noise method, or any other protection method, greatly over-protects the data, it may be possible to adjust some parameters so that a second application will less over-protection while still keeping the alpha error small. However, due to the inherent randomness in the EZS noise process, there will always be some amount of over-protection. We define the 'relative perturbation' (rel-pert) in the cell value as: ('post' is the post-perturbed value, and 'pre' is the pre-perturbed value).

$$\text{rel-pert} = |(\text{post} - \text{pre})| / \text{pre}$$

We currently define "beta error" as the weighted average of these relative changes over all non-sensitive cells (also called 'safe' cells) and over all sensitive cells that receive full protection. Actually, for such sensitive cells we use the relative over-perturbation ('rel-over-pert') defined as {|(post-pre)| – |(nominal-pre)|}/ pre. The contribution of this subset of sensitive cells to the beta error will likely be very small. We include it since there is some data quality cost to over-perturbing sensitive cells (i.e., perturbing them more than their nominal amount). Expressed in a formula, this definition of beta error is:

{(#SAFE) * (average of rel-pert for safe cells) +
(#P-SEN) * (average of rel-over-pert for protected sensitive cells)} /
(#SAFE + #P-SEN) .

where #SAFE is the number of safe cells and #P-SEN is the number of sensitive cells which receive full protection. If sensitive cells comprise a small percentage of cells in a given table, the beta error may be approximated as the average value of rel-pert for safe cells or the average for all cells. We used this latter approximation to compute beta errors for the five tables discussed in section 3. Note that beta lies in the interval [0,1] and the acceptable upper bound might be based on data use considerations since such a bound affects overall quality of the tabular data. For example, a value such as 0.05 might be low enough for certain uses.

The Impact of EZS Perturbation on Coefficients of Variation. Data reliability is measured at the cell level by using Coefficients of Variation estimates (abbr. CV). A CV is computed by estimating a variance for the cell and dividing it by the estimated cell value. There are a variety of ways to estimate variances. The method of random groups was chosen by the branch at Census with the responsibility for selecting computational algorithms for use in economic surveys. The random groups method (Wolter, Chapter 2, [6]) computes variances directly from the microdata rather than using formulas. It's likely that no modification of the random groups procedure is required to estimate variances for EZS modified microdata. However, in general the CV for a pre-perturbed cell will differ from the CV for the post-perturbed cell. It is hard to predict how much CV's will be modified by an application of the EZS noise method. In section 3, we report on the distribution of the CVs for several CFS tables after EZS noise has been applied to the CFS microdata. The resulting modifications to CVs are minimal; e.g., often over 85% of the cells in a table have CVs that differ by less than 1% from their value based on pre-perturbed cell values.

Overview of Operations That Contribute to Protecting the Confidentiality of Data in Tables. There are several ways to protect the confidentiality of data presented as cell values in tables. The sole purpose of the EZS noise method is the protection of tabular data. There are various other operations applied to the data in both microdata and tabular form during production that are not explicitly designed to protect the data, but is nevertheless a positive side effect. For example, CFS tables are protected by weights and scaling factors that are applied to the microdata and by the process of rounding cell values (Massell, [5]).

Protection Due to Weighting. There is an extension of the p% rule that includes the protection provided by weights. According to Working Paper 22 on Disclosure (WP22, [7]), when one determines the sensitivity of a cell value one needs to compute a cell total that uses all the relevant weights (e.g., sampling weights, adjustment weights, etc.). We denote this as 'TA'. Then one computes the largest two company values contributing to a cell value, denoted X1 and X2. Usually, these values are computed **without** weights. One computes the remainder ('rem') as follows:

$$\text{rem} = \text{TA} - (\text{X1} + \text{X2}).$$

Finally one computes the protection required (prot) where

$$\text{prot} = (p/100)*\text{X1} - \text{rem}.$$

Here, 'p' is the value used in the p% rule; it's chosen by the statistical office. If prot > 0, the cell is defined as sensitive and the amount of perturbation required equals prot.

To see more clearly how weights protect cell values, consider a cell with contributions from only two companies. Then the expression for 'rem' above yields:

$$\text{rem} = (\text{w1} - 1) * \text{X1} + (\text{w2} - 1) * \text{X2}.$$

Here, the weights w1 and w2 generally satisfy $\text{w1} \geq 1$ and $\text{w2} \geq 1$. For fixed X1 and X2, as w1 and/or w2 increase, rem increases; therefore 'prot' decreases. Even for small weights, prot may become zero; e.g., if w2 = 1, and w1 = (1 + p/100), prot = 0.

For the CFS, there are 7 weights for each shipment. Some of these might be described as 'adjustment factors'. Four of these weights are generally unknown to data users, but 3 of these weights are often known. We believe the correct way to handle known weights is to multiply each shipment for company 1 by the product of its 'known' weights and then sum these values to form X1, and similarly for company 2 and X2. (Since these weights are known, table users could apply these weights to their own contributed values before subtracting the product from the cell value.) As before, TA should involve all the weights for each shipment.

There is an interesting relationship involving confidentiality, data quality, and sample size. An agency has limited funds for conducting any survey and this generally limits the size of the available sample. In past cycles of the CFS, many cells were suppressed due to high CVs. If sample sizes were to increase in future cycles of the CFS, we would expect the number of cells that have high CVs to decrease, allowing more cells to be published. Some of the cells that previously were suppressed due to having high CVs, might still have to be suppressed, but now because they are sensitive. This occurs because as sampling weights decrease, cells that receive protection from those weights, especially one and two-company cells, now may become sensitive.

Protection due to Rounding. In many CFS tables, cell values are rounded to the nearest million. This rounding prevents table users from seeing a very precise estimate of a company's or establishment's value. The exact amount of protection provided by such rounding is expressed in a formula derived by Laura Zayatz and described in a note (Zayatz, [8]).When data are to be rounded to six digits (i.e., to the nearest multiple of 1,000,000), the formulas below are used. For this type of data, the rounding protection is added to the protection reflected in 'rem' (where rem = TA - (X1 + X2)); thus the cell is sensitive if:

$$\text{rem} + |500{,}000 - |\text{TA} - \text{round}(\text{TA})|| < \text{X1} * (p/100).$$

If it is sensitive, then the following amount of additional protection is required to protect it:

$$\text{prot} = X1 * (p/100) - [\text{rem} + |500{,}000 - |TA - \text{round}(TA)||] .$$

One can view this complicated expression as the difference in perturbation required and the sum of uncertainties contributed by various sources.

The Interaction of Weights and Noise in Determining Cell Value Changes. The EZS method is described in a paper that appeared in a special issue on disclosure of the Journal of Official Statistics (Evans, Zayatz, Slanta, [3]). There is also a recent note on the analysis of the interaction of weight and noise in the EZS method (Massell, [4]). Specifically, it is important to note that even though the noise distribution always creates pure noise factors that involve at least a k% change (e.g., k might be in the interval [5,15]) for each company's contribution to a cell value, the weights associated with the company's shipment usually lower the actual effect of the noise, sometimes substantially. Thus even for a cell with all its data from a single company, the relative change in the cell value is often less than half of the change due to the pure noise factor; this will happen if $w > 2$. This can be shown with simple algebra using the expression:

$$\text{joint noise-weight multiplier} = \{ \text{(noise multiplier)} + (\text{weight} - 1) \}$$

that is applied to each contribution to a cell value. Typically the noise multiplier has a value between say, 0.5 and 1.5, but is not very close to 1. Weights are usually ≥ 1.

In some of our test tables, the weights fully protect many of both the one-company and the two-company cells. This is in contrast to census data (or more precisely, data for which the weights equal 1) for which cells that consist of data from only one or two companies always require protection. Thus, survey data in which most records have weights larger than $(1 + p/100)$ are likely to have few sensitive cells assuming sensitivity is determined using the extended p% rule.

2.3 Calibration

Finding a good noise distribution to use for the EZS method may require some calibration. Initially, the statistical office must select the shape and location of the noise distribution. Experimentation will often lead to parameter values for the distribution that produce acceptably small values of alpha and beta error. If an office has the time, it can continue experimentation to find parameters that produce an optimal distribution, such as one which minimizes some given linear combination of alpha and beta error. There is no theoretical reason why the same noise distribution should be optimal for all surveys. Although we have done only a modest amount of experimentation on this problem, one would expect the optimal distribution to vary somewhat among even surveys involving the same type of data. In fact, the optimal amount of noise is a function of 1) the survey microdata, 2) the set of tables to be generated from it, 3) the amount of protection required for each cell, and 4) the acceptable levels of alpha and beta errors (or similar error measures). For convenience, the statistical office would probably want to use the same distribution for different microdata sets as long as the errors were not far from optimal.

In our testing to date, we've used two noise distributions which we denote [D1] and [D2]. For some tables we used both (on different runs) to compare results. These two distributions have the same structure. They are "split" triangular distributions, such that the density function is described below and illustrated in figure 1.

Let $1 < a < b < 2$.

to the left of 1: $f(x) = k \cdot (x - (2-b))$ for $2-b < x < 2-a$
around 1: $f(x) = 0$ for $2-a < x < a$
to the right of 1: $f(x) = (- k) \cdot (x - b)$ for $a < x < b$.

$f(x) = 0$ otherwise (i.e., for $x < (2-b)$ and for $x > b$)

Here $k = (1/(b-a)^2)$ since the area under the density curve must equal 1.

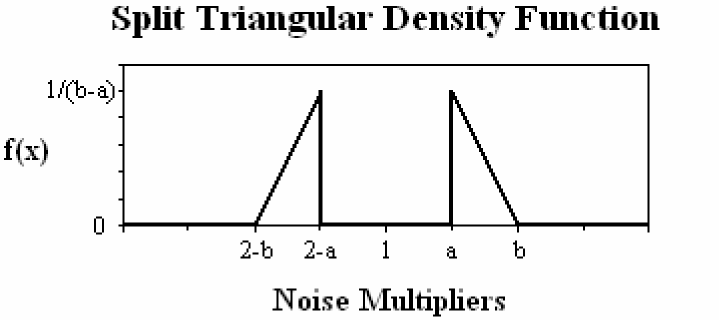


Fig. 1. The Split Triangular Distribution

Thus the density is piecewise linear and is symmetric about 1 ([3]). The following distributions were used for the pure noise multiplier on different test runs.

[D1] noise: split triangular distribution on [0.8, 0.9] and [1.1, 1.2]
[D2] noise: split triangular distribution on [0.85, 0.95] and [1.05, 1.15]

One can develop approximate relationships between the noise distribution used and the resulting alpha and beta errors. One can show using simple algebra that if one uses the split triangular distribution with $a \geq 1 + (p/100)$, then all single company cells will be perturbed by an amount large enough to guarantee that they will be safe. For example, if p=10, the [D1] distribution, since it has a = 1.1, satisfies this inequality. (They will be perturbed by an amount greater than the 'nominal protection' = sum over company records i of $(p/100 + 1 - w_i) \cdot x_i$; where $w_i \geq 1$ and $x_i \geq 0$). Since for p = 10, all single company cells are safe (i.e., non-sensitive), we expect [D1] noise to produce a very low alpha error. The beta error may be a bit too large reflecting excess noise that leads to excess perturbation of safe cells, and over-protection of sensitive cells. Moreover, since a substantial percentage of the microdata records have large weights, and large weights provide inherent data protection, the perturbation required to protect cells that contain such data is either small or zero. Thus it likely that a noise distribution which modifies values by less than 10% prior to the weight adjustment would produce acceptably small values for both alpha and beta errors. This idea motivated our selection of [D2] noise in which the pure noise multiplier modifies values sometimes by as little as 5%. We plan to experiment with other simple noise distributions until we are confident we have found one that produces a near-optimal combination of alpha and beta errors. In practice, a statistical office need not find a distribution that produces near-optimal errors, but simply one that produces acceptably small errors.

3 The EZS Noise Method: Results of Testing on Selected 2002 CFS Tables

In this section we present a summary of outputs from an analysis program (written in SAS) that was run on five 2002 CFS tables selected to represent a variety of levels of geographic detail (referred to below as CFS tables 1 through 5).

Two of the important statistical quantities reported are defined as follows: the percentage change in the cell estimate and cell CV due to noise perturbation. We define the change in cell estimate as (rel-pert * 100) where ‘rel-pert’, the ‘relative perturbation,’ is defined above. Recall the expression for the coefficient of variation (CV) for a cell value, given in the section on beta error above. We compute CVs using perturbed values for both the numerator and denominator.

Sensitivity of cells is determined using the p% rule; we use the version that takes into account weights assigned to the lowest level microdata records. For CFS data, these are the shipment level records.

For each table we report the number of cells; in general, for a fixed number of dimensions, as the detail of the table increases so does the number of cells. We report the distribution of the number of companies that contribute to a cell. It is possible that there is more than one establishment for some companies contributing to a cell and typically there are many shipment records for each establishment. Because weights play a major role in protecting against disclosure for CFS data, it is not uncommon for a cell with data from only a single company not to be considered sensitive. Sparse data has been a major problem for the CFS, such that in tables with even a moderate level of detail the CV’s for cells are often so large that the cells are suppressed. Because of this high level of uncertainty it was not unusual for a table to have more cells suppressed due to poor data reliability (i.e., large CV’s) than for disclosure protection purposes.

For those tables that contain sensitive cells, we present an analysis of the set of sensitive cells similar to that for all cells; e.g., we compute the number of companies contributing to each such cell. We also compute a ‘protection multiplier’ for each sensitive cell. This we define as the ratio of the (absolute value of the) perturbation assigned to a sensitive cell to the desired protection. For detailed tables, there are usually a small percentage of cells that do not receive full protection; however, in most of those cases, they do receive at least 50% of the desired protection. We report the number of cells not receiving full protection and the lowest value of the protection multipliers. We use the alpha error as a global measure for the amount of underprotection in a given table. We report the range of values for the protection multipliers that includes about 90% of the middle part of the distribution to give some idea of the typical amount of “over-perturbation”; i.e., perturbation that exceeds the desired protection. To assess the decrease in data quality due to perturbation we also report an approximated beta error for each table.

3.1 Analysis of Selected CFS Tables

Table 1 above represents the changes in cell estimates and CVs for the national 2-digit commodity table, a relatively high level table with only 43 cells, none of which are sensitive. There was very little data distortion added to the table due to noise addition at this level of detail. Since the D2 noise distribution generates less noise than the D1 distribution, one would expect changes to both cell estimates and CVs to be smaller under D2. This was the case in each of our five test tables. Of course, if one uses a distribution that generates very little noise, there is the risk of greatly under-protecting the sensitive cells. When this occurs, it may be reflected in a large alpha error.

Table 1. CFS Table US 5a (USA by 2-digit Commodity) (43 cells, none sensitive)

Perturbation Interval	Change in Cell Estimate (% of All Cells)		Change in Cell CV (% of All Cells)	
(% Change)	D1 Noise	D2 Noise	D1 Noise	D2 Noise
0 to 1%	95.3%	95.3%	86.0%	93.0%
1 to 2%	2.3%	4.7%	9.3%	4.7%
2 to 3%	2.3%	0.0%	4.7%	2.3%

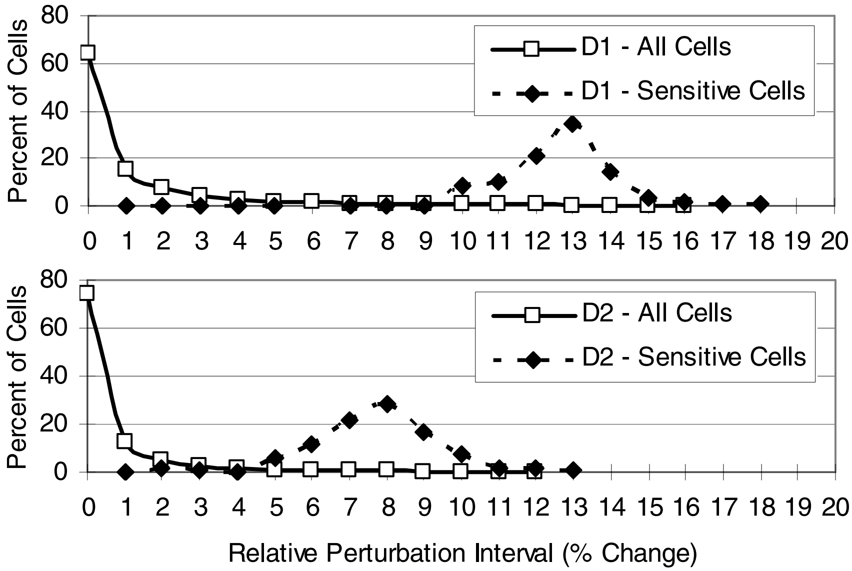


Fig. 2. The above plots represent approximations of the distribution of noise added to table cells for CFS Table 3 (Origin State by Destination State by 2-Digit Commodity), for noise distributions D1 and D2 respectively. A point (x, y) can be interpreted as the percentage y of cells that received a relative perturbation that lies in the interval [x, x+1) percent.

Figure 2 above represents the amount of perturbation added to cell estimates due to the noise procedure for D1 and D2 distributions respectively. Note the peak change for sensitive cells is in the range 13% to 14% for D1 noise and in the range 8% to 9% for D2 noise. This corresponds to the fact that the D1 distribution is shifted 5% to the right relative to the D2 distribution. The smaller perturbations created by D2 are probably sufficiently protective when p=10 is used, since only 5 of the 230 sensitive cells fail to receive the nominal perturbation, with the lowest protection multiplier (PM) being 0.8. For D1 noise, 2 PM's are less than 1. For PM's greater than 1, the range of values for D1 is roughly shifted by 1.0 compared to that for D2; this represents a significant amount of over-perturbation caused by using the D1 distribution.

3.2 Analysis of Remaining CFS Tables

For all cells in a given table, we looked at the distribution of the number of companies that contribute to the cell. We also report the percentage of cells that are suppressed due to having a CV value that exceeds 50%. (It turns out that CVs for these tables are affected very little by either [D1] or [D2] noise). We give the number of sensitive cells, the percentage that have CVs exceeding the threshold, and the distribution of the number of companies that contribute to them. For each noise distribution we present information about the protection multipliers, which are defined as the ratio of the absolute value of the perturbation for a cell divided by the perturbation required for that cell by the (extended) p% rule. We give a middle range of values that includes about 90% of the distribution, the number of values less than 1, and the lowest value for these protection multipliers.

We tested the EZS noise protection method on five CFS tables, each of which is derived from the full CFS microdata. The tables range from 2 to 4 dimensions, including a national table with no detail and no sensitive cells to tables with much detail and many sensitive cells. Certain general observations can be made from an analysis of these five tables. First we give the structure of each table followed by the number of its cells to get an idea of the size and level of detail of each. Table 1 is "U.S. by 2-digit commodity" (43). Table 2 is "Origin State by 2-digit commodity" (2,108). Table 3 is "Origin State by Destination State by 2-digit Commodity" (61,174). Table 4 is "Origin Metropolitan Area by Commodity by Mode" (31,622). Table 5 is the largest and most detailed; it is "Origin State by Destination State by 2-Digit Commodity by Mode" (389,632). It is not surprising that the percentage of cells with a CV value exceeding the data quality threshold increases as the level of detail increases. In fact, the percentages of cells that exceed the CV threshold for the 5 tables, are (0, 27, 70, 58, 75) % respectively.

As the number of cells increases, the number of sensitive cells increases. For the five tables, the numbers of sensitive cells are (0, 9, 230, 224, 1197). This phenomenon is certainly understandable for unweighted data for which any cell with contributions from only one or two companies ("one-company" or "two-company" cells) is sensitive according to the p% rule. Our results show that it appears to occur even when using weighted data. In fact, for the five tables, the percentages of all cells that are either "one-company" or "two-company" cells are (0, 8, 37, 33, 47) %. For sensitive cells, the percentages of such cells are sometimes very high; for Tables 2 through 5, the percentages are (55, 87, 94, 93) %.

We performed a careful analysis of the sensitive cells. For Table 1, there were no sensitive cells. For Table 2 with D1 noise protection, most protection multipliers (PMs) were in the range (2, 4). There were none below 1, in fact the lowest = 2.3. Thus alpha =0. With D2 noise protection, most PMs were in the range (1.4, 2.4) but again there were none below 1, and the lowest was 1.2. For Table 3 with D1 noise, most were in the range (2, 4) with just two PMs below 1, the lower being 0.44. For D2 noise, most PMs were in the range (1.1, 2.4) with five PMs below 1, the lowest being 0.80. For Table 4 with D2 noise, most PMs were in the range (1.1, 2.2) with only one PM below 1, its value was 0.52. For Table 5 with D2 noise, most PMs were in the range (1.0, 4.0) with 31 PMs below 1 with the lowest being 0.08.

Table 2 below represents calculated alpha errors and approximated beta errors for all applications of the EZS noise method on our test tables. Alpha is 0 in table 1 for both distributions in table due to the fact that there are no sensitive cells. In table 2, all sensitive cells receive sufficient protection and so the alpha errors are both 0 here as well. Since beta error is designed to measure over-perturbation we would expect it to increase both with increased table detail and increased amount of noise added. This appears to be the case in our results, such that D1 always produces larger beta errors than D2, and in general the beta errors increase with the number of cells in the table. It is important to keep in mind the random nature of this method, and therefore a simulation study involving several iterations of the random noise multipliers would be appropriate in order to make concrete conclusions from this type of analysis.

Table 2. Alpha and Beta Errors for Five CFS Test Tables

CFS Table	Alpha Error		Beta Error Approximation	
	D1	D2	D1	D2
1	0	0	0.0057	0.0055
2	0.0000	0.0000	0.0122	0.0089
3	0.0031	0.0028	0.0164	0.0113
4	NA	0.0021	NA	0.0103
5	NA	0.0063	NA	0.0109

3.3 Comparison of EZS Noise with Cell Suppression

Certain aspects of the noise protection process are a matter of agency policy. In particular, an agency needs to decide on what "publication rules" are suitable for the data at hand. An agency may choose to publish a value in all cells, and simply put an extra symbol in those cells whose values have undergone a percentage change that exceeds some data quality threshold due to the added noise. We say these cells have been 'flagged.' Most sensitive cells would typically be in this category. A small fraction of the non-sensitive cells would also. Finally the agency would probably choose to include all sensitive cells in this category simply as a way to protect these values by discouraging their use. Rather than flagging such cells, an agency might decide to suppress all sensitive cells in a noise protected table. Even in this case, there are several differences between using a noise protected table versus one protected with cell suppression. In a suppression protected table, the sensitive cells are suppressed and there are usually some (and sometimes many) secondary suppressions whose role is to make it impossible to recover the exact value of any sensitive cell or even a good estimate of any one. There is nothing comparable to secondary suppression in noise protected tables. In this scenario, the most obvious improvement to table users occurs when there are many secondary suppressions; under noise protection these are replaced with values. However, the advantage to the table user is not always clear because these formerly suppressed cells, while available, may differ from the "true" values by several percent due to the noise added. In fact, many published values in the table will typically be changed by a few percent and any small set of values of interest, even if they are not in flagged cells, may be changed by an amount just below the (data quality) threshold for flagging. For certain technical reasons, we will not report on the secondary suppressions for the set of CFS tables mentioned above. However, in general, for recent CFS cycles, many of the tables have had a number of secondary suppressions that is of the same order of magnitude as the number of sensitive cells. Thus, if noise protection is used for future CFS tables, we expect to see a significant number of cells with publishable values that would have been suppressed under cell suppression.

4 Conclusions and Next Research Steps

Based on the results from the five 2002 CFS tables that we selected for testing the effectiveness of EZS noise protection, it appears that EZS noise should be considered as an alternative to cell suppression for protecting future CFS table releases. We say this because the alpha and beta errors are quite low for both distributions tested. The current implementation seems to produce high quality tables that protect the data and are still very useful to data users.

For the CFS tables that we protected with noise, the only magnitude variable we considered was 'shipment value'. There are other magnitude values that are released for each cell, e.g., the shipment physical weight (tonnage). The question here is how these magnitude variables should be treated together with noise. Currently we intend to generate a single noise multiplier for each microdata record and apply that multiplier to each magnitude variable in that record. A single multiplier suffices for the CFS since the ratio of these two quantities is not a sensitive value. For other surveys, with two or more magnitude variables in which some of the ratios need to be protected, we may need to generate separate multipliers for different variables involved in these sensitive ratios.

There are various topics we need to explore. We need to decide how much to tell the data users about the noise procedure in order to maximize the usefulness of the tables for them while not compromising on confidentiality protection. Can users be told the technical details of the algorithm or just its most important aspects? We also need to decide whether we should publish a special symbol or flag along with the value in cells whose values have been perturbed by more than a certain percentage.

We also need to explore the best way to protect longitudinal data that are generated from successive cycles of the CFS. Specifically, does there need to be consistency in the selection of a perturbation direction for each company in the CFS from one conduct of the survey to the next? Does this consistency need to be maintained for "all time", or could the direction be changed after a number of years? Longitudinal data are often used for computing trends. Along these lines, there was a careful analytical study carried out by Evans regarding the effect of EZS (multiplicative) noise on the computation of trends (Evans, [2]). This paper not only presents interesting results but also implicitly shows the way that additional studies on the effect of noise on various types of models might be carried out.

There are some general research questions for the EZS method that are being explored currently. Some of these affect the use of EZS for CFS over time. For example, how do noise multipliers need to be adjusted when companies in the sample go out of existence or merge with other companies that have already been assigned noise multipliers? Preliminary indications are that such questions can be resolved in a way that maintains the effectiveness and ease of use of the EZS method.

Acknowledgments. Thanks to John Slanta of the Census Bureau and John Nelson (formerly at Census) for much help with the implementation of the EZS method. Thanks also to Neil Russell, formerly at the Bureau of Transportation Statistics (BTS), and Jonaki Bose (BTS) who led the BTS group that is working closely with a Census Bureau group on this project. Neil contributed some of the background information in section one and both he and Jonaki made numerous helpful comments on another version of this paper.

References

1. (CFS05) Commodity Flow Survey (CFS) Conference (2005), July 8-9, Boston, Massachusetts. Participant research questions at: http://www.trb.org/conferences/cfs/Workshop-DataProducts-Question.pdf
2. Evans, B. Timothy (1997), "Effects on Trend Statistics of the Use of Multiplicative Noise For Disclosure Limitation", ASA Proceedings of the Section on Government Statistics.
3. Evans, Timothy, Laura Zayatz, John Slanta (1998), "Using Noise for Disclosure Limitation of Establishment Tabular Data", Journal of Official Statistics http://www.jos.nu/Articles/abstract.asp?article=144537
4. Massell, Paul B. (2005), "The Interaction of Noise and Weighting in Protecting Company Data from Disclosure", unpublished note.
5. Massell, Paul B. (2006), "Using Uncertainty Intervals to Analyze Confidentiality Rules for Magnitude Data in Tables", http://www.census.gov/srd/papers/pdf/rrs2006-04.pdf
6. Wolter, Kirk (1985), Introduction to Variance Estimation, Springer,
7. (WP22) Federal Committee on Statistical Methodology (FCSM) (revised 2005), Working Paper 22, http://www.fcsm.gov/working-papers/spwp22.html
8. Zayatz, Laura (2000), "How rounding should be incorporated into the p% rule", unpublished note.

Italian Household Expenditure Survey: A Proposal for Data Dissemination

Mario Trottini[1], Luisa Franconi[2], and Silvia Polettini[3]

[1] Departamento de Estadística e I.O.
Universidad de Alicante
Apartado de Correos 99, 03080, Alicante, Spain
mario.trottini@ua.es

[2] Istat
Servizio Progettazione e Supporto Metodologico
nei Processi di Produzione Statistica
Via Cesare Balbo 16, 00184 Roma, Italy
franconi@istat.it

[3] Dipartimento di Scienze Statistiche
Università di Napoli Federico II
Via L. Rodinò 22 – 80128 Napoli, Italy
spolettini@unina.it

Abstract. In this paper we define a proposal for an alternative data dissemination strategy of the Italian Household Expenditure Survey (HES). The proposal moves from partitioning the set of users in different groups homogeneous in terms of needs, type of statistical analyses and access to external information. Such a partition allows the release of different data products that are hierarchical in information content and that may be protected using different data disclosure limitation methods. A new masking procedure that combines Migroaggregation and Data Swapping is proposed to preserve sampling weights.

Keywords: Disclosure Limitation, Sample Survey Data, Dissemination Policies, Risk-Utility Assessment.

1 Introduction

The Household Expenditure Survey (HES) is a sample survey on household expenditure carried out, every year, by the Italian National Statistical Institute (Istat). Household purchases of goods and services are the object of the survey. The goal of data collection is to provide the structure and level of expenditure according to the main social, economic and geographical characteristics of the Italian households. This information is essential to a wide range of statistical analyses of great interest for academic researchers, decision makers in the Public Administration, marketing and consulting companies, and other users.

The current data dissemination strategy comprises a set of tables and a masked microdata that are released to all users after completion of an application form and signing a pledge of confidentiality. The original microdata without direct identifiers and regional geographical detail may be accessed also for

J. Domingo-Ferrer and L. Franconi (Eds.): PSD 2006, LNCS 4302, pp. 318–333, 2006.

research purposes in the Data Analysis Centre at Istat. Finally, there are the institutional users. Official statistics in Italy is organized through a system (SISTAN - National Statistical System) whose entities may access, for institutional purposes, the original microdata without direct identifiers.

This paper reports on an ongoing work aimed to define a proposal for an alternative strategy of HES data release. The underlying idea is that HES users are very diverse and an efficient data dissemination strategy should take full account of this heterogeneity[1]. An alternative strategy could be designed by: (i) partitioning the set of users in different groups in order to increase the degree of homogeneity within each group; (ii) releasing a different data product to each group.

The paper is organized as follows. Section 2 describes the HES data. In section 3 we report on the results of a descriptive study that we realised to get a deeper understanding of HES data users and their needs. Section 4 reviews the current HES data dissemination strategy and explains why, we believe, this can be improved. A possible alternative is outlined in section 5. Section 6 describes the construction of a Public Use file as a part of the general strategy in Section 5 and a new masking procedure that preserves sampling weights essential for correct data users' inferences. Disclosure risk and data utility assessment for the Public Use file is also discussed here. Section 7 summarizes the main contributions of the paper and outlines open problems and ideas of future work.

2 The HES Data Set

HES considers a cross-sectional sample of Italian households selected from the current Italian Population Register. The HES relies on a two-stage sampling design with stratification of the municipalities, and systematic selection of the households. 479 municipalities have been selected for the year 2004 with roughly 2330 households each month for a total of around 28000 households in a year. The final sampling design weight of each household is derived from the basic weight by means of a calibration process in order to preserve known population totals. In the HES such totals are the population of each macro region (first level of the Nomenclature of Territorial Units for Statistics - NUTS1 - 5 classes) by sex and four age-classes (0-14,15-29, 30-59, 60 and more) and the population and number of households by administrative region (NUTS2, 20 classes). Starting from 1997 data are collected using two different techniques: completion of a diary reporting daily purchases on a variety of goods in a predefined seven days period and a face to face interview where socio-demographic information on household members are collected together with expenditures on housing, clothing, health, transport, leisure activities, education. When needed the household is requested to complete also another diary of goods produced by the household and used in the given week. Along with expenditure variables (approximately 480 classified

[1] The idea is not new in the statistical confidentiality literature, see for example, [9,2,1].

according to the divisions and groups of COICOP classification - Classification of Individual Consumption According to Purpose) sampling weights that are needed to derive population estimates from the survey sample are also supplied. For more information on the methodology of the survey and HES questionnaires see [6].

3 Heterogeneity of HES Data Users

Users of HES data are very diverse. An analysis of the application forms received by Istat in the period 2002-2004 (a total of 119) shows how HES users include academic researchers, decision makers in the Public Administration, marketing and consulting companies, public and private research institutes, international organizations and individuals. Type and sophistication of statistical analyses that need to be performed, skill and resources required to perform the analyses as well as motivation for disclosure are very diverse across users. Typical applications of the data among researchers include definition of equivalence scales, studies on poverty, definition and estimation of econometric models, changing of food consumption habits to mention just a few. Aims of the analyses for Public Administrations range from monitoring of prices to microsimulation models as well as analysis of consumption of specific services. Marketing and consulting companies naturally claim for marketing studies, but also for analysis of pattern of high expenditure and financial and welfare planning.

Despite the inaccuracies in the application forms (description of the planned analysis is often very generic), the non-exhaustive nature of the information represented in it[2], and the bias due to the reporting of analyses feasible solely with the provided masked version of the data, we believe that the information collected well illustrates the diversity of HES data users.

Another source of users' heterogeneity is their permission to access external data files for record linkage and re-identification. There are two main types of external registers in Italy: the Population Register (*Anagrafe*), i.e. the collection of information on all the people and households living in the municipality, and the electoral roll, the official list of people who are allowed to vote in the municipality. According to the Italian legislation the whole Population Register can be used only by Public Administrations for reason of public utility, whereas information on the presence of a *named* household in the municipality and its composition may be given to all who ask for it. Finally, the entire electoral roll of a municipality can be released for research purposes.

Taking into account these different aspects is essential for the design of efficient strategies of HES data dissemination, as well as for a correct evaluation and comparison of alternative strategies in terms of the *risk-utility* trade-off that these yield. The discussion of current HES data dissemination in the next section should illustrate the validity of this proposition.

[2] As we describe later there are alternative ways to access the data that do not require completion of an application form.

4 Current Data Dissemination Strategy of HES Data

Istat current data dissemination policy of HES data comprises:

(1) A set of marginal tables, HES_{1A}, published in a volume [6] and available on Istat web site, accessible by all users. The volume contains tables of specific expenditures at NUTS1 level as well as socio-demographic analyses of expenditures showing tables on purchases per number of household components, per type of household and per type of employment of the household reference person. Besides these, a whole series of marginal tables is given for monthly mean expenditure on goods and services collected in the survey by macro regions and number of components of the household, by type of occupation of household reference person and type of household. Finally, users may require Istat to release further *ad hoc* tables on specific matters; these can be provided if disclosure control rules allow.

(2) A masked microdata (also known as *File Standard*), HES_{1B}, obtained through a combination of suppression and global recoding also accessible by all users after completion and approval of an application form and a confidentiality pledge that users must submit to Istat.

(3) A microdata, HES_2, obtained by suppressing the direct identifiers in the original microdata and maintaining only NUTS2 geographical detail, accessible for research purposes in the Data Analysis Centre at Istat where the output of the analysis is reviewed for confidentiality checks.

(4) Finally, there are the statistical institutional users. Inside SISTAN the exchange of microdata[3] aims at the realization of direct surveys and other projects within the National Statistical Program along with studies for institutional purposes. In this framework statistical offices may ask for the HES microdata without direct identifiers; each request is processed centrally by Istat.

When, in 1992, Istat started releasing masked microdata the only users were academic researchers. Choice of transformations to be used and data dissemination strategy were intended to find an optimal balance between confidentiality protection and quality of released data for statistical analyses of interest to legitimate data users. Although the process has worked quite smoothly so far, we believe it is not completely satisfactory for at least two reasons. The analysis reported in Section 3 shows that, nowadays, HES data users and their needs are very diverse. As a result, the current data dissemination procedure that relies on a single transformed HES microdata for all users seems inefficient, since the masked data set and the set of tables produced are likely to be of insufficient detail for the more sophisticated data users, while disclosing information that is perhaps unnecessary for more basic research.

In addition, the current data dissemination strategy makes a rigorous assessment of disclosure risk and data utility associated with the procedure quite difficult. The set of tables, HES_{1A}, and the masked microdata, HES_{1B}, released to all

[3] Istituto Nazionale di Statistica. Deliberazione 20 aprile 2004, n. 9, *Criteri e modalità per la comunicazione dei dati personali nell'ambito del Sistema statistico nazionale.*

users, in fact, contain different information and are obtained using different disclosure limitation techniques. Sophisticated users intending to take full account of the information in the released data for the inferences of interest should combine the information in HES_{1A} and HES_{1B}. Combining these two sources of information, however, is usually a complex task due to different disclosure limitation techniques that produce the two data products. As a result a rigorous assessment of disclosure risk and data utility is usually very difficult even assuming the agency knowledge of users' targets, prior information, estimation procedures, etc.[4].

5 An Alternative Data Dissemination Strategy

Building on the limitations of the existing procedure, we propose an alternative strategy. In order to take account of users' heterogeneity, we partition HES data users into three groups: *Academic Researchers* (AR), *Public Administration* (PA), *Other Public Users* (OPU) and propose to replace the current masked microdata, HES_{1B}, with three different releases: a *Public Use File* ($\mathrm{HES}_{\mathrm{PU}}$) accessible to all users, a *Public Administration File* ($\mathrm{HES}_{\mathrm{PA}}$) accessible only to Public Administration offices, and a *Research file* ($\mathrm{HES}_{\mathrm{RE}}$), accessible only to researchers. In proposing such partition and defining the three data user groups we heuristically solved a decision problem with multiple objectives, where the action space is the set of all possible partitions of data users and the best action must be selected by taking into account the two conflicting objectives: "maximise users' homogeneity within groups", and "minimise the number of groups". The interpretation of the first objective should be clear from the discussion in the previous sections. The second objective reflects, instead, concerns about data dissemination cost. Ideally Istat could provide targeted data sets for each pair of data users and data user analyses. However, the larger the number of data sets to be released the larger the cost of data dissemination. Given Istat resources this partition was not further refined[5]. Choice of three groups reflects the existance of three different categories of intruders and types of access.

Within our framework the three files, $\mathrm{HES}_{\mathrm{PA}}$, $\mathrm{HES}_{\mathrm{PU}}$, and $\mathrm{HES}_{\mathrm{RE}}$, must be defined hierarchically in terms of information content. We must have $\mathrm{HES}_{\mathrm{PU}} \subset \mathrm{HES}_{\mathrm{PA}} \subset \mathrm{HES}_{\mathrm{RE}}$ whith $\mathrm{HES}_i \subset \mathrm{HES}_j$ meaning all the information in masked data set HES_i is also contained in HES_j. The hierarchical structure of the three data sets greatly simplifies assessment of the disclosure risk and information loss associated with the proposed strategy. Because of the hierarchy, in fact, there is

[4] Current Istat assessment of disclosure risk and data utility considers the two data products HES_{1A} and HES_{1B} independently assuming that users do not try to combine the information in both. This assumption seems quite realistic due to the difference in information content of the two data products and the complexity of combining the information in both. However, such an assessment underestimates both disclosure risk and data utility.

[5] Because of space constraints an operational definition of the three groups PA, PU and OPU is not reported here. For details on this definition the interested reader is referred to [10].

no gain for a user to access data in a lower level than the one he/she is granted. The agency can just assume that user h will use only data HES_h for his/her (legitimate or illegitimate) inferences. The next section describes a proposal for designing the Public Use File.

6 The Public Use File

The design of the Public Use file, HES_{PU}, was based on two general considerations. First of all, although the survey covers households as well as household members, we deemed appropriate for a public use file to release information at household level only. All personal information about individuals belonging to the sampled households was therefore removed from the file, except for the household reference person.

In addition, it was recognized that HES variables have a different role and carry different implications for both disclosure risk and data utility, given their heterogeneity in terms of *relevance* for intruder's goals, *usefulness* for legitimate statistical analyses, and *existence of alternative means* to disclose their values. As a result of this consideration variables in the data were partitioned into 5 homogeneous groups: *direct identifiers* (X_{DI}), *sensitive variables* (X_{S}), *non sensitive variables* (X_{NS}), *key variables* (X_{KEY}), and *sampling weights* (W).

Only those variables whose disclosure is both *relevant* to intruder's goals and *harmful* either to Istat or to survey respondents are defined as sensitive and collected in the X_{S} group. The extent to which a variable is "relevant" and "harmful" clearly depends on the intruder's motivations and data respondent perception of disclosure. To provide an operational classification we considered what is expected to be the prevailing perception of these two aspects. Like sensitive variables, non sensitive are not available to the intruder but their value is either not relevant for him/her or not harmful. Finally, key variables are those whose value is available from external sources of information, at least to some users.

An operational definition of key, sensitive and non sensitive variables is not an easy task because of data users' heterogeneity in terms of motivations for (legitimate or illegitimate) inferences, and number and type of variables available from external sources of information. As an illustration consider a target unit in the population. The set of key variables for a journalist that has picked up randomly the target from a list, a neighbour or a long-term friend of the target are clearly very diverse. In addition, depending on the motivation for disclosure, each of these potential intruders could, to a different extent, and with a different cost, increase the set of key variables. These additional variables could be considered as either key variables or sensitive/non sensitive variables disclosed by alternative means. In this paper we use the latter interpretation, including in the set of key variables, X_{KEY}, those variables that are available from the external registers mentioned in Section 3, or from brief social contact with the target (i.e. household composition). Data users that are so close to the target to be able to gather with very little effort detailed additional information on it are not considered potential intruders. In our view, given the existence of alternative means, motivation for disclosure is indeed very low in this group.

The distinction between sensitive and non sensitive variables, outlined above, reflects a precise interpretation of disclosure that combines the notion of *attribute* disclosure and *disclosure harm* (e.g. [3,8]) and that takes into account intruders' motivations. Disclosing expenditure in bread (non sensitive), clearly, is not the same as disclosing household income (sensitive). The difference relies on different potential harm of the disclosure, on potential intruders' motivations to disclosure and, as a result of this, on efforts and amount of resources the potential intruders would dedicate to such intent. Because of these differences, we deemed appropriate to define different disclosure scenarios and use different data masking procedures (and thus different risk/utility measures) for sensitive and non sensitive variables as it is described in the next subsections.

6.1 Disclosure Scenarios for the HES$_{\mathrm{PU}}$ File

Two disclosure scenarios for non sensitive and sensitive variables in X_{NS} and X_{S}, are considered.

Scenario1: intruders' attack to non sensitive variables. It is assumed that the intruder does not know whether the target household is in the sample and can only use the key variables for re-identification.

Scenario2: intruders' attack to sensitive variables. Since motivation for disclosure is high in this case we assumed a worst case scenario. To evaluate the risk of disclosure we make the assumption that potential intruders do know that a target household is in the sample. Moreover, they can use all the key variables and possibly some of the non sensitive variables for re-identification (as we commented before, a motivated intruder can obtain additional information from external sources on values of some non sensitive variables for the target unit).

Note that under the first scenario (for non sensitive variables) only meta-knowledge[6] about key variables is available to users. Indeed, under the Italian legislation, HES$_{\mathrm{PU}}$ data users can only access targeted records in the external data files mentioned in section 3. Likewise, personal/informal knowledge about households is only available at target level. Thus, to claim a reidentification the intruder can only resort, under Scenario 1, to *common sense epidemiology*, judging when a given combination of key variables is rare by common sense social knowledge.

6.2 Disclosure Limitation for the HES$_{\mathrm{PU}}$ File

In designing the HES$_{\mathrm{PU}}$ file, we decided to avoid those protection techniques that require users to adjust their inferences. Given the non sophisticated nature of users' analyses, global recoding was used whenever possible. Indeed global recoding yields microdata that can be analyzed by the users with the same statistical tools as the original data, no special care being required.

[6] As defined in [4], "meta-knowledge about keys is not information about actual values of key variables, but knowledge about the variables themself. An obvious example is social knowledge about baselines".

Direct identifiers i.e. variables in X_{DI} were clearly suppressed, whereas the sampling weights W were released unchanged, as they are required for proper inferences. Non sensitive variables, X_{NS}, were not perturbed. Apart from a subgroup of variables that were judged not relevant for the inferences of legitimate users of the $\mathrm{HES}_{\mathrm{PU}}$ data and were therefore suppressed, most non sensitives are released unchanged. The rationale for releasing the first set of variables unperturbed is that for these variables either intruder's motivation for disclosure is low, and/or they could be obtained with less effort by other means than the released data. Based on these considerations we deemed that, under Disclosure Scenario 1, a suitable masking of X_{KEY} variables aimed to avoid rare combinations was sufficient to protect confidential information in the non sensitive variables. Key variables masking was done as follows. Key variables containing personal information about household members were suppressed. The other key variables were masked using global recoding. In recoding we took into account the trade-off between analysis of legitimate data users and potential threats to confidentiality. An additional constraint was the existence of the set of tables, HES_{1A} that Istat traditionally releases as one of its standard products (see section 4). The public use file must agree with the released tables and, as discussed in section 5, the latter cannot provide the intruder an additional source of information for re-identification. Key variables were therefore recoded in a way that allowed users to recover, directly or by aggregation, the published tables. As a result of the recoding the initial set of key variables, X_{KEY}, is replaced by a new one, $X_{\mathrm{KEY}'}$ that comprises four keys: *Geography*, at regional level (20 regions), *Occupation* of household reference person (6 classes), *Household Type* (11 categories, six containing information on household reference person age in three broad classes), and *Household Size* (topcoded at 5 components).

The masking procedure for key variables that we just described, is not sufficient to guarantee confidentiality protection for the sensitive variable. Under Disclosure Scenario 2, in fact, the intruder does know that a target is in the sample and release of key variables $X_{\mathrm{KEY}'}$ would allow re-identification (and thus exact disclosure of sensitive variables) for several respondents (which are sample unique in terms of key variables $X_{\mathrm{KEY}'}$). As a result of these considerations the sensitive variables in X_{S} were either masked with an *ad-hoc* procedure that we call *Microaggregated Swapping*, or completely removed from the file. In this respect, sometimes the precision of the corresponding estimates motivated exclusion from the $\mathrm{HES}_{\mathrm{PU}}$ file.

Due to the presence of sampling weights, special care is required in masking sensitive variables. The technique we implemented for continuos variables, which is discussed in detail in the Appendix, mimics swapping but also implements a microaggregation that accounts for the presence of unmasked sampling design weights. Our target was to let users build the same estimates as with the original data over the subpopulations defined by the most important socio-demographic survey variables in X_{KEY}. Confidentiality considerations led us to define these subpopulations according to a broader classification than the one used to produce $X_{\mathrm{KEY}'}$, namely: *Geography*, at an aggregate level (5 macro regions), *Household*

Type (6 categories) and *Household Size* (topcoded at 5 components). We refer to these variables as $X_{\mathrm{KEY}''}$.

As far as disclosure risk is concerned, the above protection technique reduces the set of variables that can be used for re-identification to the set of masked key variables $X_{\mathrm{KEY}''}$ and makes any intruder's attempt to use other key or non sensitive variables useless. Masking was performed independently for each sensitive variable, thus producing a new set of variables that we denote by $\tilde{X}_{\mathrm{S}}$. As anticipated, it was designed so that the marginal distribution of sensitive variables is unaffected and so is their conditional distribution over subgroups defined by cross-tabulating the variables in $X_{\mathrm{KEY}''}$. The minimum detail at which the masked data are to be analyzed corresponds by construction to the cells defined by cross-tabulation of $X_{\mathrm{KEY}''}$ variables.

6.3 Disclosure Risk for the HES$_{\mathrm{PU}}$ File

As described in section 6, to assess disclosure risk associated with the Public Use File we used the interpretation of disclosure as *attribute disclosure* and *disclosure harm*. The different scenarios adopted for non sensitive and sensitive variables require two different approaches to disclosure risk assessment. Denote by m a generic data masking. Let $X_{\mathrm{A}}^{(s,m)}$, be the set of variables available for household re-identification under Disclosure Scenario s, $s = 1, 2$ and data masking m, and let $T_{\mathrm{A}}^{(s,m)}$ be the contingency table built by cross-classifying the variables $X_{\mathrm{A}}^{(s,m)}$. Consider a targeted household I^* in the sample. Observing the value of $X_{\mathrm{A}}^{(s,m)}$ variables in such an household will classify the household into a cell of $T_{\mathrm{A}}^{(s,m)}$. Denote by $k_{I^*}^{(s)}$ the index of the cell into which I^* is classified based on the values of $X_{\mathrm{A}}^{(s,m)}$. Under Disclosure Scenario 2 (for sensitive variables) the relevant distribution for attribute disclosure or disclosure harm evaluation is the estimated sample distribution of the sensitive variable in cell $k_{I^*}^{(2)}$. Under Disclosure Scenario 1 (for non sensitive variables), instead, it is the estimated population distribution of the non sensitive variable in cell $k_{I^*}^{(1)}$. For the masking $\bar{m}$, described in the previous section, we have

$$X_{\mathrm{A}}^{(1,\bar{m})} = X_{\mathrm{KEY}'}, \quad X_{\mathrm{A}}^{(2,\bar{m})} = X_{\mathrm{KEY}''},$$

while $T_{\mathrm{A}}^{(1,\bar{m})}$ ($T_{\mathrm{A}}^{(2,\bar{m})}$) is the contingency table formed by cross-classifying the key variables in $X_{\mathrm{KEY}'}$ ($X_{\mathrm{KEY}''}$). We denote the total number of cells in $T_{\mathrm{A}}^{(s,\bar{m})}$ by K_s, $s = 1, 2$.

For a generic sensitive variable, S_j, disclosure risk assessment proceeds as follows. Let $\hat{G}_{S_j}^{(k)}$ be the estimated c.d.f. of S_j for the k-th cell of $T_{\mathrm{A}}^{(2,\bar{m})}$ that is the empirical distribution of the un-weighted masked S_j data. Let $Spread(\hat{G}_{S_j}^{(k)})$ and $\int_{c_{1k}}^{\infty} d\hat{G}_{S_j}^{(k)}$ ($\int_{-\infty}^{c_{2k}} d\hat{G}_{S_j}^{(k)}$) be, respectively, measures of spread[7] and tail probability for $\hat{G}_{S_j}^{(k)}$, $k = 1, 2, \ldots, K_2$. We measure disclosure risk for S_j at the file level as

[7] The measures of spread actually used are: variance, range, and mean variation.

$$DR_{S_j} = min\{Spread(\hat{G}^{(k)}_{S_j}),\ k = 1, 2, \ldots, K_2\} \tag{1}$$

or

$$\begin{cases} DR_{S_j} = max\{\int_{c_{1k}}^{\infty} d\hat{G}^{(k)}_{S_j},\ k = 1, 2, \ldots, K_2\}, \\ DR_{S_j} = max\{\int_{-\infty}^{c_{2k}} d\hat{G}^{(k)}_{S_j},\ k = 1, 2, \ldots, K_2\}. \end{cases} \tag{2}$$

Choice between (1) and (2) depends on whether attribute disclosure or disclosure harm of the type $P(S_j^{(i)} > c_{1k})$ $(P(S_j^{(i)} < c_{2k}))$ is of concern. Table 1, shows, in bold, the values of (1) and (2) evaluated for the sensitive variable *Imputed Rent* (IR). For (2) only the first expression is considered. For comparison the table also reports (in parentheses) the value of the measures of disclosure risk that we would obtain by using in (1) and (2) the true sample distribution i.e. the empirical distribution of the un-weighted original data. Note as the masking provides a poor estimate of the tails of the distribution for the un-weighted data thus providing additional protection for attribute disclosure.

Table 1. Disclosure risk for imputed rent

Variance	Range	Mean Variation	$max_k \int_{c_{1k}}^{\infty} d\hat{G}^{(k)}_{S_j}$
18313 (19073)	**601** (601)	**58** (58)	**0.13 (0.06)**

Choice of the constants c_{1k} and c_{2k} in (2) is not trivial and should depend on the differences in the distribution of the sensitive variable across cells. For the sensitive variable IR in our example we used two different values of c_{1k} depending on *Geography*.

For the non sensitive variables, given scarcity of data available to estimate their population distribution in each cell of $T_{\mathrm{A}}^{(1,\bar{m})}$ we considered the probability of re-identification of targeted households based on $X_{\mathrm{KEY}'}$ variables as a proxy for attribute disclosure and disclosure harm. Let k be the $k-th$ cell in $T_{\mathrm{A}}^{(1,\bar{m})}$, and let f_k^y, F_k^y be the sample and the population frequency of cell k for year y, $k = 1, 2, \ldots, K_1$. Following the re-identification scenario in [5], we define the risk of re-identification at the household level for the $\mathrm{HES}_{\mathrm{PU}}$ 2004 file as:

$$DR = max\{1/F_k^{2004},\ k = 1, 2, \ldots, K_1\}.$$

For HES sample frequencies f_k^y are very stable over the period 2002-2004 we estimated the population frequency F_k^{2004} for the year 2004 using the available Census Data for the year 2001, i.e.

$$DR = max\{1/\hat{F}_k^{2001},\ k = 1, 2, \ldots, K_1\} = 0.053$$

where $\hat{F}_k^{2001}$ are the estimated population frequencies obtained from the 2001 Italian Population Census.

6.4 Data Utility

The disclosure risk assessment described in the previous section constraints the detail at which the masked non sensitive and sensitive variables are to be analyzed. As a result of these constraints, the descriptive study on the HES data users' needs outlined in section 3, and the additional constraints imposed by the hierarchical structure of the proposed data dissemination procedure, we identified three data-utility goals in the dissemination of the HES_{PU} data: GOAL 1) preserve the joint distribution of the non sensitive variables and the key variables $X_{\text{KEY}'}$; GOAL 2) Preserve the mean of each of the sensitive variables within each cell of the contingency table $T_{\text{A}}^{(2,\bar{m})}$; GOAL 3) preserve the marginal distribution of each of the sensitive variables within each cell of the contingency table $T_{\text{A}}^{(2,\bar{m})}$ (or, otherwise stated, preserve the joint distribution of each sensitive variable and the key variables $X_{\text{KEY}''}$). We used measures that quantify achievement of these three objectives as measures of data utility. Since the proposed HES_{PU} file achieves GOALS 1 and 2 by construction, our measures of data utility reduces to measure achievement of the third goal. Considering the typical queries of HES_{PU} data users, we defined data utility of the HES_{PU} file as a measure of discrepancy between decile estimates of sensitive variables under the original data and the corresponding estimates under the MS-masked data. For each sensitive variable S_j decile estimates of the distribution of S_j across cells of the contingency table $T_{\text{A}}^{(2,\bar{m})}$ were obtained using both the empirical distribution function (for the original and MS-masked weighted data), and kernel density estimators. For each cell discrepancies between decile estimates were quantified by summary statistics of their absolute and relative differences across cells. The MS-masking seems to perform quite well, discrepancies in decile estimates being usually very small. As expected the largest discrepancies are observed when deciles are estimated from the empirical distribution function and the number of observations in the cell is small (see, for example, Table 2 in the Appendix which illustrates evaluation of Data Utility for the sensitive variable *imputed rent*). To explore the effect of sampling weights in data users' inferences, and to appreciate the advantage of the proposed MS-masking with respect to other procedures that do not preserve sampling weights, weighted and unweighted inferences were also qualitatively compared. Fig. 1 in the appendix shows the type of plots that we used for such an assessment. Regardless of the sensitive variable considered, we observed that for several cells in $T_{\text{A}}^{(2,\bar{m})}$ inferences that ignore weights were misleading thus confirming that preserving sampling weights is an essential property for masking procedures designed to allow correct data users' inferences.

7 Conclusions and Future Work

Recent research in data disclosure limitation has recognized that users' heterogeneity is a crucial component in designing efficient data dissemination strategies. It is argued that, because of this heterogeneity, data dissemination procedures that use a portfolio of protection methods (and thus that comprise different data products targeted to different data users and data users' needs) would perform better than those relying on a single protection method.

In this paper we reported on an ongoing work to define a proposal for an alternative data dissemination strategy of the Household Expenditure Survey carried out by the Italian National Statistical Institute that tries to implement these ideas. Building on a descriptive study of HES data users and their needs we propose a differentiated data dissemination strategy that comprises three different protection methods and thus three different data products depending on the type of user. It is argued that the standard partition of the variables in *key* and *sensitive*, is data user dependent and, as a result, the masking procedure and the assessment of the performance of the masking in terms of risk and utility should be user dependent, too. We illustrated the rationales for partition, data masking and risk/utility assessment only for the Public Use File targeted to all users but researchers and users in the Public Administration. For such file, in order to preserve sampling weights, we designed a new masking procedure that we called *Microaggregated Swapping*. We are currently working on the implementation of the other two files that complete the proposed data dissemination procedure and on the definition of a global measure of disclosure risk and data utility that can be applied to the overall procedure. This requires to take into account multiple components of risk and utility that include those already described for the Public Use File and those to be defined for the Public Administration and Research file. Ranking procedures in multiple objective decision theory that do not require formalized preference structures (see [7], chapter 3) can be useful in this respect.

The preliminary results presented and the complete implementation of the proposed HES data dissemination will hopefully contribute in establishing a methodology for differentiated data dissemination procedures useful to take into account the diverse data users and data users' needs. Despite the convincing arguments of the portfolio approach, in fact, data dissemination procedures targeted to the different data users are still very rare in practice (an example of such implementations can be found in [2]). This is due to the cost associated with a differentiated data dissemination and the complexity of its implementation. Real data examples are needed to establish a solid methodology and make portfolio strategy the norm rather than an exception in the practice of Data Disclosure Limitation.

Acknowledgment. The authors would like to thank Maria Polsoni for providing application forms for HES masked data set received by Istat and Maurizio Lucarelli for process of census data. The first author was supported by the U.S. National Science Foundation under Grant EIA-0131884 to the National Institute of Statistical Sciences. The contents of the paper reflects the authors' personal opinion. Neither the National Science Foundation nor Istat are responsible for any views or results presented.

References

1. J. M. Abowd and J. Lane. The economics of data confidentiality. In *Unpublished paper presented at the National Research Council's Committee on National Statistics Workshop on Confidentiality and Access to Research Data Files*, Washington DC, October 2003.
2. J. M. Abowd and J. Lane. Synthetic data and confidentiality protection. In *Workshop on Microdata*, Stockholm, Sweden, August 2003.
3. G. T. Duncan and D. Lambert. Disclosure-limited data dissemination (with comments). *Journal of the American Statistical Association*, 81:10–27, 1986.
4. M. J. Elliot and A. Dale. Scenarios of attack: The data intruder's perspective on statistical disclosure risk. *Netherlands Official Statistics*, 14:6–10, 1999.
5. L. Franconi and S. Polettini. Individual risk estimation in μ-Argus: A review. In J. Domingo-Ferrer and V. Torra, editors, *Privacy in Statistical Databases 2004*, volume 2316 of *Lecture Notes in Computer Science*, pages 262–272. Springer, Berlin, 2004.
6. Istat. *I Consumi delle Famiglie.* Annuari, n. 11, 2006. Istituto Nazionale di Statistica, Roma, 2006.
7. R.L. Keeney and H. Raiffa. *Decisions with Multiple Objectives.* New York:Wiley, 1976.
8. D. Lambert. Measures of disclosure risk and harm. *Journal of Official Statistics*, 9:313–331, 1993.
9. C. Mackie and N. Bradburn. *Improving Access to and Confidentiality of Research Data: Report of a Workshop.* Committee on National Statistics (CNSTAT). Commission on Behavioral and Social Sciences and Education, Washington DC: National Academies Press, 2000.
10. M. Trottini, L. Franconi, and S. Polettini. *Italian Household Expenditure Survey: a Proposal for Data Dissemination.* Collana Contributi Istat. 2006. In preparation.

Appendix: Microaggregated Swapping (MS)

Consider a sensitive variable Y. Let $l_1^k, l_2^k, \ldots, l_{n_k}^k$ be the set of indices corresponding to those records/households in the original data whose key variables combination belongs to the k-th cell of the contingency table $T^{(2,\bar{m})}$ formed by cross-classifying the key variables in $X_{\mathrm{KEY}''}$. Let $y_{l_d^k}$ be the value of Y for the l_d^k-th record in the original data and $w_{l_d^k}$ be the corresponding sampling weight, $d = 1, 2, \ldots, n_k$. Define the matrix

$$A^{(k)} = \begin{pmatrix} y_{l_1^k} & w_{l_1^k} \\ y_{l_2^k} & w_{l_2^k} \\ \cdots & \cdots \\ y_{l_{n_k}^k} & w_{l_{n_k}^k} \end{pmatrix}.$$

The $k - th$ step of the MS procedure for the masking of Y is as follows:

1. Sort the records of $A^{(k)}$ according to the value of Y. Denote by s the permutation that produces the sort and by $B^{(k)}$ the sorted matrix

$$B^{(k)} = [\mathbf{y}^{(k)} \vdots \mathbf{w}^{(k)}]$$

where $\mathbf{y}^{(k)}$ is the vector of Y values obtained sorting in ascending order the first column of $A^{(k)}$, and $\mathbf{w}^k$ is the vector of the corresponding weights,

$$\mathbf{y}^{(k)} = s(y_{l_1^k}, y_{l_2^k}, \ldots, y_{l_{n_k}^k}), \;\; \mathbf{w}^{(k)} = s(w_{l_1^k}, w_{l_2^k}, \ldots, w_{l_{n_k}}^k).$$

2. Generate a random permutation, p, of the integers $1, 2, \ldots, n_k$ from the uniform distribution (on the set of all possible permutations of the first n_k integers).
3. Apply p to the vector of weights $\mathbf{w}^{(k)}$. This will yield a new vector of weights $\tilde{\mathbf{w}}^{(k)}$,
$$\tilde{\mathbf{w}}^{(k)} = p(\mathbf{w}^{(k)}) = (\tilde{w}_1, \tilde{w}_2, \ldots, \tilde{w}_{n_k}).$$
4. Apply univariate microaggregation to Y in the microdata corresponding to the "frequency" distribution $B^{(k)}$ using n_k as the number of groups and $\tilde{\mathbf{w}}^{(k)}$ to define the groups size. This will yield a new microdata corresponding to a new "frequency" distribution $\tilde{B}^{(k)}$,
$$\tilde{B}^{(k)} = [\tilde{\mathbf{y}}^{(k)} \vdots \tilde{\mathbf{w}}^{(k)}]$$
where $\tilde{\mathbf{y}}^{(k)} = (\tilde{y}_1, \tilde{y}_2, \ldots, \tilde{y}_{n_k})$ and $\tilde{y}_d$ is the average of Y in the d-th group, $d = 1, 2, \ldots, n_k$.
5. Let p^{-1} be the inverse of the permutation p in step 4. Apply p^{-1} to the vector $\tilde{\mathbf{y}}^{(k)}$. This will yield a new vector of Y values, $\tilde{\tilde{\mathbf{y}}}^{(k)}$.
6. Let s^{-1} be the inverse of the permutation that defines the sort s in step 3. Apply s^{-1} to $\tilde{\tilde{\mathbf{y}}}^{(k)}$. This will yield a new vector of Y values, $\tilde{\tilde{\tilde{\mathbf{y}}}}^{(k)}$
$$\tilde{\tilde{\tilde{\mathbf{y}}}}^{(k)} = s^{-1}(p^{-1}(\tilde{\mathbf{y}}^{(k)})) = (\tilde{\tilde{\tilde{y}}}_1, \tilde{\tilde{\tilde{y}}}_2, \ldots, \tilde{\tilde{\tilde{y}}}_{n_k}).$$
7. In the original data replace $y_{i_d}^k$ by $\tilde{\tilde{\tilde{y}}}_d$, $d = 1, 2, \ldots, n_k$.

The MS procedure for the masking of Y is then as follows:

```
a) Set k=1
b) Perform Step k
c) If k<K2 set k=k+1 and return to b)
      else  END.
```

Example

The *Swapping* and *Microaggregation* steps of the MS procedure might be no apparent from the algorithm describing the procedure. An illustration of the MS masking procedure might help to clarify. Suppose that: $n_k = 3$, the matrix A^k is given by:

$$A^{(k)} = \begin{pmatrix} 32 & 2 \\ 40 & 4 \\ 25 & 3 \end{pmatrix},$$

and the random permutation generated at step 4 of the MS procedure is $p = (2, 3, 1)$. Denote by A' the "microdata" corresponding to $A^{(k)}$,

$$A' = (32, 32, 40, 40, 40, 40, 25, 25, 25).$$

The MS procedure is equivalent to replace $A^{(k)}$ with a new matrix, $\ddot{A}^{(k)}$, such that the values in the second column of $A^{(k)}$ (the weights) are preserved and the values in the first column (i.e. de values of Y) are modified as follows:

1. Create a new "microdata" shuffling the observations in A'. For this example the shuffling induced by the permutation p is:

$$A'' = (40, 40, 25, 25, 25, 32, 32, 40, 40).$$

2. Apply microaggregation to A'' with number of groups equal to 3 and groups size given by the vector of sampling weights in $A^{(k)}$, i.e. $(2, 4, 3)$ and replace the $i-th$ value in the first column of $A^{(k)}$ with the average of the i-th group.

The new matrix $\ddot{A}^k$ is given by: $\ddot{A}^{(k)} = \begin{pmatrix} 40.00\ 2 \\ 26.75\ 4 \\ 37.33\ 3 \end{pmatrix}$.

Assessment of the MS Method

Table 2. Data Utility for imputed rent (IR): summary statistics for relative differences in estimates of deciles obtained using the empirical distribution function for the original and MS-masked data by number of observations (in the cells of $T_{\mathrm{A}}^{(2,\bar{m})}$). Corresponding summaries for absolute differences (in euros) are also reported in parentheses.

Number of Observations in the Cell	min	1^{st} Quartile	Median	Mean	3^{rd} Quartile	max
0-99	0.017 (5.6)	0.039 (29.1)	0.059 (33.2)	0.078 (39.7)	0.099 (38.6)	0.216 (197.2)
100-249	0.000 (0.0)	0.017 (5.2)	0.050 (19.5)	0.044 (23.6)	0.055 (36.7)	0.148 (79.2)
250-349	0.000 (0.0)	0.000 (1.0)	0.004 (5.0)	0.027 (7.0)	0.039 (11.9)	0.091 (18.2)
≥ 350	0.000 (0.0)	0.000 (0.0)	0.000 (0.0)	0.008 (2.2)	0.010 (1.0)	0.051 (19.0)

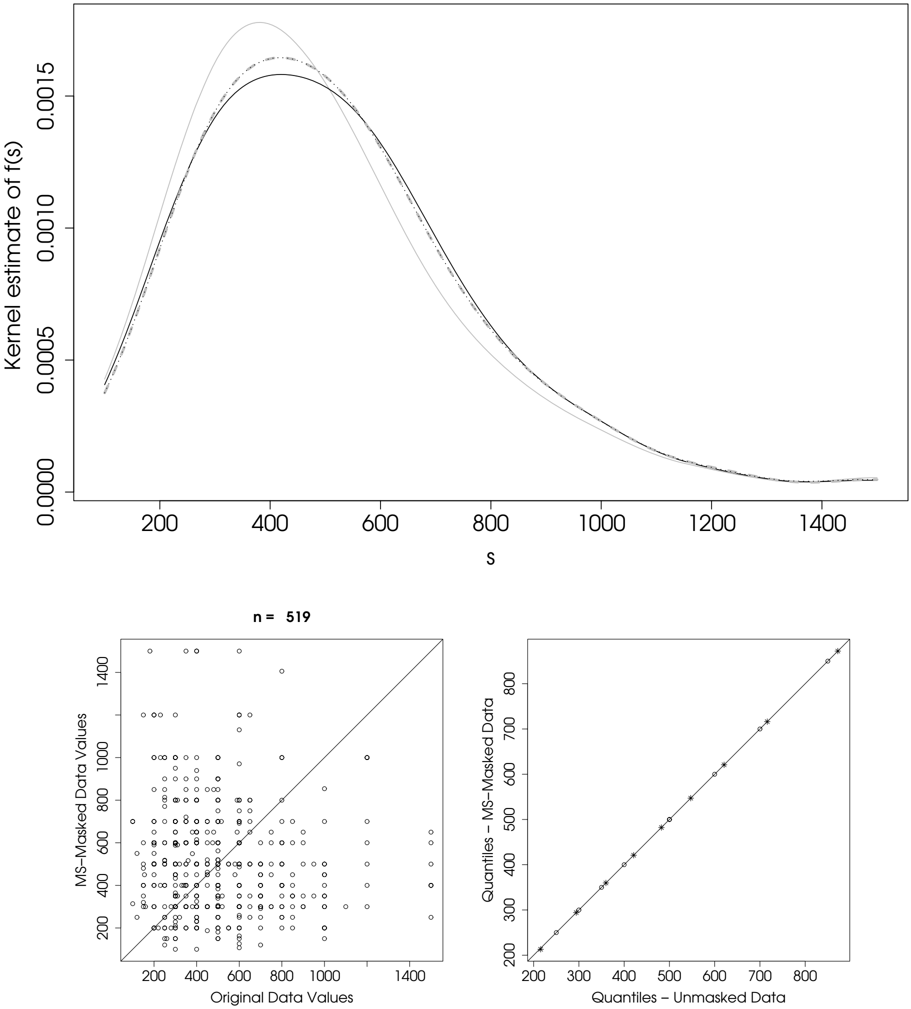


Fig. 1. Comparison of inferences for imputed rent (IR) using original and MS-Masked data in a cell of the contingency table $T_{\mathrm{A}}^{(2,\bar{m})}$ (all values are in euros). **Top**: Comparison between Kernel density estimation using the original data (in gray) and the MS-Masked data (in black). Weighted estimates are represented by dotted lines, un-weighted estimates by solid lines. **Bottom left**: Original data of IR versus MS-Masked data. **Bottom right**: Estimates of deciles obtained using the empirical distribution function for the original data versus corresponding estimates using the MS-masked data. Weighted estimates are denoted by stars, un-weighted estimates by circles.

The ARGUS Software in CENEX

Anco Hundepool

Department of Methods and Informatics
Statistics Netherlands,
P.O. Box 4000
2270 M Voorburg, Netherlands
ahnl@rnd.vb.cbs.nl

Abstract. In this paper we will give an overview of the CENEX project and concentrate on the current state of affairs with respect to the ARGUS-software twins. The CENEX (Centre of Excellence) is a new initiative by Eurostat. The main idea behind the CENEX-concept is to join the forces of the national NSI's and together bring the skills of the NSI's on a higher level. The CENEX on Statistical Disclosure Control is a first pilot CENEX-project both aiming at testing the feasibility of the CENEX idea and working on SDC. This project will make a start of writing a handbook on SDC, after an inventory and extend the ARGUS software with an emphasis on issues of practical use. Within this CENEX we will organise the transfer of technology via courses, a WEB-site and this conference. Finally a roadmap for future work will hopefully lead to a follow-up CENEX.

In this paper we will summarise this CENEX-project and give a short overview of the current versions of ARGUS.

Keywords: European cooperation. Statistical Disclosure Control, ARGUS.

1 The CENEX-SDC Project

In the recent years there has been already a lot of fruitful cooperation in Europe in the field of Statistical Disclosure Control. As a result of the 5th Framework project CASC (Computational Aspects of Statistical Confidentiality) much progress has been made, both in the field of research as well as the development of practical tools, i.e. the ARGUS software. These achievements also showed that international cooperation can work very well.

Recently Eurostat has launched new initiatives to promote the cooperation between the NSI's. This has led to the development of the concept of Centres and Networks of Excellence (CENEX). The main idea is of course that not all NSI's in Europe have the same level of skills on all subjects and that NSI's more advanced on a certain topic should join the forces to both further develop these skills but also work on the transfer of these skills to the NSI's that are less developed in this topic. This should lead to improvements in the European Statistical System.

In order to test the CENEX-concept two pilot projects were initiated. SDC was selected as one of the topics. As only small groups per country are working in this field joining forces is a very efficient idea. The SDC-groups have showed already in

J. Domingo-Ferrer and L. Franconi (Eds.): PSD 2006, LNCS 4302, pp. 334 – 346, 2006.

the past that European cooperation can work very well. This made the field of SDC an obvious choice for this CENEX pilot. Although only a smaller part of the CENEX has been completed when this paper was written, the first conclusions are very positive and we are looking forward to continue to cooperate in a CENEX concept in the future.

As the CENEX concept is mainly aiming at cooperation between NSI's the contribution from the universities is restricted. But as universities have been playing an important role in this field we should continue to initiate research-projects e.g. in the EU framework projects FP7 parallel to the CENEX-initiatives

The work in the CENEX will be summarised in the next sections.

1.1 The Inventory

As a starting point of this CENEX project an inventory of existing methods and techniques in Europe is being carried out. A rather large ambitious questionnaire has been designed and sent to all NSI's of the EU-member states. This questionnaire include questions on legal issues, the SDC-measures with respect to Public Use Files and Micro data files for researchers, the SDC-aspects of magnitude tables and frequency tables. At this moment the answers are being analysed.

1.2 The Handbook

One of the outcomes of the CENEX project is also a first version of a handbook on SDC. It is not to be expected that a fully completed handbook will be written within one year, but we will make a first version. That version will be further discussed on several platforms. We hope that the outcome of this process will be a useful standard on SDC.

After an introduction we will summarise the regulations and then pay attention to SDC-methods for microdata. Risk assessment will be the starting point and then we will describe all the methods available. The pro's and con's of these methods will be highlighted and we will introduce the tools available, i.e. μ-ARGUS. The same format will be used for tabular data.

After the completion of this version of the handbook we foresee discussions on this handbook at various platforms. These discussions could lead to improvements in the near future.

1.3 ARGUS Software

The ARGUS software is a tool in constant development. As the aim of the ARGUS development is to make ARGUS a control centre for all the methods available in SDC, this requires a continuous development process. But in the CENEX project we will mainly pay attention to the practical applicability in the software in the daily practice of the NSI's. The extensions during this CENEX will therefore mainly focus in these aspects, leaving the implementation of new methodology to other future projects.

For this several NSI's are testing ARGUS and will provide us will their comments. When possible and feasible we will include these remarks in a new version of ARGUS during this CENEX. The more complex wishes will be input for the future work list (see 1.5).

1.4 Communication

Communication is an essential aspect of this CENEX. Therefore we have taught a successful course on SDC. Thanks to the hospitality of the Hungarian Statistical Office this course was held in Budapest (12-14 June 2006). 25 representatives of almost all member states were present. The content of the course was a mix of theoretical subjects, manual exercises and also training sessions in the use of the ARGUS software packages. According to the reactions of the participants this was a very useful happening.

A second activity in this topic is organisation of the conference PSD'2006 in Rome. This conference is more directed at the methodological aspects of SDC and will keep the contacts alive between the universities and the NSI's. These contacts and cooperation have been very effective in the past and have led to many methodological improvements in the practice of SDC.

Besides this the CENEX team will maintain a website, where the results of this project will be made publicly available. This website can be found at http://neon.vb.cbs.nl/cenex

1.5 Future Work

The work on SDC will not be finished after the CENEX. We look forward to continue our cooperation and hopefully have the opportunity to continue the CENEX work. Therefore this CENEX will also draw a roadmap for the future work of cooperation. A two-day internal conference is planned for drawing this roadmap.

This roadmap will pave the way for the new cooperation in this field.

1.6 Management

The key participants of this CENEX are the major participants in the CASC Project. Besides Anco Hundepool (Statistics Netherlands) as project leader Luisa Franconi (IStat), Sarah Giessing (DeStatis), Jane Longhurst(ONS) and Josep Domingo-Ferrer (University Rovira i Virgili) participate as the major partners in this project. Besides these partners the statistical offices of Sweden, Slovenia, Estonia and Austria participate as testers of the ARGUS software and will contribute to the roadmap for future work.

2 The ARGUS Software

2.1 Introduction

The CASC-project has made a major step forward in the development of software for Statistical Disclosure Control. The resulting packages μ-ARGUS and τ-ARGUS have made the results of several research efforts of the CASC project readily available or the users. In the CENEX project we will concentrate not on the implementing new methodology in ARGUS, but mainly try to make the software more easily applicable in the daily routine of the users in the different NSI's.

2.2 μ-ARGUS

2.2.1 Introduction

μ-ARGUS is based on a view of safety/unsafety of microdata that is used at Statistics Netherlands. In fact the incentive to build a package like μ-ARGUS was to facilitate data protectors at Statistics Netherlands to apply the general rules for various types of microdata easily and to relieve them from the tedious taks that producing a safe file in practice can involve. Not only should it be easy to produce safe microdata, it should also be possible to generate a logfile that documents the modifications of a microdata file.

The aim of statistical disclosure control is to limit the risk that sensitive information of individual respondents can be disclosed from data that are released to third party users. In case of a microdata set, i.e. a set of records containing information on individual respondents, such a disclosure of sensitive information of an individual respondent can occur after this respondent has been re-identified. That is, after it has been deduced which record corresponds to this particular individual. So, the aim of disclosure control should help to hamper re-identification of individual respondents represented in data to be published.

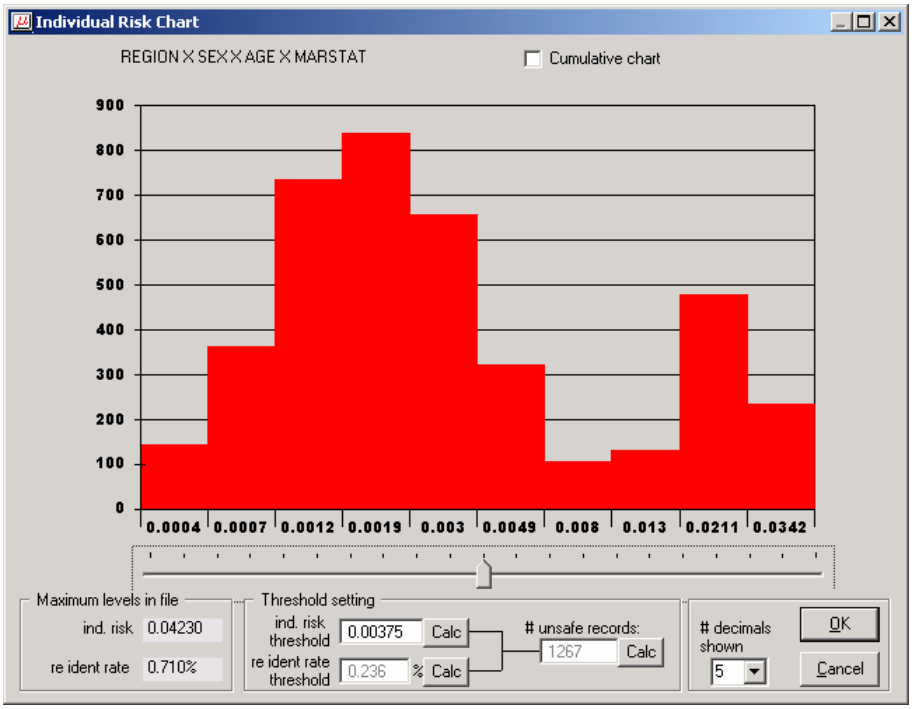
2.2.1.1 A Simple Threshold Rule. An important concept in the theory of re-identification is a key. A key is a combination of (potentially) identifying variables. An identifying variable, or an identifier, is one that may help an intruder re-identify an individual. Typically an identifying variable is one that describes a characteristic of a person that is observable, that is registered (identification numbers, etc.), or generally, that can be known to other persons. This, of course, is not very precise, and relies on one's personal judgement. But once a variable has been declared identifying, it is usually a fairly mechanical procedure to deal with it in μ-ARGUS.

Re-identification of an individual can take place when several values of so-called identifying variables, such as ‘Place of residence’, ‘Sex’ and ‘Occupation’, are taken into consideration. The values of these identifying variables can be assumed to be known to relatives, friends, acquaintances and colleagues of a respondent. When several values of these identifying variables are combined a respondent may be re-identified.

2.2.2.2 The Franconi-Benedetti Risk Model. To be able to distinguish safe from unsafe microdata, it is necessary that a disclosure risk model is specified. Disclosure models can differ greatly in their degrees of sophistication. The basic model in μ-ARGUS is a fairly simple such model, namely one based on a thresholding rule. The understanding is that a combination of values is safe only if the (estimated) frequency of its occurrence in the population (or in the file) is above a certain threshold value.

An individual risk of disclosure allows one to estimate a measure of the chance of identification of each record in the released file on the basis of the actual values observed on the public variables. In the last few years a number of proposals have been made. Benedetti and Franconi (1998) propose a methodology for individual risk estimation based on the sampling weight, which is the approach used in this version of μ-ARGUS.

In extension to the individual risk model implemented already also a risk model based on groups of records (holdings or households) has been introduced.



□ Cumulative chart

Maximum levels in file
ind. risk 0.04230
re ident rate 0.710%

Threshold setting
ind. risk threshold 0.00375
re ident rate threshold 0.236 %
# unsafe records: 1267

decimals shown 5

Fig. 1. A graph showing the risk distribution in μ-ARGUS

2.2.2 Statistical Disclosure Control Measures

To avoid re-identification several techniques are available in μ-ARGUS, like global recoding (grouping of categories), local suppression, PostRAndomisation Method (PRAM), adding noise and microaggregation.

2.1.2.1 Global Recoding. In case of global recoding several categories of a variable are collapsed into a single one. The effect will be that the number of records with the same key will rise. And the risk of re-identification will diminish. On the one hand side it is a very powerful instrument in μ-ARGUS to make a safe datafile. Many unsafe keys/records with a high disclosure risk will disappear, but on the other hand a lot of detail can disappear as well. The data protector should use this method carefully and also keep in mind that if he cannot protect the unsafe records here, he will have to apply many local suppressions (i.e. impute missing values). Future users of a dataset might not like this perspective and prefer a more aggregated categorisation for a variable without all these missing values.

It is important to realise that global recoding is applied to the whole data set, not only to the unsafe part of the set. This is done to obtain a uniform categorisation of each variable. Finding an optimal balance between global recoding and local suppression is not an easy task, but requires statistical insight in the needs of the users of the protected data sets.

2.1.2.2 Local Suppression. When local suppression is applied one or more values in an unsafe combination are suppressed, i.e. replaced by a missing value. This removes the possibility to use this key any longer for re-identification. As keys often consist of several variables there is a freedom to select one of them for local suppression. Also several unsafe keys can be found in one record. μ-ARGUS offers two methods to do this efficiently. One is based on the minimising of the reduction of the entropy (i.e. preserving as much as possible the information), but as an alternative the user can specify his own priorities, based on his views on the importance of preserving as much as possible certain variables..

2.1.2.3 Top and Bottom Coding. Global recoding is a technique that can be applied to general categorical variables, i.e. without any requirement of the type. In case of ordinal categorical variables one can apply a particular global recoding technique namely top coding (for the larger values) or bottom coding (for the smaller values). When, for instance, top coding is applied to an ordinal variable, the top categories are lumped together to form a new category. Bottom coding is similar, except that it applies to the smallest values instead of the largest. Top and bottom coding for categorical variables can be seen as special case of global recoding.

Top and bottom coding can also be applied to continuous variables. What is important is that the values of such a variable can be linearly ordered. It is possible to calculate threshold values and lump all values larger than this value together (in case of top coding) or all smaller values (in case of bottom coding). Checking whether the top (or bottom) category is large enough is also feasible.

2.1.2.4 The Post RAndomisation Method (PRAM). PRAM is a disclosure control technique that can be applied to categorical data. Basically, it is a form of deliberate misclassification, using a known probability mechanism. Applying PRAM means that for each record in a microdata file, the score on one or more categorical variables is changed. This is done, independently of the other records, using a predetermined probability mechanism. Hence the original file is perturbed, so it will be difficult for an intruder to identify records (with certainty) as corresponding to certain individuals in the population. Since the probability mechanism that is used when applying PRAM is known, characteristics of the (latent) true data can still be estimated from the perturbed data file. See De Wolf (2006).

2.1.2.5 Microaggregation. Microaggregation is a family of statistical disclosure control techniques for quantitative (numeric) microdata, which belong to the substitution/perturbation category. The rationale behind microaggregation is that confidentiality rules in use allow publication of microdata sets if records correspond to groups of k or more individuals, where no individual dominates (i.e. contributes too much to) the group and k is a threshold value. Strict application of such confidentiality rules leads to replacing individual values with values computed on small aggregates (microaggregates) prior to publication. This is the basic principle of microaggregation.

The method for multivariate fixed-size microaggregation implemented in μ-ARGUS tries to form homogeneous groups of records by taking into account the distances between records themselves and between records and the average of all records in the data set; this method will be called MDAV (multivariate microaggregation based on Maximum Distance to Average Vector).

2.1.2.6 μ-ARGUS software. All these above mentioned methods have been implemented in the current versions of μ-ARGUS. We will continue to extend and improve μ-ARGUS, as our goal is to make all the SDC methodology easily available for the data-protectors. However it must be stressed that this software tools can only be applied by people with a basic understanding of the SDC-theory. They are not 'black-boxes', which will automatically produce a safe file.

The methods available in μ-ARGUS can be used to produce datafiles for different purposes. We make a basic distinction between datafiles that will be made available to established researchers at universities et al. (possibly with a contract) and datafiles which will be made available to the general public. It goes without saying that in this case the much more strict rules have to be applied.

For more information on the μ-ARGUS software we refer the μ-ARGUS (4.0)-manual (Hundepool et al, 2003)

2.2.3 Recent Extension of ARGUS

Till now μ-ARGUS could only read ASCII files, both fixed and free format. In general this is a quite flexible format that can be exported from many tools used in the statistical production at the NSI's. However users do not like exporting and importing the data between various packages. They rather prefer to stick with one package and want the other tools to operate on these files. So recently we have added the option to protect SPSS system files, enabling μ-ARGUS to read and update directly the SPSS-systemfiles. SPSS is a tool that at used in several NSI's, a.o. Statistics Netherlands. This extension has proved to be quite successful. However this might lead the wishes to build links to other packages.

2.2.4 Future Developments

During the writing of this paper the process of testing μ-ARGUS was not yet finished. The outcome of this process will be input of the further development of μ-ARGUS. Possible extensions that are foreseen now are the introduction of a batch-version like has been introduced in τ-ARGUS, the investigation of linking to SAS, the inclusion of record linkage software as has been developed during the final phase of the CASC project. Also special output formats, like the one needed for the transmission of data to Eurostat is considered. Depending on the conclusions of the testing some parts will be realized during this CENEX, other parts will be left to the future work.

2.3 τ-ARGUS

2.3.1 Introduction

Even in moderate sized tables there can be large disclosure risks. Take e.g. a cell in a table where there is only one contributor. The published cell value is clearly the contribution of one respondent/enterprise. However the situation is more complex. Several rules for identifying the unsafe cells have been proposed in the past. This still remains the easy part. The protection of the unsafe cells by suppressing so-called secondary cells in a table is an even more complex task.

2.3.2 Sensitive Cells in Magnitude Tables

Although there is a long tradition of using the so-called dominance rule, we propose to use the p% rule to identify the primary unsafe cells.

The basic idea is that a contributor to a cell has the best chances to estimate the contributions of competitors in a cell than an outsider and also that this kind of intrusions can occur rather often. The precision with which a competitor can estimate is a measure of the sensitivity of a cell. The worst case is that the second largest contributor will be able to estimate the largest contributor. If this precision is more than p% the cell is considered unsafe. An extension is that also the global knowledge about each cell is taken into account. In that case we assume that each intruder has a basic knowledge of the value of each contributor of q%.

Traditionally the well-known dominance rule is still often used to find the sensitive cells in tables, i.e. the cells that cannot be published as they might reveal information on individual records. More particularly, this rule states that a cell of a table is unsafe for publication if a few (n) major contributors to a cell are responsible for a certain percentage (k) of the total of that cell. The idea behind this rule is that in that case at least the major contributors themselves can determine with great precision the contributions of the other contributors to that cell. The choice n=3 and k=70% is not uncommon, but τ-ARGUS will allow the users to specify their own choice.

Internationally and also in the Netherlands there is a shift from the dominance rule towards the prior-posterior rule. The reasons for this are the more intuitive approach and the better numerical properties like the protection levels. Also waivers (contributors giving permission to publish their results) can be taken into account more easily. See Loeve (2001).

With these rules as a starting point it is easy to identify the sensitive cells, provided that the tabulation package has the facility not only to calculate the cell totals, but also to calculate the number of contributors and the n individual contributions of the major contributors. With this information τ-ARGUS can apply the sensitivity rules ands also perform the table redesign very easily. τ-ARGUS can produce the tables from the microdata files, also calculating the necessary additional information.

Traditionally τ-ARGUS could only read microdata files, but because of so many requests to be able to protect ready-made tables as well the next version of τ-ARGUS will have this facility. However with the restriction that the options for table redesign cannot be used any more and that only the sensitivity rules can be applied if also the largest contributions are provided.

A problem, however, arises when also the marginals of the table are published. It is no longer enough to just suppress the sensitive cells, as they can be easily recalculated using the marginals. Even if it is not possible to exactly recalculate the suppressed cell, it is possible to calculate an interval that contains the suppressed cell. This is possible if some constraints are known to hold for the cell values in a table. A commonly found constraint is that the cell values are all nonnegative.

If the size of such an interval is rather small, then the suppressed cell can be estimated rather precisely. This is not acceptable either. Therefore it is necessary to suppress additional information to achieve that the intervals are sufficiently large.

Several solutions are available to protect the information of the sensitive cells:

- Combining categories of the spanning variables (table redesign). Larger cells tend to protect the information about the individual contributors better.
- Suppression of additional (secondary) cells to prevent the recalculation of the sensitive (primary) cells.

The calculation of the optimal set (with respect to the loss of information) of secondary cells is a complex OR-problem. τ-ARGUS will be built around this solution and takes care of the whole process. A typical τ-ARGUS session will be one in which the users will first be presented with the table containing only the primary unsafe cells. The user can then choose how to protect these cells. This can be the combination of categories, equivalent to the global recoding of μ-ARGUS. The result will be an update of the table with presumably less unsafe cells (certainly not more). At a certain stage the user requests the system to solve the remaining unsafe cells by finding secondary cells to protect the primary cells.

2.3.3 Sensitive Cells in Frequency Count Tables

In its simplest way sensitive cells in frequency count tables are defined as those cells that contain a frequency that is below a certain threshold value. This threshold value is to be provided by the data protector. This way of identifying unsafe cells in a table is the one that is implemented in the current version of τ-ARGUS It should be remarked, however, that this is not always the adequate way to protect a frequency count table. A greater risk in frequency tables is the so called group disclosure. If from a table it can be deduced that all contributors to a cell have a certain characteristic, this characteristic is revealed for contributors to a cell (even many). This is also an undesirable situation. Current research at Statistics Netherlands aims at establishing better rules for these frequency tables.

2.3.4 Secondary Cell Suppression

Once the sensitive cells in a table - either of magnitude or a frequency count type - have been identified and there are not too many of these it might be a good idea to suppress these values. In case no constraints on the possible values in the cells of a table exist this is easy: one simply removes the cell values concerned and the problem is solved. In practice, however, this situation hardly ever occurs. Instead one has constraints on the values in the cells due to the presence of marginals and lower bounds for the cell values (typically 0). The problem then is to find additional cells that should be suppressed in order to protect the sensitive cells. The additional cells should be chosen in such a way that the interval of possible values for each sensitive cell value is sufficiently large. What is “sufficiently large” is to be specified by the data protector by specifying the protection intervals.

In general the secondary cell suppression problem turns out to be a hard problem, provided the aim is to retain as much information in the table as possible, which, of course, is a quite natural requirement. The optimisation problems that will then result are quite difficult to solve and require expert knowledge in the area of combinatorial optimisation.

2.3.4.1 Information loss in terms of cell weights. In case of secondary cell suppersssion it is possible that a data protector might want to differentiate between the candidate cells for secondary suppression. By specifying a cost-function he can influence the choice of the secondary suppressions. The cellvalue is a possibility, but also the cell frequency could be chosen or any other variable in the datafile. The aim of secondary cell suppression can be summarised by saying that a safe table should be produced from an unsafe one, by minimising the information loss, expressed as the sum of the weights associated with the cells that have secondarily been suppressed.

2.3.5 Solving the Secondary Cell Suppression Problem

Several approaches to solve this problem have been implemented in τ-ARGUS characteristics and advantages and disadvantages

- The hypercube method
- The optimal solution
- The partial optimal solution
- The network solution

2.3.5.1 The Hypercube Method. The approach builds on the fact that a suppressed cell in a simple n dimensional table without substructures cannot be disclosed exactly if that cell is contained in a pattern of suppressed, nonzero cells, forming the corner points of a hypercube.

The algorithm subdivides n-dimensional tables with hierarchical structure into a set of n-dimensional sub-tables without substructure. These sub-tables are then protected successively in an iterative procedure that starts from the highest level. Successively, for each primary suppression in the current sub-table, all possible hypercubes with this cell as one of the corner points are constructed.

For each hypercube, a lower bound is calculated for the width of the suppression interval for the primary suppression that would result from the suppression of all corner points of the particular hypercube. To compute that bound, it is not necessary to implement the time consuming solution to the Linear Programming problem. If it turns out that the bound is sufficiently large, the hypercube becomes a feasible solution. For any of the feasible hypercubes, the loss of information associated with the suppression of its corner points is calculated. The particular hypercube that leads to minimum information loss is selected, and all its corner points are suppressed. See Giessing (2002).

An implementation of this method by R. D. Repsilber offers a quick heuristic solution. The method has been implemented in τ-ARGUS. The advantages are the speed of the solution even for very large tables and the fact that this method does not require a licence for commercial OR-software like the other solutions. A disadvantage might be that the solution will not be the optimal one, leading to over-suppression

2.3.5.2 The Optimal Solution. Fischetti and Salazar (1998) has developed complex optimisation models to find the optimal solution for the secondary cell suppression. The models take into account the primary cells to be protected but also see to it that the cells cannot be recalculated to a given upper and lower protection level. These models have the flexibility to allow for different optimisation criteria, so it is possible to minimise the sum of the values of the cells to be suppressed, the sum of the frequencies of the individual cells of merely the number of cells to be suppressed.

The original Salazar models could only protect simple unstructured tables, but recently the models and the implementation have been extended for hierarchical and linked tables. Due to all the sub-totals present in these tables the intruder has many more options to recalculate suppression pattern and so the optimisation models have become much more complex.

It is to be expected that for very large tables the required computing time to find the optimal solution might be prohibitive in real life situations. But then alternatives are available in τ-ARGUS

The solution of these problems requires high performance OR-solvers which are only available commercially. In τ-ARGUS we have made provisions to solve the Salazar models with either Cplex or Xpress, two major solvers available.

2.3.5.3 The Modular Partitioning. In real life situation most tables of NSI's tend to have one or more hierarchical spanning variable. And real life tables tend to be very large. Given the numerical complexity the Salazar model can only handle moderate sized tables. To overcome these restrictions an approximation has been built which breaks down the large hierarchical table into many unstructured sub-tables. This results in a whole tree of small sub-tables. Starting at the top this method then protects all these tables. As sometimes the suppression pattern influences a higher level of the tree a backtracking procedure will be carried out.

At the end of this procedure the whole table is protected. It proves to be a reasonable quick procedure, which has enabled us to protect very large table. See De Wolf (2002).

2.3.5.4 The Network Solution. Networks are often used in optimisation problems as an approximation of the full optimal solution. The advantages are that the solutions are obtained rather quickly, often at high quality. Therefore networks have been studied in the SDC area for a longer time. However the conclusions were that networks can only be used properly for 2-dimensional; tables. On the first sight this might be a serious drawback, but many very large tables produced by the NSI's are 2-dimensional, e.g. the foreign trade statistics.

Jordi Castro (2003) has developed a network based solution, which is now available in τ-ARGUS. The first implementation only allowed for non-hierarchical tables, but an extension for hierarchical tables has been build as well, provided that only the first variable is hierarchical.

2.3.6 The τ-ARGUS Software

All the above mentioned solutions have been built in τ-ARGUS. The aim of τ-ARGUS is to make it into a control centre for tabular SDC. This will facilitate the users to apply the most appropriate method available for the problems he faces. Like with μ-ARGUS, τ-ARGUS is not a black box, which will just protect a table for you. τ-ARGUS is a control centre, which helps you to apple the appropriate SDC, measures and performs the complex computations involved.

For more information on τ-ARGUS software we refer the τ-ARGUS-manual (Hundepool et al, 2004)

2.3.7 New Developments in τ-ARGUS

Recently a lot of work has been done with respect to τ-ARGUS. Although τ-ARGUS has started as an interactive program with a GUI more and more people were asking for a batch version, mainly for automating the routine jobs. Secondary cell suppression works quite well for magnitude tables, for frequency tables often rounding is preferred. For users of the SuperCross tabulation package an interface between SuperCross and τ-ARGUS has been developed, making the power of τ-ARGUS directly available to the SuperCross users.

2.3.7.1 Batch Version. Often the same tables have to be protected each month or quarter. This led to the need of a batch-version of τ-ARGUS. It has been added recently and now most of the functionality of is available is batch. A batch instruction file could look like:

```
<OPENMICRODATA>  "…\tau_testW.asc"
<OPENMETADATA>   "…tau_testW.rda"
<SPECIFYTABLE>   "Size""Region"|"Var2"|""|""
<SAFETYRULE>     NK(3,75)|FREQ(3,30)
<READMICRODATA>
<SUPPRESS>       MOD(1)
<WRITETABLE>     (1,2,3,"… TestB.txt")
<SUPPRESS>       GH(1,100)
<WRITETABLE>     (1,1,3,"…\TestBGH.txt")
```

This instruction will read the data, using a RDA file, make a table with a NK-rule and a freq. rule, protect it first with the modular version and then with the hypercube. These instruction file can be generated automatically form the GUI-version of τ-ARGUS

An other advantage of the batch-version is that this enabled the users of τ-ARGUS that are not using Windows-PC's but e.g. Linux/Unix, to use τ-ARGUS as a remote service via some Windows server in their network

2.3.7.2 Rounding. Secondary cell suppression is a well known practice for magnitude tables. Rounding is often preferred for frequency tables. Thanks to a research project sponsored by ONS. JJ Salazar (2006) has developed a routine to controlled-round multidimensional tables using his optimisation models, taking into account specified protection intervals as well. The outcome of this project has led to an extension of τ-ARGUS, making this rounding procedure available to everyone. The current implementation can easily handle tables up to 150K cells. But as the ONS had to round even larger tables, we have introduced a partitioning method, breaking down very large tables to smaller (< 150K) subtables. In this way we have been able to round tables up to 6 Million cells.

Of course this rounding procedure can not only be used as a protection method for frequency tables, but also generally in all situations, when large multidimensional tables have to be rounding while preserving the additivity.

2.3.7.3 Link to SuperCross. At many NSI's SuperCross is a popular tabulation tool. For these NSI's SuperCross is a standard production tool. However it lacks the capacity to properly protect the tables. In order to meet these needs we have investigated a possible link between τ-ARGUS and SuperCross. This has led to a fruitful cooperation. SuperCross has now the capacity of exporting a table in a format suitable for τ-ARGUS. Via the batch-version the power of τ-ARGUS is invoked to protect a table. The table is then protected or rounded. The results are read back to SuperCross and are available in the SuperCross environment.

References

Benedetti, R. and Franconi, L. (1998), 'An estimation method for individual risk of disclosure based on sampling design'

J. Castro, "User's and programmer's manual of the network flows heuristics package for cell suppression in 2D tables" , research report DR 2003/07, Statistics and Operations Research Dept., Universitat Politècnica de Catalunya, 2003.

Fischetti, M. and J.J. Salazar-González (1998). *Models and Algorithms for Optimizing Cell Suppression in Tabular Data with Linear Constraints.* Technical Paper, University of La Laguna, Tenerife.

Giessing, S. and Repsilber, D. (2002), 'Tools and Strategies to Protect Multiple Tables with the GHQUAR Cell Suppression Engine', in '*Inference Control in Statistical Databases'* Domingo-Ferrer (Editor), Springer Lecture Notes in Computer Science Vol. 2316.

Anco Hundepool, Aad van de Wetering, Ramya Ramaswamy, Luisa Franconi, Alessandra Capobianchi, Peter-Paul de Wolf, Josep Domingo, Vicenç Torra, Ruth Brand and Sarah Giessing (2003), μ-ARGUS user manual version 4.0, *Statistics Netherlands, Voorburg.*

Anco Hundepool, Aad van de Wetering, Peter-Paul de Wolf, Sarah Giessing, Matteo Fischetti, Juan-José Salazar and Alberto Caprara (20043), τ-ARGUS user manual 3.1, *Statistics Netherlands, Voorburg*

Anneke Loeve (2001), Notes on sensitivity measures and protection levels, Research paper 0129, *Statistics Netherlands, Voorburg*

J.-J. Salazar (2006), Finding Good Rounded Tables, *paper submitted to the conference PSD'2006, Rome*

P.-P. De Wolf (2006), Risk, Utility and PRAM, *paper submitted to the conference PSD'2006, Rome*

P.-P. de Wolf (2002). HiTaS: a heuristic approach to cell suppression in hierarchical tables. in '*Inference Control in Statistical Databases'* Domingo-Ferrer (Editor), Springer Lecture Notes in Computer Science Vol. 2316

Software Development for SDC in R

M. Templ[1,2]

[1] Statistics Austria, Guglgasse 13, A-1110 Vienna
[2] Dept. of Statistics & Probability Theory, Vienna University of Technology, Wiedner Hauptstraße 8–10, A-1040 Vienna

Abstract. The production of scientific-use files from economic microdata is a major problem. Many common methods change the data in a way which leaves the univariate distribution of each of the variables almost unchanged towards the distribution of the variables of the original data, the multivariate structure of the data, however, is often ruined.

Which method are suitable strongly depends on the underlying data. A program system with which one can apply different methods and evaluate and compare results from different algorithms in a flexible way is needed. The use of methods for protecting microdata as an exploratory data analysis tool requires a powerful program system, able to present the results in a number of easy to grasp graphics. For this purpose some of the most populare procedures for anonymising micro data are applied in a flexible R-package. The R system supports flexible data import/export facilities and advanced developement tools for the development of such a software for disclosure control.

Additionally to existing algorithms in other software (MDAV algorithm for microaggregation, ...) some new algorithms for anonymising microdata are implemented, e.g. a fast algorithm for microaggregation with a projection pursuit approach. This algorithm outperforms existing other algorithms for most of real data.

For all this algorithms/methods print, summary and plot methods and methods for validation are implemented.

In the field of economics suppression of cells in marginal tables is likely to be the most popular method to protect tables for statistical agencies. The use of linear programming for cell suppression seems to be the best way of protecting tables and hierarchical tables.

Some R-packages for various fields of disclosure control are being developed at the moment. It is easy to learn the applications of disclosure control even with little previous knowledge because of its integrated online-help with examples ready to be executed.

1 Using R for Disclosure Control

R [35] is a open source statistics software package subjected to the GPL and therefore free and extendable for companies. R can be downloaded from the following website:

http://cran.r-project.org/

J. Domingo-Ferrer and L. Franconi (Eds.): PSD 2006, LNCS 4302, pp. 347–359, 2006.

Many methods and papers have been presented on disclosure control over the last years, but the underlying code is rarely available. The implementation of methods in software is required to evaluate the quality of the methods. For this purpose the free available, open-source, object-oriented, high-level language software R seems to be perfect.

Nowadays, R has become the standard statistical software. Thousands of people are involved in the development of R both at universities and companies and more than 700 add-on packages have been built in the last years.

R can deal with many different data formats and have very flexible data import/export facilities which is quite important when dealing with data from various formats. R can also communicate with various popular software and data bases. A major advantage of using R for disclosure control is that the facilities of R and also already implemented algorithms and graphical tools can be used easily. We do not need develop things new.

Applying different methods for disclosure control on data and evaluate and compare the results is a kind of explorative data analysis. For this purpose a object-oriented language like R is quite important.

Also very important is the reproducibility of results [28] when applying algorithms for disclosure control on data and when doing a validation of the results or comparing different results and making some nice graphics. For this, R is very well designed and in combination with LaTeX someone can produce dynamical reports, where LaTeX-code and R-code can be written and executed in/from one document together with the help of `Sweave` [26, 27].

Additionally it is very easy to develop your own packages with online helpfiles in R. With some developement tools of R one can make nice graphical user interfaces (GUI) like the Argus GUI's [21] as well.

Everybody can contribute code to the packages and help to make the project a success. The code is freely available under GPL and legally protected for commercial use of others (nobody has the right to use the code for an commercial implementation in software). The open source status and free use of the code should help to make the code better and better. Everybody is invited to check and upgrade the code.

The R system provides a powerful programming language and existing Fortran or C-code can easily embedded. R is the most powerful program system in the statistical world and in my opinion we don't have an alternative to R in the near future.

2 Microdata Protection

The methods for making microdata confidential vary considerably, depending on the scaling of the data. We will concetrate on continuous microdata only in order to stay within the limit of pages of this paper.

We want to give data to researchers and preserve confidentiality at the same time. There are few general concepts to do this. Some statistical agencies have decided to design confidentiality preserving model servers. E.g. the United States Bureau of the Census operates a number of *Research Data Centers* where researchers with special sworn status have access to specified microdata (see e.g. [39]). Researchers can apply models on data which can not be seen and the results of the models are checked by the Census staff. This approach preserves confidentiality but is not flexible, and in my point of view not compatible with a modern statistical world. When working with model servers the result of the model is the object of interest not the underlying data. But it is really difficult to apply methods and it is problematic to choose and evaluate models without seeing the data. The model can e.g. be influenced by outliers. Additionally, only very few methods are available on such model servers.

Much more flexible are remote access facilities. Researches can look at the data and can choose a suitable method for analysing the underlying data. Finally, the output should be checked by the staff of the statistical agencies and is, confidentiality preserved, usually to sent per e-mail the researchers (see e.g. [20] or [3]).

It is expensive and time-consuming to implement one of these two concepts.

The third approach for preserving confidentiality is to produce scientific use files by perturbation of microdata. The main goal is to produce a data set from the original data which preserves confidentiality and has a same structure as similar to the original data set as possible.

There are some concepts for this approach. The well-established concepts are *Microaggregation* [1], *Adding Noise* (see e.g. [24, 25]), *Rank Swapping* [7], *Blanking and Imputation* [17] and the generation of *synthetic data* with the same stucture as the original data (e.g. *Latin Hypercube Sampling* [23, 40, 42]).

All these methods have been implemented in various R-Packages.

3 Micoraggregation

On *http://neon.vb.cbs.nl/casc/Glossary.htm* we can find the "official" definition of Microaggregation: *"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."* [12]. More references can be found in [1, 9, 8, 10].

While for the proximity measure very different concepts can be used, microaggregation is naturally done with the mean. Nevertheless, other measures of location can be used for aggregation, especially when the group size for aggregation has been taken higher than 3. Since the median seems to be unsuitable for microaggregation due to it's rather high *breakdown point*, other measures like an *onestep from median* (see e.g. [19]) can be chosen. The *breakdown point* of an estimator measures the maximal percentage of the data points that may be contaminated before the estimates becomes completely corrupted.

3.1 Clustering at First

The package contains also a method with which the data can be clustered with a variety of different clustering algorithms. Clustering the observations before applying microaggregation might be useful. There is quite a lot of algorihms to do a clustering with, but for most of these cases Mclust [15] provides the best results. These technique, which is not based on distance measures, usually find the clusters by optimizing a maximum likelihood function. Avoid using hierarchical or classical partitioning cluster algorithms because hierarchical clustering algorithms result in worst partitions and classical partitioning algorithms, such as kmeans, result in spherical clusters with nearly the same size. Note that in our approach the data should be standardised before clustering, especially when the variables are of unequal scaling. Without standardising the data one variable might have the highest influence in each cluster and this is not what we want. Cluster analysis in general does not need normally distributed data. However, it is advisable that heavily skewed data are first transformed to a more symmetric distribution. If a good cluster structure exists for a variable, we can expect a distribution which has two or more modes. A transformation to more symmetry will preserve the modes but remove large skewness.

3.2 Methods Based on Sorting of Variables

We have developed a package called `Microaggregation` which contains methods like *individual ranking* [8], sorting based on a single variable [9, 1] and some related methods.

Cluster analysis can be applied before applying these methods. The clustered data can be sorted in each cluster depending on the most important variable in each of the clusters (in the package we call this method *influence*).

3.3 Projection Methods and MDAV

Package `Microaggregation` contains a good method called *mdav* **M**aximum **D**istance to **A**verage **V**ector (*mdav* is in turn an evolution of the multivariate fixed-size microaggregation in [10] proposed by the same authors). This method was first implemented in the μ-Argus software [22].

Another approach is to sort the data according to the first principal component (see e.g. in [38]) which is a well-documented method in SDC. Classical Principal Component Analysis (PCA) [33] is very sensitive to outlying observations since it is computed from eigenvalues and eigenvectors of the non-robust sample covariance matrix. Therefore applying PCA to sort the observations on the first principal component before aggregation may provide worst results in context of microaggregation.

In addition to that, package `Microaggregation` contains two major types for a robustification of this approach.

The first one calculates eigenvectors and eigenvalues based on robust estimates of the covariance matrix. The MCD-estimation [37] is the default for the estimation of the covariance matrix. Others, like M-estimation [30], the MVE estimator [37], the orthogonalized Gnanadesikan-Kettering (OGK) estimator [31] and some more can also be used. High breakdown point estimators for the covariance matrix are to be prefered. When using classical PCA or these robustified PCA all principal components must be estimated, but in the context of microaggregation we need only the first principal component.

The second approach avoids this and estimates the first (robust) principal component without covariance estimation. [6] has developed a method based on *projection pursuit* (PP) [29, 18]. With PP we search for directions with maximal variance of the data projected on it. Instead of using the classical variance estimator they use a robust scale estimator S_n as *projection pursuit index*. For a sequence of observations $x_1, \ldots, x_n \in \mathbb{R}^p$, the first "eigenvector" is defined as

$$v_{S_n,1} = \underset{||a||=1}{\text{argmax}} S_n(a^t x_1, \ldots, a^t x_n) \ . \tag{1}$$

The associated "eigenvalue" is then, by definition,

$$\lambda_{S_n,1} = S_n^2((v_{S_n,1})^t x_1, \ldots, (v_{S_n,1})^t x_n).$$

Li and Chen proposed working with an M-estimator of scale for S_n, and applied a general projection-pursuit algorithm for maximizing Formula (1), leading to an iterative and complicated computer intensive method. Nowadays there is a renewed interest in the projection-pursuit approach to PCA. Filzmoser [14] for instance applied it to a geostatistical problem. Perhaps the best known robust dispersion measure is the Median Absolute Deviation (MAD). For a sample $\{x_1, \ldots, x_n\} \subset \mathbb{R}$ it is defined as

$$\text{MAD}_n(x_1, \ldots, x_n) = 1.486 \ \underset{i}{\text{med}} |x_i - \underset{j}{\text{med}} x_j| \ , \tag{2}$$

where the constant 1.486 ensures consistency at normal distributions.

Primarily when having mixed structures in your data it is a good idea to cluster the data and apply the projection methods on each cluster. This can be easily done by setting up an optional parameter in a function in package `Microaggregation`.

It is really easy to use the package `Microaggregation` and apply its methods, because of the online-help files, the included examples and the simple handling of objects in R.

4 Adding Noise

4.1 S4-Class Style

A S4-class `R`-Packages called `AddNoise` was developed. Normally in other statistical softwares you have only few classes, like *numeric*, *character*, or *data set*. In `R` there are much more classes and additionally one can design your own classes (for e.g. class *addNoise*). The concept of S4-classes [5] in `R` is new and an extension of the traditional S3-class system in `R` [5]. Only few packages are written in S4 style up to now. S4 style code is very formal and you can define classes, plot-, print-, and summary methods. The advantages for the user of an S4-class package are that the packages are really flexible in use and mostly easy extendable with your own code. Additionally the user gets very precise error messages when operating the implemented functions in an incorrect way.

4.2 Methods

Beside the implementation of simple adding normal distributed noise and correlated noise [24], there are an implementation of Random Orthogonal Matrix Masking, called ROMM [41]. Note that this procedure preserve no confidentiality, e.g. the output of *biplots* [16] from the masked data and the original data are the same.

Additional there is a another concept of adding noise. Presume that observations which can be identified with diagnostic plots are confidential, then we can detect such observations with robust outlier detection methods. So, only these observations should be perturbed (and of course observations which can be identified with the help of key variables as well), depending on the outlyingness of these observations.

For the detection of critical observations, which may be identified by an data intruder, there are some (robust) outlier detection tools, like mahalanobis distances, robust mahahlanobis distances (see e.g. in [30]), jackknifing for results on the first 2 eigenvalues [11] and also on some univariate statistics which can be applied on Box-Cox transformed data [4]. A very good R-package for outlier identification is also package `mvoutlier` from Filzmoser [13].

5 Other Approaches for Continuous Microdata

A package called `rankSwapp` (*rank swapping*) and a few methods, like *latin hypercube sampling* and *blanking and imputation* have been developed.

While with the rank swapping approach the univariate structures of the data are nearly the same as for the original data, the multivariate structure is modified dramatically.

The results of latin hypercube sampling are not satisfactory, even not when doing some iterations.

6 Validation of the Results from Microdata Protection

First we want to give a short overview about existing validity measures, which are nearly almost univariate measures of information loss. After these, we want to propose other measures of information loss, which evaluates the multivariate structure of the original and the perturbed data.

One measure of information loss which is proposed by [32] is given by the original and the perturbed version of a observation i

$$IL1 = \frac{1}{d} \sum_{i=1}^{p} \frac{|x_{ij} - x_{ij}^{'}|}{\sqrt{2}S_j} \tag{3}$$

where S_j is the standard deviation of the j-th variable in the original data set. This measure of information loss does not evaluate how well univariate or multivariate statistics are preserved. This is a real disadvantage of this kind of measures and we want to show other measures which takes univariate and multivariate statistics into account.

In [32] there is also proposed a measure of disclosure risk, which based on distances and assumes that an intruder has additional information (disclosure scenarios) so that one can link the masked record of an individual to its original version [32]. Given the value of a masked variable, they check 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 [32].

Applying these measures on real data from a subset of the structural business statistics in Austria (one economic sector and 5 variables) we will see the IL1 and disclosure risk on the following graphics resulting from a part of the implemented algorithms for disclosure control. The algorithms which are chosen are M3 - M11 (microaggregation with aggregation level from 3 to 11 on different methods, ROMM with different parameters for the magnitude of perturbation [41], rank swapping with different maximum rank differences, which is expressed as a percentage of the total number of records (1%, 5%, 10%, 20%) and latin hypercube sampling [23].

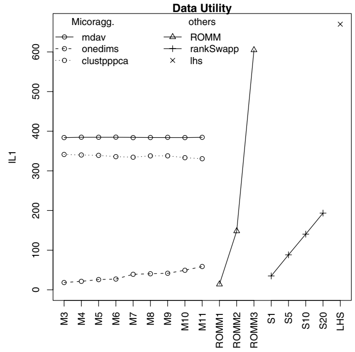


Fig. 1. Data utility resulting from different perturbation methods

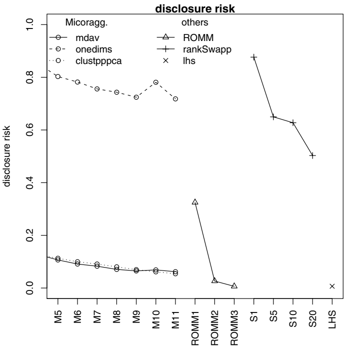


Fig. 2. Disclosure risk resulting from different perturbation methods

Previous Figures shows the polarity from data utility (Figure 1) and disclosure risk (Figure 2). Unfortunately, the previous measure of information loss tells us nothing about the quality of the perturbation, but we can anticipated that our proposed method with the projection pursuit approach (`clustpppca`) e.g. is more suitable than method `mdav`. Note, that we want to have anonymised data, which have the same statistical properties as the original data.

In the following we will evaluate some methods on real data from the structural business statistics in Austria on some statistical measures of information loss. The results based only on microaggregation methods, because our experience shows that only microaggregation fulfills the required amount of anonymisation and we want to be in the limit of pages in the paper.

```
> method <- c("simple", "single", "onedims", "pca", "pppca", "influence",
+     "clustpppca", "mdav")
> g <- valTable(x = x, method = method, measure = "onestep")
> g[, -c(9, 12, 13, 14)]

      method amean amedian aonestep devvar  amad   acov   acor  adlm apcaload
1     simple 1.809   0.538    0.186  2.859 0.582  1.430  0.606 0.012    0.099
2     single 0.995   0.322    0.301  2.933 0.318  1.466  2.573 0.017    0.051
3    onedims 0.100   0.007    0.004 66.549 0.007 33.274 39.750 0.194    1.416
4        pca 0.782   0.289    0.225  1.676 0.239  0.838  0.285 0.035    0.101
5      pppca 1.064   0.363    0.293  1.919 0.322  0.959  0.894 0.089    0.236
6  influence 0.943   0.207    0.231  2.169 0.289  1.084  1.107 0.041    0.222
7 clustpppca 0.996   0.226    0.161  2.025 0.229  1.013  1.577 0.057    0.369
8       mdav 1.225   0.267    0.226  3.131 0.364  1.566  7.022 0.016    0.288
```

The three columns of the previous table represents univariate measures of information loss, the following four columns represent multivariate measures of information loss followed by one columns representing differences in results from a classical regression models and followed by a mean difference of the loadings getting from a classical principal component analysis. A detailed description can be found in the online-help files from package `Microaggregation`. You can easily see that the individual ranking method (`onedims`) preserves univariate statistics quite well, but fails completely in the multivariate case. In many cases the principal component analysis method via projection pursuit on each cluster (found with a model based clustering algorithm) performs best.

Note that the data must not be aggregated with the mean. Sometimes it is useful to aggregate it with an another measure of location, like *onestep from median*. Observations, which are outside $\text{med}(x) \pm c.\text{mad}(x)$ have been put to these limits and then the mean is calculated. c is a constant to be chosen as in Formula (2). Additionally other robust measures like M-estimates [18] can be used, where data outside an robust interval are down weighted.

To evaluate differences between the original data and the perturbed data (or between two perturbation methods) in the univariate and multivariate structure of these data a variety of univariate and multivariate comparison plots is implemented. One plot is e.g. to compare the covariance structure of the original data. In Figure 3 you can see the covariances of the original data (black lines) in comparison with the covariances of the perturbed data (blue lines). Here the perturbation is done with the pca microaggregation method. The robustified pca with projection pursuit in each cluster are shown in Figure 4. While the results with sorting based on first classical principal component looking not really good, the perturbed data resulting from the robustified version of pca looks very well.

7 Protection of Hierarchical Tables

For this purpose a package called `disclosure` is under development.

The ordinary *primary cell suppression*, like *dominance rules*, *p*-percent rule, *pq*-percent rule (see e.g. in [12]) are implemented and additional cells can be easily suppressed by the user.

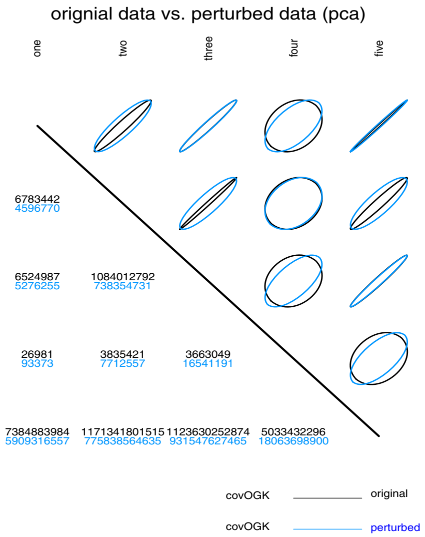


Fig. 3. Comparison plot of original and microaggregated data via sorting on the first classical principal component

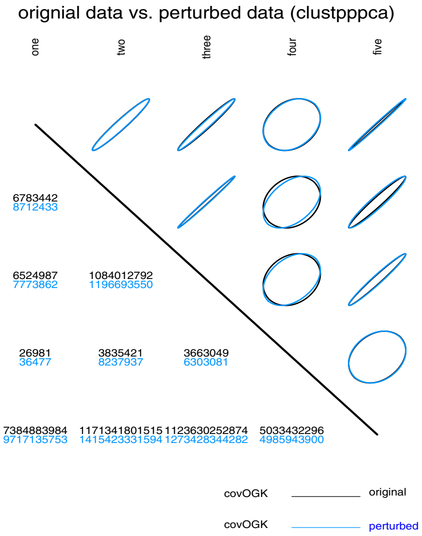


Fig. 4. Comparison plot of original and microaggregated data via sorting on the first robust principal component with projection pursuit in each cluster

For *secondary cell suppression* there is a method called *disc* [34] implemented. It is similar to the Hypercube method [36]. The disadvantage of this method is that after secondary cell suppression some primary protected cells can be computed too accurately and there is sometimes a little bit of over-suppression. A better strategy is to do a secondary cell suppression based on *linear programming*. The aim is e.g. to minimise the amount of suppressed cells (or other approaches) in (hierarchical) tables under the constraint that each primary suppressed cell can be computed only on a predefined interval (can not be computed too accurately). In R there is a wrapper function in C for the freely available (under LGPL2) linear program solver `lpSolve` [2] included, which can solve general linear/integer problems and more.

Perturbation and controlled rounding methods are not implemented yet (you can do this with τ-Argus [21]).

The usage of this package should be very easy, but note that there is no general solution yet been implemented for applying the tools of package *disclosure* automatically on varying hierarchical structures. Even for intruding some tables the package can be used in a very simple way. After loading data this should be carried out in one statement with some optional arguments. The functionality should be seen clearly from the online-help files of the package.

In the following I will show only a attacker problem from data aggregated at European Union level. Here is an example of such an table which is already hidden from member states and rules for suppression from Eurostat:

```
[1] "Table 1: EU level"

   C1  C2  C3  C4  C5  C6   EU
1  20  70  90  30  20  30  260
2 500   1   1  50   1  70  623
3  10   3   6  25  50  20  114
4  70  80  10 100  30  40  330
5 600 154 107 205 101 160 1327
```

```
[1] "Table2: Supp. from member
states + rules for agg."

   C1  C2  C3  C4  C5  C6   EU
1  20  70  90  30  NA  30   NA
2 500  NA  NA  NA  NA  70  623
3  10   3   6  NA  NA  20  114
4  70  NA  NA 100  30  40   NA
5 600 154 107 205 101 160 1327
```

We can easily attack these table with our implemented functions based on linear programming: In the following the table on the left side shows us the attacker solution from Table 2. On the right side we can see the solution when all additional EU-aggregates have been hidden (note that only aggregates can be hidden and cells from member states must not be hidden).

```
> library(disclosure)
> i <- c(1, 2, 4, 5)
> lp2.hier(e, e1)$lp.out2[, i]

      min max nrow ncol
 [1,]  18  71    1    5
 [2,] 258 311    1    7
 [3,]   0  53    2    2
 [4,]   0  11    2    3
 [5,]   0  53    2    4
 [6,]   0  53    2    5
 [7,]  22  75    3    4
 [8,]   0  53    3    5
 [9,]  28  81    4    2
[10,]   0  11    4    3
[11,] 279 332    4    7
```

```
> e1[2, 7] <- NA
> e1[3, 7] <- NA
> lp2.hier(e, e1)$lp.out2[, i]

      min max nrow ncol
 [1,]   0  71    1    5
 [2,] 240 311    1    7
 [3,]   0  81    2    2
 [4,]   0  11    2    3
 [5,]   0  75    2    4
 [6,]   0  71    2    5
 [7,] 570 808    2    7
 [8,]   0  75    3    4
 [9,]   0  71    3    5
[10,]  39 185    3    7
[11,]   0  81    4    2
[12,]   0  11    4    3
[13,] 240 332    4    7
```

You can easily compare the protection intervals (on the output of function lp.hier you can see their row number and column number) with the true values. We can see that the table is still not protected.

8 Conclusions

Using R for disclosure control has many advantages. The data import/export facilities are very flexibility and powerful and this is really useful in context of disclosure control. With R you can see the perturbation of microdata as a kind of explorative data analysis. We have a powerful system to analyse and evaluate the results and methods during the process of masking data. For all methods there are print, summary and plot methods implemented. Diagnostic plots and plots for the comparison of the original and the perturbed data are very useful during the process of perturbation. Several methods can be evaluated on basis of your data, and diagnostic tools can be used at the same time. The open source packages are highly extendable and can be well documented by use of online-help pages, vignettes and integrated examples. When calculation time is important the code can be written in C or Fortran and can be included in an R package easily. Everybody is invited to contribute to these packages or to make your own R packages for disclosure control.

The robustification of the pca approach for microaggregation with projection pursuit on each cluster leads often to the best results compared with other microaggregation methods.
Also the extension of the single axis method by sorting in each cluster by the most influencial variable can provide good results. Adding noise can be also provide good results and the perturbation must not be applied on all observations but rather on observations which can be identified probalby.

While cell perturbation and controlled tabular adjustment seems to be very good methods for protecting hierarchical tables some statistical agencies will keep hold on traditional suppression methods for a variety of reasons.

In my point of view statistical agencies will have commercial software with guaranteed support for disclosure control or they want to have free available and modifiable open-source software for disclosure control. Having open source code in an attractive software system for doing disclosure control might be a first step forward to harmonize the application of methods for all european statistical agencies.

Bibliography

[1] N. Anwar. Micro-aggregation - the small aggregates method. In *Internal report.* Luxembourg: Eurostat, 1993.

[2] M. Berkelaar, J. Dirks, K. Eikland, and P. Notebaert. lpsolve ide v5.5, 2006.

[3] L. Borchsenius. New developements in the danish system for access to micro data. In *Monographs of official statistics, Work session on statistical data confidentiality.* Eurostat, Luxembourg, 2005.

[4] G.E.P. Box and D.R. Cox. An analysis of transformations. *Journal of the Royal Statistical Society*, pages 211–252, 1964.

[5] J.M. Chambers. *Programming with Data.* Springer, New York, 1998. ISBN 0-387-98503-4.

[6] C. Croux and A. Ruiz-Gazen. High breakdown estimators for principal components: the projection-pursuit approach revisited. *Journal of Multivariate Analysis*, 95:206–226, 2005.

[7] T. Dalenius and S.P. Reiss. Data-swapping: A technique for disclosure control. In *Proceedings of the Section on Survey Research Methods*, volume 6, pages 73–85. American Statistical Association, 1982.

[8] D. Defays and Anwar M.N. Masking microdata using micro-aggregation. *Journal of Official Statistics*, 14(4):449–461, 1998.

[9] D. Defays and P. Nanopoulos. Panels of enterprises and confidentiality: the small aggregates method. In *Proceedings of the 1992 Symposium on Design and Analysis of Longitudinal Surveys*, pages 195–204. Statistics Canada, Ottawa, 1993.

[10] J. Domingo-Ferrer and J.M. Mateo-Sanz. Practical data-oriented microaggregation for statistical disclosure control. *IEEE Trans. on Knowledge and Data Engineering*, 14(1):189–201, 2002.

[11] R.G. Efron and R.G. Tibshirani. *An Introduction to the Bootstrap.* Chapman and Hall, New York, 1993.

[12] M. Elliot, A. Hundepool, E.S. Nordholt, J-L. Tambay, and T. Wende. Glossary on statistical disclosure control, 2005.

[13] P. Filmoser. A multivariate outlier detection method. In S. Aivazian, P. Filzmoser, and Y. Kharin, editors, *Proceedings of the Seventh International Conference on Computer Data Analysis and Modeling*, volume 1, pages 18–22. Belarusian State University, Minsk, 2004.

[14] P. Filzmoser. Robust principal component and factor analysis in the geostatistical treatment of environmental data. *Environmetrics*, 10:363–375, 1999.

[15] C. Fraley and A.E. Raftery. How many clusters? which clustering method? answers via model-based cluster analysis. *The Computer Journal*, 41(8):578–588, 1998.

[16] K.R. Gabriel. The biplot graphic display of matrices with application to principal component analysis. *Biometrika*, 58(3):453–467, 1971.

[17] R. Griffin, A. Navarro, and L. Flores-Baez. Disclosure avoidance for the 1990 census. In *Proceedings of the Section on Survey Research Methods*, pages 516–521. American Statistical Association, 1989.

[18] P.J. Huber. Projection pursuit. *Ann. Statist.*, 13:435–525, 1985.

[19] B. Hulliger. Simple and robust estimators for sampling. In *Proceedings of the Survey Research Methods Section*, pages 54–63. American Statistical Association, 1999.

[20] A. Hundepool and P-P. de Wolf. Onsite@home: Remote access at statistics netherlands. In *Monographs of official statistics, Work session on statistical data confidentiality.* Eurostat, Luxembourg, 2005.

[21] A. Hundepool, R. Ramaswamy, de Wolf P-P., L. Franconi, S. Giessing, D. Repsilber, J.J. Salazar, C. Castro, G. Merola, and P. Lowthian, 2003.

[22] A. Hundepool, A. Van deWetering, Ramaswamy R., L. Franconi, A. Capobianchi, P-P. DeWolf, J. Domingo-Ferrer, V. Torra, R. Brand, and S. Giessing. μ-argus version 3.2 software and users manual, 2005.

[23] R.L. Iman and W.J. Conover. A distribution-free approach to inducing rank correlation among input variables. *Communications in Statistics*, B11:311–334, 1982.

[24] J.J. Kim. A method for limiting disclosure in microdata based on random noise and transformation. In *Proceedings of the Section on Survey Research Methods*, pages 303–308. American Statistical Association, 1986.

[25] J.J. Kim and W.E. Winkler. Masking microdata files. In *Proceedings of the Section on Survey Research Methods*, pages 114–119. American Statistical Association, 1995.

[26] F. Leisch. Sweave: Dynamic generation of statistical reports using literate data analysis. In Wolfgang Härdle and Bernd Rönz, editors, *Compstat 2002 — Proceedings in Computational Statistics*, pages 575–580. Physica Verlag, Heidelberg, 2002. ISBN 3-7908-1517-9.

[27] F. Leisch. Sweave, part I: Mixing R and $\LaTeX$. *R News*, 2(3):28–31, December 2002.

[28] F. Leisch and A.J. Rossini. Reproducible statistical research. *Chance*, 16(2):46–50, 2003.

[29] G. Li and Z. Chen. Projection-pursuit approach to robust dispersion matrices and principal components: primary theory and monte carlo. *J. Amer. Statist. Ass.*, 80:759–766, 1985.

[30] R.A. Maronna. Robust m-estimators of multivariate location and scatter. *The Annals of Statistics*, 4(1):51–67, 1976.

[31] R.A. Maronna and R.H. Zamar. Robust multivariate estimates for highdimensional datasets. *Technometrics*, 44:307–317, 2002.

[32] J.M. Mateo-Sanz, F. Sebe, and J. Domingo-Ferrer. Outlier protection in continuous microdata masking. *Lecture Notes in Computer Science, Vol. Privacy in Statistical Databases, Springer Verlag*, 3050:201–215, 2004.

[33] K. Pearson. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, 6(2):559–572, 1901.

[34] K. Piker. Geheimhaltung - allgemeiner programmablauf, 1995.

[35] R Development Core Team. *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria, 2006. ISBN 3-900051-07-0.

[36] R. D. Repsilber. Preservation of confidentiality in aggregated data. In *paper presented at the Second International Seminar on Statistical Confidentiality.* Luxembourg, 1994.

[37] P. Rousseeuw. Multivariate estimation with high breakdown point. In *Mathematical Statistics and Applications*, pages 283–297. Akademiai Kiado, Budapest, 1985.

[38] M. Schmid. The effect of single-axis sorting on the estimation of a linear regression., 2006.

[39] P. Steel and A. Reznek. Issues in designing a confidential preserving model server. In *Monographs of official statistics, Work session on statistical data confidentiality.* Eurostat, Luxembourg, 2005.

[40] M.L. Stein. Large sample properties of simulations using latin hypercube sampling. *Technometrics*, 29:143–151, 1987.

[41] D. Ting, S. Fienberg, and M. Trottini. Romm methodology for microdata release. In *Monographs of official statistics, Work session on statistical data confidentiality.* Eurostat, Luxembourg, 2005.

[42] G.D. Wyss and K.H. Jorgensen. Sandia's latin hypercube sampling software. Technical report sand98-0210, Sandia National Laboratories, Albuquerque, NM, 1998.

On Secure e-Health Systems

Milan Marković

Banca Intesa ad Beograd, III Bulevar 1c,
11070 Belgrade, Serbia
Milan.markovic@bancaintesabeograd.com

Abstract. This paper is devoted to e-healthcare security systems based on modern security mechanisms and Public Key Infrastructure (PKI) systems. We signified that only general and multi-layered security infrastructure could cope with possible attacks to e-healthcare systems. We evaluated security mechanisms on application, transport and network layers of ISO/OSI reference model. These mechanisms include confidentiality protection based on symmetrical cryptographic algorithms and digital signature technology based on asymmetrical algorithms for authentication, integrity protection and non-repudiation. User strong authentication procedures based on smart cards, digital certificates and PKI systems are especially emphasized. We gave a brief description of smart cards, HSMs and main components of the PKI systems, emphasizing Certification Authority and its role in establishing cryptographically unique identities of the valid system users based on X.509 digital certificates. Emerging e-healthcare systems and possible appropriate security mechanisms based on proposed Generic CA model are analyzed.

Keywords: E-healthcare systems, Multilayered security systems, PKI systems, smart cards.

1 Introduction

The low-cost nature of the Internet coupled with the ease of making transactions has led to an explosive growth in e-business but trust in this medium is still a major concern. E-security is the foundation that enables trust in e-business [1], [2]. In this sense, main cryptographic aspects of modern TCP/IP computer networks are: digital signature technology based on asymmetrical cryptographic algorithms, data confidentiality by applying symmetrical cryptographic systems, and Public Key Infrastructure (PKI) systems.

The Internet is also changing the way the healthcare industry does business. It offers astounding opportunities to share information between healthcare professionals and to reduce the costly paper trail. However, organizations must create secure architecture to protect the privacy of patient records since main security requirements in healthcare, as well as in emerging mobile healthcare, systems include privacy and integrity of information related to patients. Such information includes information related to person, medical service given, and e.g. social status, and should be kept out of reach of unauthorized persons. Healthcare security benefits are: protect patient confidentiality from network-based violations, securely provide information to remote

J. Domingo-Ferrer and L. Franconi (Eds.): PSD 2006, LNCS 4302, pp. 360 – 374, 2006.

physicians, partners, and branch offices, and comply with government regulations on network security.

This paper is devoted to e-healthcare security systems based on PKI systems. We signified that only a general and multi-layered security infrastructure could cope with possible attacks to e-healthcare systems. We evaluated security mechanisms on application, transport and network layers of ISO/OSI reference model and gave examples of the today most popular security protocols applied in each of the mentioned layers. These mechanisms include confidentiality protection based on symmetrical cryptographic algorithms and digital signature technology based on asymmetrical algorithms for authentication, integrity protection and non-repudiation.

2 Multilayered Security Infrastructure in e-Healthcare Systems

Like in all the other electronic business systems, key security features that should be included in modern medical computer networks are: user and data authentication, data integrity, non-repudiation, and confidentiality. This means that in secure e-healthcare systems, the following features must be realized:

- strong user authentication both for doctors and other medical employees, as well as for patients,
- integrity of medical data transferred either via wired or wireless IP networks should be ensured and
- the non-repudiation function should be implemented.

These features are to be implemented by using digital signature technology based on asymmetrical cryptographic algorithms. Besides, the confidentiality and privacy protection of transferred data must be preserved during whole transmission and they are to be done by using symmetrical cryptographic algorithms. Also, strong user authentication techniques based on smart cards are to be implemented.

In this Section, we will give the overview of modern security mechanisms with particular emphasis on their use in medical electronic business systems and classical and mobile healthcare systems. The considered security mechanisms are based on PKI systems, digital certificates, digital signature technology, confidentiality protection, privacy protection, strong user authentication procedures and smart card technology. An overview of these techniques is given in [3].

In order to preserve the potential malicious attacks to the particular network, the multilayered security architecture has to be implemented. Modern computer networks security systems consist of security mechanisms on three different ISO/OSI reference model layers:

- Application level security (end-to-end security) based on the strong user authentication, digital signature, confidentiality protection, digital certificates and hardware tokens (e.g. smart cards),
- Transport level security based on establishment of a cryptographic tunnel (symmetric cryptography) between network nodes and strong node authentication procedure,
- Network IP level security providing bulk security mechanisms on network level between network nodes – protection from the external network attacks.

These layers are projected in a way that a vulnerability of the one layer could not compromise the other layers and then the whole system is not vulnerable.

2.1 Application Level Security Mechanisms

Application level security mechanisms are based on asymmetrical and symmetrical cryptographic systems, which realize the following functions:

- Authenticity of the relying parties (asymmetrical systems),
- Integrity protection of transmitted data (asymmetrical systems),
- Non-repudiation (asymmetrical systems),
- Confidentiality protection on application level (symmetrical systems).

The most popular protocols in domain of application level security are: S/MIME, PGP, Kerberos, proxy servers on application level, SET, crypto APIs for client-server applications, etc. Most of these protocols are based on PKI X.509 digital certificates, digital signature technology based on asymmetrical algorithms (e.g. RSA) and confidentiality protection based on symmetrical algorithms (e.g. DES, 3DES, IDEA, AES, etc.) [4]. Most of the modern application level security protocols, such as: S/MIME and crypto APIs in client-server applications are based on digital signature and digital envelope technology.

In modern e-commerce and e-business systems, asymmetrical algorithms (e.g. RSA) are mainly used according to PKCS#1 standard. PKCS#1 standard [5] describes a method for encrypting data using the RSA public-key cryptosystem. Its intended use is in the construction of digital signatures and digital envelopes, according to the syntax described in PKCS#7 standard. There is a lot of work on optimization of RSA algorithm implementation in hardware security module (HSM) based on signal processor [6], [7], [8], [9]. For digital signatures, the content to be signed is first reduced to a message digest with a message-digest algorithm (such as MD5), and then an octet string containing the message digest is encrypted with the RSA private key operation of the signer of the content. The content and the encrypted message digest are represented together according to the syntax in PKCS#7 to yield a digital signature. For digital envelopes, the content to be enveloped is first encrypted by a symmetric encryption key with a symmetric encryption algorithm (such as DES, 3DES, IDEA, AES, ...), and then the symmetric encryption key is encrypted with the RSA public key of the intended recipient of the content. The encrypted content and the encrypted symmetric encryption key are represented together according to the syntax in PKCS#7 to yield a digital envelope.

Security systems on application level consist also of the user authentication procedure which could be one, two or three-component authentication procedure.

2.2 Transport Level Security Mechanisms

Security mechanisms on transport level generally include confidentiality protection of transmitted data based on symmetrical cryptographic algorithms. These systems are mostly based on establishing the cryptographic tunnel between two network nodes on transport level. The establishment of the tunnel is preceded by strong authentication procedures. In this sense, the systems are based both on symmetrical algorithms for realization of cryptographic tunnel and a bilateral challenge-response authentication procedure based on asymmetrical algorithms and PKI digital certificates for authentication of the nodes and for establishing the symmetrical session key for this tunnel session. The transport level security system is mostly used for communication protection between client with Internet browser programs (Internet Explorer, Netscape Navigator, etc.) and WEB server, and the most popular protocols are: SOCKS (used earlier), SSL/TLS and WTLS. Between them, the most popular is SSL (Secure Sockets Layer) protocol (or Transport Layer Security (TLS)), which is used for protection between client browser program and WEB server. Furthermore, the SSL is the most popular and the far widest used security protocol today.

SSL protocol consists of two phases: authentication phase with bilateral exchanging of PKI digital certificates of the WEB server and the client (optional) and establishing the symmetrical session key and secure communication based on symmetrical algorithm and established session key (cryptographic tunnel). SSL protocol is placed just below the application layer of the ISO/OSI reference model and just on top of the TCP/IP layer (transport layer). This means that the SSL is not necessarily used only under the HTTP protocol but could be used also under some other application level protocols, such as: POP3, SMTP, etc.

WTLS (Wireless Transport Layer Security) protocol is a kind of wireless version of SSL protocol and serves for transport level protection between microbrowsers on WAP (Wireless Application Protocol) enabled GSM mobile phones and WAP servers, based on the same principles and functionality as the SSL protocol. This way, WTLS protocol is intended to use for secure communication in wireless networks (GSM), and is implemented in most of microbrowsers and WAP servers. WTLS protocol uses special digital certificates for wireless communication (WAPCerts).

2.3 Network Level Security Mechanisms

Network level security mechanisms include security mechanisms implemented in communication devices and firewalls, as well as operating system security mechanisms, etc. These methods represent the basis for realization of Virtual Private Networks (VPN). Security protection is achieved by encrypting the complete IP traffic (link encryption) between two network nodes. The most popular network layer security protocols are: IPSec (AH, ESP, IKE), packet filtering and network tunneling protocols, and the widest used is IPSec. Like transport level security protocols, IPSec consists also of network node authentication based on asymmetrical cryptographic algorithms and link encryption based on symmetrical algorithms. IPSec represents a group of protocols consisting of Authentication Header (AH), Encapsulated Security Payload (ESP) and Internet Key Exchange (IKE) protocols in transport and tunnel modes. AH is used for authentication IP packets, ESP is used for encryption and authentication the payload of the IP packets and IKE is used for authentication of the communication nodes and IPSec session key establishment. The most secure IPSec protocol is ESP in tunnel mode, since attacker does not know internal addresses (source and destination) – only addresses of IPSec gateways could be seen externally.

Firewalls also belong to network security mechanisms and could be computers, routers, workstations. Their main characteristics are to define which information and services of internal network could be accessed from the external world and who from internal network is allowed to use information and services from the external network. Firewalls are mostly installed at breakpoints between insecure external networks and secure internal network. Depending of the needs, firewalls consist of the one or more functional components from the following set: packet filter, application level gateway, and circuit level gateway. In this sense, there are four traditional examples of firewalls: Packet Filtering Firewall, Dual-Homed Firewall (with two network interface), Screened Host Firewall, Screened Subnet Firewall (with DeMilitarized Zone (DMZ) between internal and external networks). Nowadays, firewall devices are in fact multifunctional devices that include very sophisticated security mechanisms, such as: several firewall interfaces allowing more detailed secure separation of the network, antivirus, Intrusion Prevention, content filtering, VPN concentrator functionalities, etc.

3 PKI Systems

Public-key cryptography uses a combination of public and private keys, digital signature, digital certificates, and trusted third party Certification Authorities (CA), to meet the major requirements of e-business security. Before applying the security mechanisms you need the answers for the following questions: Who is your CA? Where do you store your private key? How do you know that the private key of the person or server you want to talk to is secure? Where do you find certificates?

A Public Key Infrastructure (PKI) provides the answers to the above questions. In the sense of X.509 standard, the PKI system is defined as the set of hardware, software, people and procedures needed to create, manage, store, distribute and revoke certificates based on public-key cryptography.

PKI system provides a reliable organizational, logical and technical security environment for realization of the four main security functions of the e-business systems: authenticity, data integrity protection, non-repudiation and data confidentiality protection. PKI systems are based on digital certificates as unique cryptographic based electronic IDs of relying parties in some computer networks.

PKI system consists of the following components: Certification Authority (CA) – responsible for issuing, renewing and revoking certificates, Registration Authorities (RAs) – responsible for acquiring certificate requests and checking the identity of the certificate holders, Systems for certificate distribution – responsible for delivering the certificates to their holders, Certificate holders (subjects) – people, machines or software agents that have been issued with certificates, CP, CPS, user agreements and other basic CA documents, systems for publication of issued certificates and Certificate Revocation Lists (CRLs), and PKI applications (secure WEB transactions, secure E-mail, secure FTP, VPN, secure Internet payment, secure document management system – secure digital archives, access control system, etc.)

The method defined in X.509 for revoking certificates involves the use of a certificate revocation list (CRL). This list identifies revoked certificates and is signed and timestamped by the CA. Normally, each certificate is identified by a unique serial number that is assigned when the CA issues it. The CA publishes the CRL, at regular intervals, into the same public repository (e.g. LDAP) as the certificate themselves (only certificates with owner permissions could be published).

There are several types of CA: corporate CAs, closed user group (CUG) CA, CA of vertical industries, and public CAs. Regarding the implementation approach, CAs could be divided to: outsourced CA – when some organization use certification services from the earlier established CA, and insourced CA – when some organization establishes its own CA services (bought on the market or inhouse developed). In all cases, all CA organization mostly used the CA software-hardware technology from the established CA technology vendors, such as: Entrust, Cybertrust, Cryptomathic, Utimaco, SmartTrust, RSA Data Security, etc.

In the following, a brief description is given of the generic model of the Certification Authority software-hardware system which is realized as a web multitier architecture. The described system is similar to the most modern and most secure PKI systems today. Also, some possible variants of system realization depending on the set of the requests that should be fulfilled are discussed. This generic CA represents a solution which could be fully customized to be adapted to the customer requirements.

3.1 Main Features of the Generic CA System

The generic CA is a WEB-based Certification Authority system which could support both closed PKI systems with strictly defined users of usually only one or two different user profiles, as well as public PKI systems with more user profiles and more different ways of user registration. The generic CA system represents the public CA system fully customizable to the particular requests of different users. Main features of the generic CA system are the following:

- The system fulfils all worldwide PKI standards and could be customized according to both adding new features and customizing the applied cryptographic algorithms.
- The generic CA is WEB multitier CA application which is based on smart cards for users.
- Generic CA system supports different database servers, such as: MS SQL Oracle and IBM DB2.
- The generic CA supports a working system with one asymmetrical keypair, with two keypairs and combined system.
- The generic CA supports a hierarchical PKI structure and has the off-line Root CA and more on-line Intermediate CAs. As an example, each user profile should have its Intermediate CA server.
- The generic CA supports different ways of the user registration, such as: through registration authorities (RA) and RA operators (RAO), as well as directly (for specific user profiles) via WEB CA server.
- The generic CA has implemented a procedure of distributed responsibilities (secret sharing, necessity of presence of number of specific users) in sense of creating the Root CA asymmetrical private key for generating the new Intermediate CA certificate.
- The generic CA has a support for life cycle certificate management (renewal, suspension, revocation).
- The generic CA has possibilities for electronic personalization of the smart cards and this could be done by client themselves, RAO or CA Operators (CAO).
- The generic CA system has a support for printing PIN code (lettershop) for accessing the cards which should be sent to the user separately from the smart card.
- The generic CA system provides the printing of different reports depending of the user needs.

3.2 System Architecture of the Generic CA System

A system architecture of the described generic CA system is given on Fig. 1. What missing on the Fig. 1 are application servers from different business processes which use the generic CA system. For example, WEB server in DMZ zone could be a business WEB server which will eventually realize strong authentication procedure of the users with smart cards, issued by the described generic CA system.

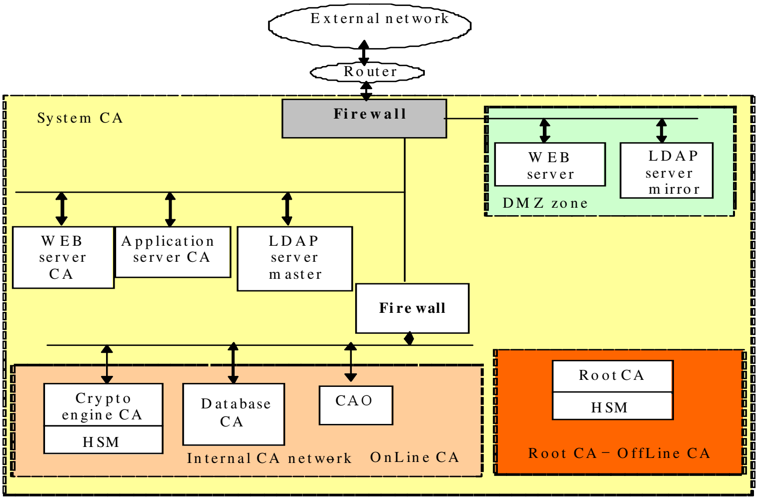


Fig. 1. A simplified network configuration of the generic CA system

As it could be seen from the Fig. 1, the generic CA system consists of OnLine and OffLine parts. OffLine part represents RootCA which is used only in rare cases when the Root CA asymmetrical private key should be activated for a purpose of generating a new Intermediate CA certificate in the hierarchical structure shown on Fig. 2, which is the most popular PKI structure in the modern PKI systems. Root CA is located in totally separated room from the rest part of the CA where there exist a vault in which individual activation parts of the Root CA asymmetrical private key are securely stored. These parts are used according to the defined procedure of “distributed responsibilities” (or “secret sharing”) in cases of generating new Intermediate CA certificates (this procedure is called “CA ceremony”). Eventually, the Root CA could be also in the same room (if necessary) as the OnLine CA but, as mandatory request, outside the LAN network and with mandatory vault for storing the activation parts of the Root CA private key.

In the CA ceremony procedure, it must be present a corresponding minimum number of special CA employees (custodians) who have access to the corresponding individual activation parts, stored on smart cards in special separated boxes of the vault, for activating the private key. Namely, a corresponding pre-defined number of smart cards must be present in order to activbate the Root CA private key in HSM device of the Root CA server, fully in accordance with General and Internal CA practices. After that, a new Intermediate CA asymmetrical keypair is generated and the Intermediate CA certificate is created (digitally signed) by a digital signature applying the Root CA private key in the Root CA's HSM. The encrypted private key and certificate of the Intermediate CA will be programmed (or generated directly onto

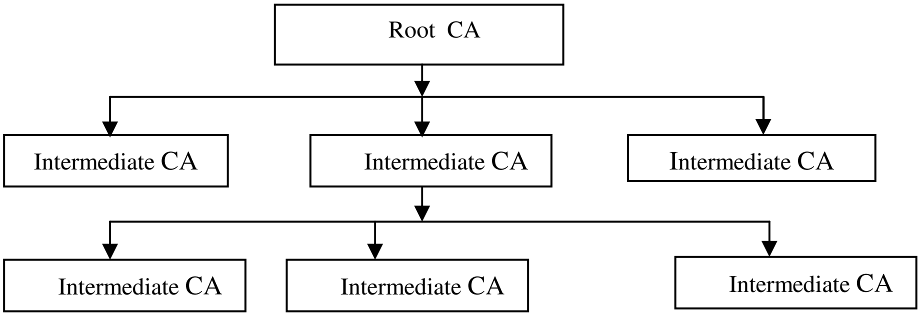


Fig. 2. Modern hierarchical structure of the Certification Authorities

the Intermediate CA's HSM) into the new smart card (Intermediate CA smart card) which will be installed into the HSM device of the new Crypto Engine server, intended for use as an OnLine CA for this Intermediate CA system. After that, Root CA private key will be deactivated from the Root CA HSM device and the smart cards with activation parts of the Root CA private key will be returned to the vault. As it could be concluded, it is possible that more Intermediate CA simultaneously work in OnLine working mode, i.e. that more Intermediate CA Crypto Engine servers are activated in the OnLine working mode for digital certificate generation (e.g. Intermediate CA for different kind of medical institutions). OnLine and OffLine parts of the generic CA system should support the use of the HSM modules, see Fig. 1.

In DMZ zone, besides WEB server, there is a LDAP mirror server which serves for publishing the CRL and ARL lists, as well as for eventual publication of issued digital certificates. This server is a copy of the master LDAP server which is located in the internal zone.

The generic CA system supports different methods for user registration, and the system is fully flexible to support different requests regarding ways of user registration according to the adopted documents (Certificate Policy and CPS) which defines the appropriate user profiles (both for individuals and legal persons). The system supports the issuing of digital certificates on different media (smart cards, mini CD, etc.) and enable functioning in the system with one or two asymmetrical keypairs. In the generic CA system, the certificate life-cycle management is implemented and comprises of the following procedures:

- Certificate renewal,
- Certificate suspension and reactivation,
- Certificate revocation.

These functions are implemented in accordance with Certificate Policy and Certificate Practise Statement of this CA system. In this case, the user will be enabled to make certificate renewal by himself, while the suspension and revocation will be done exclusively by the RAO and CA employees, according to the written procedure in Certificate Policy and CPS.

It should be mentioned that described architecture of the Generic CA system could be one example of possible realization of modern CA system and that actual implementations are more or less different depending on the way of key generation for users, the way of distribution keys and certificates, as well as on ways of CRL publishing. However, although there are differences, basic principles and concepts of the modern certification authorities are the same as in the described example. In this sense, the described Generic CA system could be a good candidate for establishing some e-healthcare PKI systems.

4 Smart Cards and Hardware Security Modules

Software only security solutions are not safe and are very vulnerable to some attacks (e.g. Trojan horse). There are several reasons why SW only security systems are not suitable: certificate and private key are stored on conventional media which is not secure, consumers are tied to their PC and are thus not mobile, and consumers are to manage certificates, which is not simple.

Hardware security modules (HSM) represent very important security issue of the modern computer networks. Main purposes of the HSM are twofold: increasing the overall system security and accelerating cryptographic functions (asymmetric and symmetric algorithms, key generation, etc.). HSMs are intended mainly for use in server applications and, optionally for client sides too in case of specialized information systems (government, military, police) [10]. For large individual usage, smart cards are more suitable as hardware security modules. However, for large usage, the best approach is in the combination of SW and smart card solutions for the best performance. Namely, smart card increases security and SW increases the total processing speed. In this sense, the most suitable large-scale solution consists of: SW for bulk symmetric data encryption/decryption plus hash calculations and smart card for digital envelop retrieval and digital signature generation. In modern and most secure PKI systems, two asymmetrical keypairs are used: one for digital envelope retrieval and the other for digital signature generation. In short, smart cards are credit-card sized plastic card with an embedded computer chip. There are several types of smart cards. Regarding the processing power, smart cards could be divided into the following categories: memory cards – containing a memory chip only with non-programmable logic, microprocessor's chip card with internal memory, microprocessor's chip card with internal memory including additional PKI capabilities (with additional RSA, 3DES, and RNG (Random Number Generator) coprocessors – called PKI smart cards). Regarding the physical contacts of the chip, smart cards could be: contact cards – chip with electrical interface, contactless cards – chip with electromagnetic interface, combo cards – with two chips: one with contact and one with contactless interface, and dual interface chip cards - with chip that have two interfaces: electrical and electromagnetic. Regarding the chip operating system, smart cards could be: proprietary operating system smart cards (with a single or multiple (MULTOS) application capabilities), and JAVA smart cards. Also, smart cards could be divided regarding the power of the implemented microprocessors (8-bit, 16-bit or 32-bit) or regarding the amount of the available memory (EEPROM) (16 KB – 128 KB).

Today's PKI smart cards are still mostly based on 8-bit microprocessors (based on the well-known Intel 80C51 microcontroller) with smaller amounts of 16-bit and 32-bit microprocessors. However, it is clear that 8-bit smart card microprocessors will be forgotten very soon and that the market will move toward more powerfull microprocessors. Also, there is a clear move toward JAVA and Multos multiapplicative smart cards instead of previously used proprietary OS – one application smart cards. JAVA and Multos smart cards enable both multiapplication and easier customization of the existing applications. Smart cards used in PKI systems provide a secure and portable way to store the private cryptographic keys and corresponding X.509 digital certificates. The smart card enhances the PKI security by enforcing an extra authentication layer at the end-user level. This extra authentication layer, coupled with the fact that cryptographic keys generated on the card never leave the card, adds an important additional security layer which increases the security of the overall solution. Actually, PKI smart cards with two X.509 digital certificates and two private asymmetric keys stored (for digital envelope retrieval/identification and for digital signature), where signature keypair is generated on the card, represents the most up-to-date security solution for large scale users which provides all four mentioned main security functions in modern information systems: authentication (X.509 digital certificate), data integrity (digital signature), non-repudiation (digital signature by asymmetric key generated and stored on the card), and confidentiality (based on asymmetric private key for digital envelope retrieval). Also, it should be emphasized that today mostly smart cards are certified according to the EAL4+ certification (certification includes: smart card (chip), chip operating system and PKI application on the card) which is a necessary condition for Secure Signature Creation Devices (SSCD) according to EU Electronic Signature legislation.

5 e-Healthcare Security Mechanisms

This Section deals with the basics of security mechanisms in e-healthcare systems. Key players in healthcare systems are: medical organizations (hospitals, clinics, pharmaceutical organizations), insurance organizations, healthcare professionals (doctors, physicians, nurses, pharmacists, etc.), and patients – end users. Most modern healthcare systems are information systems based on TCP/IP computer networks and they work fast move toward the electronic business in healthcare industry – electronic healthcare (e-healthcare). In this environment, security mechanisms for e-business must be implemented with necessary adaptation to the healthcare environments. There are a lot of technical and security issues for these systems that include, between the others: electronic patient record or electronic health record (EHR) must be fully private, central database of patient electronic records must be enabled for use from all players (medical organizations, professionals, insurance, patients), privacy protection of the patient records, secure communications between all players in the system, electronic order entry, enabling mobile healthcare, HIPAA compliance [11], etc. Thus, security mechanisms that are necessary to be implemented in these e-healthcare systems are: strong user authentication procedure, digital signature technology, confidentiality protection of data in the system on the application, transport and network layers, privacy protection of the patient personal data, strong protection of the central healthcare database based on multiple firewall architecture, and PKI systems, which issue X.509 digital certificates for all users of the system (healthcare professionals and patients) - digital identities (IDs) for the users.

5.1 Strong User Authentication

There are several types of user authentication procedures that could be based on the following components: Username/Password – PIN code – something that user know, hardware token – something that user has, and biometric characteristic (e.g. fingerprint) – something that user is.

Regarding the above components there are several types of authentication procedures which combine some of them, such as:

- Username/password based authentication – weak authentication,
- Username + dynamic password (one-time password) obtained by appropriate hardware token – stronger than previous one but not in the class of strong user authentication procedures,
- Username + dynamic password obtained by appropriate hardware token + challenge-response procedure – strong user authentication procedure,
- Username/password or PIN code + PKI smart card + bilateral challenge response procedure based on PKI X.509 digital certificate and asymmetrical cryptographic techniques – strong user authentication procedure (stronger than the previous one),
- Username/password or PIN code + PKI smart card + biometric characteristic checking + bilateral challenge response procedure based on PKI X.509 digital certificate and asymmetrical cryptographic techniques – the strongest user authentication procedure.

In other words, the class of strong user authentication procedures consists of the two or more component authentication procedures and a use of the bilateral challenge-response procedure.

Modern e-healthcare information systems must be based on the strong authentication procedure.

5.2 Digital Signature Technology

It should be pointed again that the state-of-the-art solution for all the mentioned three security functions: authenticity, data integrity and non-repudiation, could be today achieved only by use of the PKI smart cards with digital signature generation on the card with signature private key generated on the card and never leaves the card. In the modern e-healthcare systems, healthcare professionals, as well as the patients, should use the smart cards as the hardware tokens for creating digital signature. More and more EU countries demands that the signature made with e-healthcare PKI card must be qualified electronic signature according to the EU Electronic Signature Legislation.

5.3 Confidentiality Protection

Since data that is transmitted through the particular e-healthcare system contain very sensitive, often personal patient's data, its confidentiality must be fully preserved. This should be done by using digital envelope technology based on symmetrical and asymmetrical cryptographic techniques and PKCS#7 file format. This technology is based on digital certificate, symmetrical algorithms for encryption of data and asymmetrical algorithms for protection of symmetric key which is sent together with encrypted data. This technology is mainly used for application level protection and it mainly represents the protection from internal attacks to the system.

However, the transport and network level protections should be also used in the system in order to prevent external attacks. Namely, besides the application level protection based on digital signature and envelope technologies, that are based on the end users smart cards, the transport level (SSL) and network level (IPSec/VPN) security mechanisms should be used. In other words, it is strongly recommended that security mechanisms on more than one level are to be used. In this sense, the application level protection should be mandatory in combination with one or two additional level protections, transport or network based, depending of system characteristics, type of application and connections, required system throughput, other technical requirements, etc.

5.4 Privacy Protection of the Personal Patient Data

It is already emphasized that the main issue of the e-healthcare system is to protect privacy of the patient personal medical data – now processed and stored in the electronic form. This means that the data should be protected on the whole processing path in the system, i.e. from the medical professional workstation to the central database. In other words, unauthorized access to the data should be protected in the entire e-healthcare system.

5.5 Protection of the Central Database

As we already mentioned, the central e-healthcare database should be maximally protected from internal and external attacks. Normally, the new designed e-healthcare application should be multi-tier (three or more tiers) applications that could be WEB based or client-server based applications. Modern trends move toward web based applications with pretty thin clients [12]. Client's part of the application should prepare data (e.g. offline) which includes applying of the appropriate security mechanisms (digital signature and digital envelope based on smart cards) and send this data (in online mode) through web browser interface to the web site (e.g. e-healthcare WEB portal) on the central (or other) location. Before sending data to the web portal, the enduser must authenticate himself on the web portal by using the strong authentication PKI procedure based on smart card and digital certificates. Modern trends move toward establishing web portal for different medical organizations (or for central point) with single-sign-on capabilities of end-user authentication. This means that administration of the valid users should be centralized and that users cannot do any action if they are not strongly authenticated before through adequate single-sign-on function.

A possible example of Generic model of the central e-healthcare site is proposed on Fig. 3. In this model, we could see four different parts of the central medical site:

- External part for accessing the system which is on the one side of Firewall (connected to one particular interface of the firewall),
- DMZ – DeMilitarized Zone with some general purpose servers, such as: mail, ftp, http, as well as with the WEB e-healthcare portal,

- Internal zone with different applications servers – middle tiers of different multitier medical applications, and
- Most secure internal zone where the most sensitive parts of the system (e.g. central EHR database) are located.

In this model, multiple firewall architecture is applied. Between the external part and one or more DMZs, the commercial firewall (mostly based on packet filtering techniques) could be applied. However, for the protection of the most sensitive part of the system – the central database, some firewall of the application proxy level gateway type should be applied, e.g. [13], [14]. The best protection will be achieved if this second firewall will be the proprietary made firewall – and not a commercial one.

5.6 e-Healthcare PKI Systems

To enable application of the all previously mentioned security mechanisms, the appropriate PKI system must be established in advance. The e-healthcare PKI system has the following characteristics: it is based on X.509 digital certificates as digital IDs for valid users of the system, central point of the PKI system is Certification Authority (CA), and CA issues digital certificates on smart cards (patient and healthcare professionals' smart cards). In integrated e-healthcare medical systems, the CA could be truly centralized, centralized with hierarchical CA structure or decentralized. In the truly centralized system, there is only one CA (most often at some state healthcare authority) who issues all digital certificates for all kind of end-users (patients, healthcare professionals, insurance employees, etc.). In this system, individual medical organizations are not independent in defining its own PKI policy but must conform to the global healthcare PKI policy.

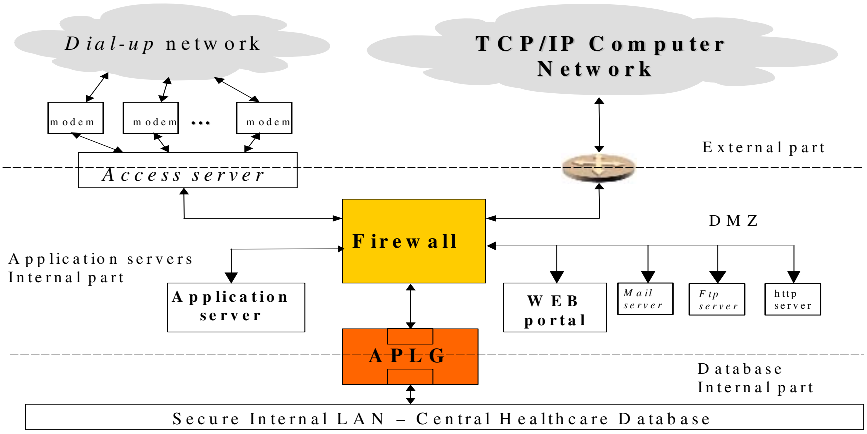


Fig. 3. A generic model of the central (or other) healthcare site

In the centralized system with hierarchical CA structure, there is a root CA at the healthcare authority and several levels of intermediate CAs. These intermediate CAs will be for different kind of end users and for individual medical organizations (e.g.

some clinic that has a large information system). The advantage of this architecture is that medical organizations are independent in creation of their own PKI subsystems and that each of the users group is under the one certificate management system. However, since all of the intermediate CAs is under the one centralized root CA, compatibility of communications between parties belonging to different intermediate CAs is completely achieved. Decentralized CA structure could be used in the case that all medical organizations have their own PKI subsystem, independent of some centralized authority. This system provides independency but there is an issue regarding the communications between parties that does not belong to the same CA. In this case, this could be only achieved by applying the cross-certification procedure between the CAs. The trend is that modern e-healthcare information systems are based on the centralized PKI system with hierarchical security infrastructure and digital certificates stored on smart cards.

6 Conclusions

In this paper, the modern computer security systems are analyzed and their possible application in e-healthcare systems is emphasized. It is concluded that only multilayered security architecture could cope with potential internal and external attacks to the modern computer networks and e-healthcare systems. The most frequently used security mechanisms on the application, transport and network layers are analyzed. It is concluded that more than one layer should be covered by the appropriate security mechanisms in order to achieve high quality cryptography protection of the e-healthcare system. It is also concluded that, between many specific conditions in the e-healthcare systems, application of security mechanisms should be considered on the client side, communication side and central database side, and that, in each of the sides, appropriate security measures should be applied. Central points of e-healthcare systems are smart cards for end users (citizens, healthcare professionals, etc.) that could be used for applying digital signature and digital envelope technology and the central PKI system. Smart cards must be used by doctors and other healthcare professionals. For patients, main point is that data should be protected from unauthorized use (privacy protection) and thus it is not mandatory to use PKI smart cards for patients. However, in order to enable some future advance features, as well as the qualified signature, it is strongly recommended that PKI smart cards be used for patients too.

References

1. Oppliger, R.: Internet and Intranet Security, Artech House, (1998), ISBN 0-89006-829.
2. Ford, W., Baum, M.S.: Secure Electronic Commerce: Building the Infrastructure for Digital Signatures and Encryption, Second Edition, Prentice Hall PTR, Upper Saddle River, NJ 07458, (2001).
3. Marković, M.: Cryptographic Techniques and Security Protocols in Modern TCP/IP Computer Networks, Short-Tutorial, in Proc. of ICEST 2002, Oct. 1-4, (2002).

4. Schneier, B.: Applied Cryptography, Second Edition, Protocols, Algorithms and Source Code in C, John Wiley & Sons, Inc., New York, Chichester, Brisbane, Toronto, Singapore, (1996).
5. RSA Laboratories: PKCS standards.
6. Marković, M., Unkašević, T., Djorđević, G.: RSA algorithm optimization on assembler of TI TMS320C54x signal processors, in Proc. of EUSIPCO 2002, Toulouse, France, Sept. 3-6, (2002).
7. Unkašević, T., Marković, M., Djorđević, G.: Optimization of RSA algorithm implementation on TI TMS320C54x signal processors based on a modified Karatsuba-Offman's algorithm, in Proc. of ECMCS'2001, 11-13 September, Budapest, (2001).
8. Djorđević, G., Unkašević, T., Marković, M.: Optimization of modular reduction procedure in RSA algorithm implementation on assembler of TMS320C54x signal processors, DSP 2002, July, Santorini, Greece, (2002).
9. Marković, M., Đorđević, G., Unkašević, T.: On Optimizing RSA Algorithm Implementation on Signal Processor Regarding Asymmetric Private Key Length, in Proc. of WISP 2003, Budapest, Sept. 2003, (2003), 73-77.
10. Marković, M., Savić, Z., Obrenović, Ž., Nikolić, A.: A PC Cryptographic Coprocessor Based on TI Signal Processor and Smart Card System, Communications and Multimedia Security Issues of the New Century, R. Steinmetz, J. Dittman, M. Steinebach, (Eds.), Kluwer Ac. Publishers, (2001), 383.393.
11. Healthcare Insurance Portability and Accountability Act: HIPAA Requirements for Technical Security, Services and Mechanisms, (1996).
12. Oppliger, R.: Security Technologies for the World Wide Web, Artech House, Boston, London. (2000).
13. Savić, Z., Nikolić, A., Marković, M.: Cryptographic proxy gateways in securing TCP/IP computer networks, In Proc. of Information Security Solution Europe, ISSE 2001, London, UK, (2001).
14. Savić, Z., Marković, M.: Development of Secure Web Financial Services in Serbia, in Proc. of ISSE 2003, Vienna, Austria, October 7-10, (2003).

IPUMS-International High Precision Population Census Microdata Samples: Balancing the Privacy-Quality Tradeoff by Means of Restricted Access Extracts

Robert McCaa, Steven Ruggles, Michael Davern, Tami Swenson,
and Krishna Mohan Palipudi

Minnesota Population Center, 50 Willey Hall
Minneapolis MN 55455 USA
rmccaa@umn.edu

Abstract. A breakthrough in the tradeoff between privacy and data quality has been achieved for restricted access to population census microdata samples. The IPUMS-International website, as of June 2006, offers integrated microdata for 47 censuses, totaling more than 140 million person records, with 13 countries represented. Over the next four years, the global collaboratory led by the Minnesota Population Center, with major funding by the United States National Science Foundation and the National Institutes of Health, will disseminate samples for more than 100 additional censuses. The statistical authorities of more than 50 countries have already entrusted microdata to the project under a uniform memorandum of understanding which permits researchers to obtain custom extracts without charge and to analyze the microdata using their own hardware and software. This paper describes the disclosure control methods used by the IPUMS initiative to protect privacy and to provide access to high precision census microdata samples.

Keywords: Census microdata samples, data privacy, data quality, IPUMS-International.

1 Introduction

In 1983, the legendary Charles M. Cawley offered the alumni association of his alma mater, Georgetown University, a deal. In exchange for its endorsement and a list of members, his fledgling credit card company, MNBA, would pay a percentage of revenues to the association. The offer was accepted and MNBA—by extending the affinity credit card offer to organizations with responsible, affluent members from the Association of Trial Lawyers of America to the Sierra Club—quickly established itself as the fastest growing, most profitable credit card company in the United States. Cawley became a billionaire. Now every successful credit card company in the world markets affinity cards.

The IPUMS project seeks neither profits nor popularity. Ours is a wholly academic initiative, but we target an affinity group, a "restricted class of individuals" [1] consisting of academic and policy researchers, who have great need to use population census microdata, but pose a vanishingly small risk of misuse.

J. Domingo-Ferrer and L. Franconi (Eds.): PSD 2006, LNCS 4302, pp. 375 – 382, 2006.

Where much disclosure control research on the privacy-quality tradeoff is focused on either “public access” at one extreme or “safe-harbor” at the other [2], the IPUMS-International initiative adopts a third way, the “trusted user” approach [3]. Access is denied to approximately one-third of those who complete the electronic application form. Four years after dissemination began in May 2002, fewer than one thousand researchers have been granted access to IPUMS-International census microdata.

We restrict access to researchers who have a defined need to use the data and who not only agree to abide by the rigorous conditions of use license but also bind their institutions as enforcing agents. With, on the one hand, the assistance of our statistical agency partners, as stipulated in the project memorandum of understanding, and, on the other, the conditions of use license, misuse will lead to punishment not only for the individual but also for the individual’s institution. Indeed, in contrast to the record of commercial companies and government agencies, where there are frequent accounts of misuse of microdata for disclosing information about individuals, there is not a single, specific allegation of misuse of population census microdata in more than four decades of use by academic researchers. By rigorously policing access, we expect to extend this unblemished record of responsible scholarly use.

2 The Case for High Precision Samples: The USA Experience

In recent years, scholars working with United States census microdata have come to rely on high-precision samples. Beginning with the 1980 census, the Census Bureau has released five-percent samples as well as the one-percent samples. The five-percent samples for the United States in 1980, 1990, and 2000 include between 12 million and 14 million individuals in each year.

The Census Bureau anticipated that the 1980 five-percent sample would be used mainly for state and local policy analysis; at the time the sample was created, it was prohibitively expensive for most researchers to process the entire set of five-percent data. By the end of the 1980s, however, data processing costs had declined dramatically and were no longer a critical constraint for researchers at major institutions. Social scientists soon developed research strategies that capitalized on the availability of very large census microdata files. Swicegood et al. [4] published the first article in *Demography* that used a five-percent national sample, an analysis of language use and fertility in the Mexican-origin population. Later that year, Odland and Ellis [5] published a second *Demography* article using the large 1980 file, a study of household size and regional outmigration rates between 1975 and 1980.

From that time on, the use of high-precision census microdata files expanded rapidly. The cost of computing declined dramatically during the first half of the 1990s with the advent of inexpensive UNIX workstations. Moreover, during the past several years the performance of Windows-based desktop computers has improved to the point that a machine costing less than $1,000 is now easily capable of processing the five-percent samples of 1980, 1990 and 2000. Since 1996, the on-line data dissemination systems developed at Minnesota and elsewhere have provided easy access to large microdata extracts. Accordingly, the largest census microdata files—once available to few researchers at great expense—are now accessible, at no cost, to virtually all social scientists and policy analysts worldwide.

Increasingly, studies that use census microdata from 1980, 1990 or 2000 have turned to the five-percent files. Since 1990, 81 percent of *Demography* articles based on recent census microdata have used the high-precision samples.[1] Most of these analyses depend on information for small population subgroups, ranging from same-sex couples to the grandchildren of immigrants. In many instances, the large samples permit the use of innovative methods; to take just one example, these files have allowed demographers to carry out multi-level contextual analyses by making it feasible to assess the characteristics of small geographic areas.

The five-percent samples of the 1980, 1990 and 2000 censuses have now become the most widely used data source in the pages of *Demography.*, as we learned from a analysis of the journal's pages in 2002. At that time, even though the United States had abundant high-quality survey data and the most recent census samples were over a decade old, high-precision census microdata files were used by a quarter of the articles on the United States that appeared in *Demography* in 2000 and 2001. In that period, the large samples were used twice as often as the next most popular data source. Clearly, the high-precision samples of the 1980 and 1990 censuses had become an indispensable component of American social science infrastructure. In 2003, with the addition of a five percent sample from the 2000 census, use skyrocketed.

It is impossible to determine an optimal size for a general-purpose sample. The number of cases needed to analyze a population subgroup depends on desired precision, type of subgroup, type of analysis, and population heterogeneity. If high precision estimates are required, many thousands of cases of the subgroup of interest may be necessary. Frequently, the relevant individuals for analysis are a small subset of the sample population. Multilevel analyses of the effects of local context on individual behavior are especially demanding since they often require data tabulated for small geographic units. The experience of the U.S. demonstrates that very large census microdata samples are among the most powerful tools available for economic and demographic analysis. As such samples become available for other countries around the world, they are becoming key components of social science and policy infrastructure.

3 The IPUMS Approach: High Precision Samples with Implicit Stratification

An important technique used to protect confidentiality of census microdata is to draw a high precision sample from all the census microdata records and then, in addition to the disclosure controls discussed below in sections 4 and 5, suppress from the sampled records all identifying information (names, addresses, and low-level geographical details). High precision samples preserve the ability to work with a large amount of microdata making it harder to identify any one person in the sample data file. In drawing high precision samples it is also important to think about

[1] This percentage excludes eight articles that did not specify sample precision.

efficient methods. By using stratification to draw a high precision sample, gains in efficiency are possible [6], [7]. To the extent the strata used to draw a high precision sample are associated with the variables of interest (e.g., orphanhood, poverty, unemployment, etc.), the resulting estimates of these variables will have lower standard errors than what would have resulted had a simple random sample of records been drawn from the complete census data [6], [7].

One of the most important stratifying variables in survey research and in drawing high precision census microdata samples is geography. Geography is related to a great number of variables researchers are interested in studying and therefore increases the efficiency of stratified samples. Many of the IPUMS-International samples capitalize on *implicit* geographic stratification. The raw census files used to create IPUMS samples are typically geographically organized within districts. Systematic random samples of the censuses capitalize on this low-level geographic sorting. By ensuring a representative geographic distribution of sampled cases, they are equivalent to extremely fine geographic stratification with proportional weighting. Since many economic and demographic characteristics are highly correlated with geographic location, this implicit stratification yields substantially greater precision than would a simple random sample of households. As part of the IPUMS project, we are developing stratification variables that allow researchers to make reliable variance estimates from implicitly stratified samples.

Almost all the statistical agency partners of the IPUMS project have endorsed the use of implicitly stratified samples of households (see Table 1, "sample design" column). Twenty-six countries (identified by "*" in Table 1) have provided complete sets of census microdata to facilitate the drawing of implicitly stratified samples by the project. In Europe, almost all the statistical agencies have drawn new samples using IPUMS specifications. IPUMS sample densities, as can be seen in Table 1, typically range between 5 and 10%. Lower densities are provided by countries where privacy matters are a greater issue than quality (Netherlands, United Kingdom) or, as in the case of 1960 round of censuses, where only low precision samples survive.

4 IPUMS-International Access Disclosure Controls

Access to the IPUMS-International database is governed, on the one hand, by the letter of understanding endorsed by the University and the National Statistical Authority, and, on the other by the license agreement between the University, the researcher, and the researcher's institution. Both are subject to amendment and enhancement as new methods are suggested. The letter of understanding grants the right to the university to disseminate microdata extracts electronically for teaching and research purposes via the project webpage: https://www.ipums.org/international, according to the authorization procedures stated in the agreement. Data may not be used for commercial purposes. Strict confidentiality of persons, households and other entities must be maintained. Alleging that a person or other entity has been identified is prohibited. The University is charged with assuring that users will guard against access to the microdata by unauthorized individuals.

The fact that IPUMS-International distributes microdata electronically as custom extracts, tailored as to country(ies), census year(s), subpopulation(s), and variables, according to the individual needs of the researcher, provides additional incentives for jealously guarding extracts. Since complete datasets are not distributed on CD or other medium, the inclination to share data with unauthorized individuals is greatly reduced, if not completely eliminated.

The electronic application form is designed to ascertain the bona fides of the applicant as well as the appropriateness of the microdata for the proposed research. A stern warning is issued against fraudulent applications, and checks are implemented to verify the identity and affiliations of the applicant (see the project home page "Apply for Access"). To confirm that the researcher understands the sensitivity of guarding the privacy of individuals, the application requests the name of the Human Subjects Protections Committee, Institutional Review Board, or similar office at the applicant's institution. A critical consideration in determining access is the proposed research. The statement must identify the data to be used and the purpose. Many applicants are denied access for failing to demonstrate that microdata are needed to address the proposed research or instructional plan. Finally the researcher must agree to seven restrictions on use: no redistribution, scholarly use only, prohibition on commercial use, strict rules of confidentiality, data security, appropriate citation, and notification of errors in the data. Approval is granted for a period of one year and may be renewed. Access to the microdata is password controlled. Remote data access is not offered. While this method might allow access to higher density, virgin microdata, our memorandum of understanding with the national statistical agencies does not authorize this form of access.

5 Technical Disclosure Controls

Where the statistical agency entrusts the anonymization procedures to the IPUMS project, we impose additional technical privacy protections. Technical controls are implemented on a subjective, ad-hoc basis as negotiated with each country for each census. Contemporary microdata, say from a census taken less than ten years ago, require more technical disclosure controls than older, historical data.

The most important technical control is the suppression of records by subsampling. All the values in the records outside the sample are suppressed. Second, is the suppression of names and geographical detail, such as place of birth or residence. Each statistical authority balances the trade-off by instructing the IPUMS project as to the minimum threshold for identifiable geographical units for the most recent census. In the case of many African and Latin American countries, the threshold is commonly set at 20,000 inhabitants in the latest census. Others place it as high as 100,000 (United States) or in the most extreme case (Netherlands) all administrative geography is suppressed. We are gratified that in some cases our statistical agency partners have reconsidered earlier decisions, offering higher precision samples (Mexico 1990 increased from one to ten percent) and greater detail. In the case of Colombia, the geographical threshold, initially set at 100,000, was reduced to 20,000 after Colombian geographers vigorously registered their dissatisfaction. The Colombian statistical agency not only reduced the threshold, but also harmonized the identifiers so that all the census microdata samples for Colombia could be disseminated with a single set of geographical codes.

Table 1. IPUMS-International: 160 microdatasets entrusted by country, subsample precision and design For current data availability, see: https://www.ipums.org/international

Datasets entrusted by subsample precision			**Country**	**Sub sample design**	**2000s**	**1990s**	**1980s**	**1970s**	**1960s**
10%	**~5%**	**<=4%**							
Release 1, May 2003 (28 datasets)									
5			**Brazil**	IPUMS	**2001**	**1991**	**1980**	**1970**	**1960**
		1	**China (only '82 'til now)**		**2000**	**1990**	**1982**		1964
3		1	***Colombia**	IPUMS		**1993**	**1985**	**1973**	**1964**
	5		**France ('99 in preparation)**	IPUMS	**1999**	**1990**	**1982**	**1975**	**1968, 2**
	2		**Kenya ('79 & '69 in process)**	IPUMS	**1999**	**1989**	**1979**	**1969**	
2		2	**Mexico ('80 in recovery)**	IPUMS	**2000**	**1990**	1980	**1970**	**1960**
	5		**United States**		**2000**	**1990**	**1980**	**1970**	**1960**
	2		**Vietnam**	IPUMS		**1999**	**1989**	1979	
Release 2, June 2006 (19 datasets)									
4		1	***Chile**	IPUMS	**2002**	**1992**	**1982**	**1970**	**1960**
3	1		***Costa Rica**	IPUMS	**2000**		**1984**	**1973**	**1963**
4		1	***Ecuador**	IPUMS	**2001**	**1990**	**1982**	**1974**	**1962**
2			**South Africa**		**2001**	**1996, 1**	**1985, 0**	**1970**	1960
3			***Venezuela**	IPUMS	**2001**	**1990**	**1981**	**1971**	**1961**
Europe (27 datasets)									
4			**Austria**	IPUMS	**2001**	**1991**	**1981**	**1971**	1961
1			**Belarus**	IPUMS		**1999**	1989	1979	1970
			Bulgaria (in process)		**2001**	**1992**	**1985**	1975	1965
	2		**Czech Republic**	IPUMS	**2001**	**1991**	**1980**	**1970**	1961
			Germany (in process)		**2001m**	**1991m**	**1987, 1**	**1970, 1**	1961
4			**Greece**	IPUMS	**2001**	**1991**	**1981**	**1971**	1961
	4		**Hungary**	IPUMS	**2001**	**1990**	**1980**	**1970**	
			Italy (in process)		**2001**	**1991**	**1981**	1971	1961
		3	**Netherlands**		**2001m**			**1971**	**1960**
			Poland (negotiating)		**2001**		**1988**	**1978, 0**	1960
	3		**Portugal**	IPUMS	**2001**	**1991**	**1981**	1970	1960
2			**Romania ('77 in recovery)**	IPUMS	**2001**	**1992**		**1977**	1965
			Russia (negotiating)		**2002**		**1989**	1979	1970
			Slovenia		**2001**	**1991**	**1981**		
	3		**Spain**	IPUMS	**2001**	**1991**	**1981**	1970	1960
			Switzerland (negotiating)		**2000**	**1990**	**1980**	**1970**	1960
			Turkey (in process)		**2000**	**1990**	1980, 5	**1970, 5**	1960, 5
		1	**United Kingdom (in process)**		**2001**	**1991**	**1981**	**1971**	**1961**
North America and the Caribbean (27 datasets)									
		3	**Canada**		**2001**	**1991**, 6	**1981**, 6	**1971**, 6	1961, 6
1	1	2	***Dominican Republic**	IPUMS	**2003**	**1993**	**1981**	**1970**	**1960**
1			***El Salvador**	IPUMS		**1992**		1971	1961
2		3	***Guatemala**	IPUMS	**2002**	**1994**	**1981**	**1973**	**1964**
3		1	***Honduras**	IPUMS	**2000**		**1988**	**1974**	**1961**
1			***Nicaragua**	IPUMS	**2005**	**1995**		**1971**	1963
5			***Panama**	IPUMS	**2000**	**1990**	**1980**	**1970**	**1960**
	4		**Puerto Rico**		**2000**	**1990**	**1980**	**1970**	1960
South America (17 datasets)									
4			**Argentina**	IPUMS	**2001**	**1991**	**1980**	**1970**	1960
3			***Bolivia**	IPUMS	**2001**	**1992**		**1976**	
4		1	***Paraguay**	IPUMS	**2002**	**1992**	**1982**	**1972**	**1962**
1			***Peru**	IPUMS		**1993**	**1981**	1972	1961
4			***Uruguay**	IPUMS		**1996**	**1985**	**1975**	**1963**

Table 1. *(continued)*

Datasets entrusted by subsample precision									
10%	**~5%**	**<=4%**	**Country**	**Sub sample design**	**2000s**	**1990s**	**1980s**	**1970s**	**1960s**
Africa (17 datasets)									
2			**Egypt**	IPUMS		**1996**	**1986**	1976	1964
2			***Guinea, Conakry**	IPUMS		**1996**	**1983**		1960
			Lesotho (in process)			**1996**	**1986**	**1976**	1966
1			***Madagascar**	IPUMS		**1993**			
2			***Malawi**	IPUMS		**1997**	**1987**	**1977**	1967
3			***Mali**	IPUMS		**1998**	**1987**	**1976**	
2			***Rwanda**	IPUMS	**2002**	**1991**			
3			***Sudan**	IPUMS		**1993**	**1983**	**1973**	
2			***Uganda**	IPUMS	**2002**	**1991**	1980		1969
Asia and Oceania (25 datasets)									
1			**Armenia**	IPUMS	**2001**		1989	1979	1970
			Bangladesh (in process)		**2001**	**1991**	**1981**	1974	1961
1			**Cambodia**	IPUMS		**1998**			1962
3			***Fiji Islands**	IPUMS		**1996**	**1986**	1976	**1966**
			Indonesia (in process)		**2000**	**1990**	**1980**	**1971**	1961
1			***Iraq**	IPUMS		**1997**	1987	1977	1967
4			**Israel**	IPUMS		**1995**	**1983**	**1972**	**1961**,7
		4	**Malaysia**		**2000**	**1991**	**1980**	**1970**	1960
1			***Mongolia**	IPUMS	**2000**		1989	1979	1970
3			***Pakistan**	IPUMS		**1998**	**1981**	**1973**	1961
1			**Palestinian Authority**	IPUMS		**1997**			
3		2	***Philippines**	IPUMS	**2000**	**1990**	**1980**	**1970**	**1960**
1			**Turkmenistan**	IPUMS		**1995**	1989	1979	1970

Note: **bold country** = Agreement signed between University of Minnesota and National Statistical Authority
Year = census; **Bold year** = microdata survive; * = 100% microdata entrusted to IPUMS; m = microcensus
IPUMS systematic subsample design for private households: every n^{th} household stratified by enumeration district.

Additional protection is provided by randomly ordering the records and swapping the geographical identifiers of an undisclosed number of households. This means that no one can state with certainty that an individual or household has been identified.

In consultation with the national statistical office, some variables may be top-coded, others may be subjected to global recoding, deletion of digits for hierarchical variables (occupation, industry, geography), or the suppression of a variable entirely. Decisions are made in consultation with the corresponding national statistical authority. Sensitive variables, if any, may be suppressed entirely at the request of the statistical agency. Weight variables are usually not an issue because most of the samples are implicitly stratified with a single weight. We do not resort to either microaggregation or Post Randomization (PRAM) methods.

6 Countering Fear, Hysteria and Paranoia with Reason

Privacy rights and statistical confidentiality of data are severely threatened by government, commercial firms, and individuals—but the threat to population census microdata is virtually nil. Fear, hysteria and paranoia are incited among official statisticians by the widespread circulation of a "pizza commercial" developed by an

American civil liberties advocacy group [8] and advertisements offering private details of individuals and entities for a price. What is striking is that none involve population census microdata. Indeed, there is no market—black, grey, gold or otherwise—for anonymized census microdata samples for the purpose of identifying individuals or linking to other data sources. Even in the United States, at a moment of shocking violations of individual rights by government agencies, there is not one allegation of access to census microdata by the Homeland Security Agency or other government agencies. The reason is obvious. Population census microdata samples, per se, do not contain sensitive or valuable political or commercial information, and without personal identifiers, statistical linkage is useless due to the high proportion of false positives [9].

7 Conclusion

The goal of IPUMS is to restore balance to the privacy-quality tradeoff by providing high precision, anonymized samples to a restricted class of researchers. In the IPUMS datasets identification is impossible for the vast majority of persons and positive identification is always impossible. Given the wealth of information readily available from private sources in most countries, it would be foolhardy to turn to census microdata to attempt to uncover imprecise and outdated information about a particular individual. We invite academics who need census microdata for research purposes to examine the offerings at the IPUMS website.

Acknowledgements. Funded in part by the National Science Foundation of the United States, Grant No. SES0433654.

References

1. Willenborg, L., de Waal, T.: Elements of Disclosure Control. New York: Springer-Verlag (2001)
2. Willenborg, L., de Waal, T.: Statistical Disclosure Control in Practice. New York: Springer-Verlag (1996)
3. McCaa, R., Esteve, A.: IPUMS-Europe: Confidentiality measures for licensing and disseminating restricted-access census microdata extracts to academic users. In Monographs of official statistics: Work session on statistical data confidentiality. Luxembourg: Office for Official Publications of the European Communities, (2006) 37-46.
4. Swicegood, G., Bean, F.D., Stephen, E.H., Opitzm, W.: Language Usage and Fertility in the Mexican-Origin Population of the United States. Demography. 25 (1988) 17–33
5. Odland, J., Ellis, M.: Household Organization and the Interregional Variation of Out-migration Rates. Demography. 25 (1988) 567-579
6. Kish, L.: Weighting for Unequal P_i. Journal of Official Statistics. 8 (1992) 183-200
7. Kish, L.: Survey Sampling, Wiley Classics Library Edition. New York: Wiley and Sons (1995)
8. American Civil Liberties Union (ACLU). Surveillance Campaign. (2005) Available online at http://www.aclu.org/pizza/
9. Dale, A., Elliot, M.: Proposals for 2001 SARS: An assessment of disclosure risk. Journal of the Royal Statistical Society. Series A. 164, part 3 (2001) 427-447

Author Index