Josep Domingo-Ferrer
Emmanouil Magkos (Eds.)

LNCS 6344

Privacy in Statistical Databases

UNESCO Chair in Data Privacy
International Conference, PSD 2010
Corfu, Greece, September 2010, Proceedings

Springer

Lecture Notes in Computer Science 6344

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

Josep Domingo-Ferrer
Emmanouil Magkos (Eds.)

Privacy in Statistical Databases

UNESCO Chair in Data Privacy
International Conference, PSD 2010
Corfu, Greece, September 22-24, 2010
Proceedings

Springer

Volume Editors

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

Emmanouil Magkos
Ionian University, Department of Informatics
Plateia Tsirigoti 7, 49100 Corfu, Greece
E-mail: emagos@ionio.gr

Library of Congress Control Number: 2010934215

CR Subject Classification (1998): H.2, D.4.6, K.6.5, C.2, E.3, E.1

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

ISSN 0302-9743
ISBN-10 3-642-15837-4 Springer Berlin Heidelberg New York
ISBN-13 978-3-642-15837-7 Springer Berlin Heidelberg New York

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India
Printed on acid-free paper 06/3180

Preface

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

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

"Privacy in Statistical Databases 2010" (PSD 2010) was held under the sponsorship of the UNESCO Chair in Data Privacy, which has provided a stable umbrella for the PSD biennial conference series since PSD 2008, held in Istanbul. Previous PSD conferences were PSD 2006, the final conference of the Eurostat-funded CENEX-SDC project, held in Rome in 2006, and PSD 2004, the final conference of the EU FP5 CASC project (IST-2000-25069), held in Barcelona in 2004. Proceedings of PSD 2008, PSD 2006 and PSD 2004 were published by Springer in LNCS 5262, LNCS 4302 and LNCS 3050, respectively. The four PSD conferences held so far are a follow-up of a series of high-quality technical conferences on SDC, which started twelve years ago with "Statistical Data Protection-SDP'98", held in Lisbon in 1998 and with proceedings published by OPOCE, and continued with the AMRADS project SDC Workshop, held in Luxemburg in 2001 and with proceedings published by Springer in LNCS 2316.

The PSD 2010 Program Committee accepted for publication in this volume 25 papers out of 37 full submissions. Furthermore, 5 additional submissions were received which were reviewed for short presentation at the conference and inclusion in the companion CD proceedings. Papers came from 16 different countries from 4 different continents. Each submitted paper received at least two reviews. The revised versions of the 25 accepted papers in this volume are a fine blend of contributions from official statistics and computer science. Topics covered include tabular data protection, microdata protection, synthetic data, differential privacy, on-line databases and remote access, privacy-preserving protocols, and legal issues.

We are indebted to many people. First, to the Government of Catalonia for financial support to the UNESCO Chair in Data Privacy, which enabled the latter to sponsor PSD 2010. Also, to the Organization Committee for making the conference possible and especially to Jesús Manjón, who helped prepare these proceedings. In evaluating the papers we were assisted by the Program Committee and the following external reviewers: Anne-Sophie Charest, Jörg Drechsler, Sara Foresti, Flavio Foschi, Gabriel Ghinita and Peter-Paul de Wolf.

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

July 2010

Josep Domingo-Ferrer
Emmanouil Magkos

Organization

Program Committee

John Abowd	Cornell University, USA
Bettina Berendt	Katholieke Universiteit Leuven, Belgium
Elisa Bertino	CERIAS, Purdue University, USA
Jordi Castro	Polytechnical University of Catalonia, Spain
Lawrence Cox	Nat. Center for Health Statistics, USA
Vassilios Chrissikopoulos	Ionian University, Greece
Josep Domingo-Ferrer	Universitat Rovira i Virgili, Catalonia, Spain
Mark Elliot	Manchester University, UK
Stephen Fienberg	Carnegie Mellon University, USA
Luisa Franconi	ISTAT, Italy
Sarah Gießing	Destatis, Germany
Anco Hundepool	Statistics Netherlands, The Netherlands
Julia Lane	National Science Foundation, USA
Emmanouil Magkos	Ionian University, Greece
Bradley Malin	Vanderbilt University, USA
Josep M. Mateo-Sanz	Universitat Rovira i Virgili, Catalonia, Spain
Krish Muralidhar	University of Kentucky, USA
Jean-Marc Museux	EUROSTAT, European Union
Silvia Polettini	University of Naples, Italy
Yosef Rinott	Hebrew University, Israel
Gerd Ronning	University of Tübingen, Germany
Juan José Salazar	University of La Laguna, Spain
Pierangela Samarati	University of Milan, Italy
Yücel Saygın	Sabancı University, Turkey
Eric Schulte-Nordholt	Statistics Netherlands, The Netherlands
Natalie Shlomo	University of Southampton, UK
Vicenç Torra	IIIA-CSIC, Catalonia, Spain
Vassilios Verykios	University of Thessaly, Greece
William E. Winkler	Census Bureau, USA
Laura Zayatz	Census Bureau, USA

Program Chair

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

General Co-chairs

Emmanouil Magkos	Ionian University, Greece
Vassilios Chrissikopoulos	Ionian University, Greece

Organization Committee

Ioannis Karydis	Ionian University, Greece
Jesús Manjón	Universitat Rovira i Virgili, Catalonia, Spain
Alexandros Panaretos	Ionian University, Greece
Glòria Pujol	Universitat Rovira i Virgili, Catalonia, Spain
Spyros Sioutas	Ionian University, Greece

Table of Contents

Tabular Data Protection

Microdata Protection

Synthetic Data

Differential Privacy

On-Line Databases and Remote Access

Privacy-Preserving Protocols

Legal Issues

Privacy Disclosure Analysis and Control for 2D Contingency Tables Containing Inaccurate Data

Bing Liang[1], Kevin Chiew[1], Yingjiu Li[1], and Yanjiang Yang[2]

[1] School of Information Systems
Singapore Management University
80 Stamford Road
Singapore 178902
[2] Institute for Infocomm Research
#21-01 Connexis South Tower, 1 Fusionopolis Way
Singapore 138632

Abstract. The 2D (two-dimensional) contingency tables have been used in many aspects of our daily life. In practice, errors may be incurred when generating or editing such a table hence the data contained by the table could be inaccurate. Even so, it is still possible for a knowledgeable snooper who may have acquired the information of error distributions to decipher some private information from a released table. This paper investigates the estimation of privacy disclosure probability for contingency tables with inaccurate data based on Fréchet bounds and proposes two optimization solutions for the control of privacy disclosure so as to preserve private information. Our estimation of privacy disclosure probability and the optimization solutions are also applicable to error-free tables which can be regarded as a special case where there are no errors. The effectiveness of the solutions is verified by rigorous experiments.

Keywords: contingency table, privacy disclosure, Fréchet bounds.

1 Introduction

Some organizations such as census bureaus and health care agencies collect statistical information about individuals; however, with the responsibility of preserving individual privacy (e.g., salary, treatment frequency of certain disease), they only publish aggregated data of individuals for public service. Typically these data of individuals are presented in the form of 2D contingency tables with non-negative integers as cell values, but only marginal values (i.e., the sum of cell values of a row, a column, or the whole table) are released as aggregated information to the public as shown in Figure 1.[1] The aggregated information is assumed privacy-preserved though, still suffers from the risks of inferential disclosure [4] by which a data snooper may infer some private information about

[1] Section 2 will show what types of privacy disclosure patterns can be found from the table shown in the figure.

J. Domingo-Ferrer and E. Magkos (Eds.): PSD 2010, LNCS 6344, pp. 1–16, 2010.

	T_1	T_2	T_3	T_4	
Bob	(5)	(1)	(0)	(0)	6
Dan	(1)	(1)	(0)	(0)	2
Tim	(3)	(0)	(1)	(1)	5
	9	2	1	1	13

Fig. 1. An example of contingency table (numbers in parentheses are not released)

individual data values. Disclosure of private and sensitive information may compromise individual and organizational privacies, confidentiality, and national interests [23,18]. Therefore, data owners need to carefully assess the potential privacy disclosure and take effective actions to protect the data before they are released to the public.

1.1 Related Work

While various techniques have been developed for limiting the disclosure of private information from these contingency tables[2] (see survey and review in [8,9]), they can be roughly classified into two categories, namely restriction-based and perturbation-based.

The restriction-based techniques put their efforts on limiting the disclosure of private information by adding restrictions on queries [2,25,24]. These restrictions include the number of values aggregated in each query, the common values aggregated in different queries [7], and the rank of a matrix representing an array of answered queries. Some other restriction-based methods, such as partition [22], microaggregation [8], suppression and generalisation [6], and k-anonymity [5], ℓ-diversity [19], and t-closeness [13], are proposed to sanitize data before releasing by imposing restrictions on data structures.

The perturbation-based techniques, on the other hand, provide methods to prevent data privacy from disclosure by adding random noises to source data [3,14,20], query answers [1], or data structures [21]. As pointed out in the early study by our research team [15] however, it has been reported in [10,11] that the original sensitive information can be estimated accurately from the perturbed data; therefore, it is suggested that the perturbed data be carefully examined before releasing to the public.

Closely related are the early study by our research team on protecting contingency tables [16,15,17] based on which this study is continued and carried out by focusing on congtingency tables with inaccurate data. The disclosure control method proposed in [17] can be classified as a restriction-based technique, and is developed based on four types of disclosure patterns that are identified first time based on Fréchet Bounds. There are other two new disclosure patterns identified in [15] for 2D dynamic tables, together with disclosure control methods developed to prevent data snoopers from inferring the data change from such

[2] We henceforth use terminologies *2D contingency tables*, *contingency tables*, and *tables* interchangeably if no confusion arises.

2D dynamic tables. Besides the discussion in [17] on extending the disclosure analysis to high-dimensional tables, in [16] the disclosure patterns for 3D tables (data cubes) are further studied with tighter bounds than Fréchet bounds and control methods are proposed to prevent data cubes from leaking of data privacy.

1.2 Our Research Scenario and Solution

An inadequacy of the existing research is that errors in contingency tables have not been taken into account. In practice, it is quite common that errors could be incurred when generating, editing, and reviewing of the tables, leading to inaccurate data in contingency tables released to the public. On the other hand, some knowledgeable snoopers may have accumulated either directly or indirectly the experiences of making such contingency tables and have acquired the knowledge of error distributions by comparing, summarizing, and inducting from the existing tables. Therefore, even with errors, it is still possible for such knowledgeable snoopers to decipher some private information from these tables.

In this paper, we investigate the estimation of privacy disclosure probability and control of disclosure for cells in a contingency table that contains inaccurate data. In detail, after revisiting *four* types of privacy disclosure patterns based on Fréchet bounds identified in our team's earlier study [15], we model the error distributions of marginal values as the normal distribution based on central limit theorem, and estimate the probability for each type of privacy disclosure pattern based on the marginal values and their error distributions. The probability estimations for four types of privacy disclosure patterns are also applicable to error-free tables which can be regarded as a special case where both the mean values and variances of errors approach to zero (i.e., $\mu \to 0$ and $\sigma^2 \to 0$). Given these estimations, we propose two methods known as *marginal value merge* and *marginal value vibration* for the control of privacy disclosure of a cell in a table. With either method, we can guarantee that the disclosure probability of a cell is within 0.5[3] even if a knowledgeable data snooper possesses full knowledge of error distributions. Both methods are developed against knowledgeable snoopers though, they are applicable to against naive data snoopers who have no idea of error distributions given the understanding that a naive snooper cannot infer more privacy from a released table than a knowledgeable snooper can. Moreover, both methods are applicable to error-free tables. We also conduct rigorous experiments to verify the effectiveness of both methods against inferential attacks from knowledgeable data snoopers. To the best of our knowledge, this paper is the first study on the estimation and control of privacy disclosure for 2D contingency tables with inaccurate data.

[3] From a viewpoint of security, we can assume that a cell in a table is safe against privacy disclosure unless the disclosure probability is higher than 0.5. This is because an attacker can give a blind guess to a cell with 50% confidence for any situation of either safe or unsafe, just like tossing a coin with any side as the outcome. Unless the attacker claims with over 50% confidence that the cell is unsafe, we do not have to worry about that he/she will win us in terms of knowing the privacy of the cell.

The remaining sections of the paper are organized as follows. We first introduce preliminaries including different types of privacy disclosure patterns and error distribution of marginal values in Section 2, followed by the estimation of privacy disclosure probability in Section 3 and methods for the control of privacy disclosure in Section 4. We present experimental results with analysis in Section 5 before concluding the paper in Section 6 by pointing out directions for future study.

2 Preliminaries

Before proceeding to estimate the probabilities of privacy disclosure based on Fréchet bounds for 2D contingency tables, we revisit four types of privacy disclosure patterns identified in our team's early study [15] and introduce the modeling of error distributions.

2.1 Revisit of Privacy Disclosure Patterns

Definition 1 (2D Contingency Table). *A 2D contingency table X is an $m \times n$ table with non-negative cell values x_{ij}, denoted as $X = \{x_{ij} | x_{ij} \geqslant 0$, where $1 \leqslant i \leqslant m$ and $1 \leqslant j \leqslant n\}$.* ∎

Definition 2 (Marginal Values). *The marginal values of a contingency table are defined as $x_{i+} = \sum_{j=1}^{n} x_{ij}$, $x_{+j} = \sum_{i=1}^{m} x_{ij}$, $x_{++} = \sum_{i=1}^{m} \sum_{j=1}^{n} x_{ij}$.* ∎

Definition 3 (Fréchet Bounds). *The Fréchet bounds of any cell value x_{ij} of a contingency table are defined as $F_\ell(x_{ij}) \leqslant x_{ij} \leqslant F_u(x_{ij})$ where Fréchet lower bound $F_\ell(x_{ij}) = \max\{x_{i+} + x_{+j} - x_{++}, 0\}$ and Fréchet upper bound $F_u(x_{ij}) = \min\{x_{i+}, x_{+j}\}$.* ∎

It can be proven that the Fréchet bounds of x_{ij} are the exact bounds (i.e., the supremum and the infimum) with which a data snooper can derive out x_{ij} given the marginal values [16]. In what follows, we revisit *four* types of privacy disclosure patterns [15] of a cell value in a contingency table based on its Fréchet bounds. For simplicity, we omit the adjective "Fréchet" when referring to the lower or upper Fréchet bound hereafter if no confusion arises.

Definition 4 (Existence Disclosure). *A cell value x_{ij} suffers from* existence disclosure *if its lower bound is positive, i.e., $F_\ell(x_{ij}) > 0$.* ∎

Definition 5 (τ-Upward Disclosure). *A cell value x_{ij} suffers from* τ-upward disclosure *if its lower bound is greater than a positive threshold τ, i.e., $F_\ell(x_{ij}) > \tau > 0$.* ∎

By these two definitions, if there is not existence disclosure for a cell value, then there is not τ-upward disclosure for this cell value. On the other hand, if there is not τ-upward disclosure for a cell value, there may be existence disclosure for this cell value.

Definition 6 (τ-Downward Disclosure). *A cell value x_{ij} suffers from* τ-downward disclosure *if its upper bound is less than a positive threshold τ, i.e., $F_u(x_{ij}) < \tau$ where $\tau > 0$.* ∎

Definition 7 (τ-Approximation Disclosure). *A cell value x_{ij} suffers from* τ-approximation disclosure *if the difference between its upper and lower bounds is less than a positive threshold τ, i.e., $F_u(x_{ij}) - F_\ell(x_{ij}) < \tau$ where $\tau > 0$.* ∎

From the above definitions, we have the following corollary.

Corollary 1. *For any cell value x_{ij} suffering from* τ-approximation disclosure, *if there is not existence disclosure for this cell value (i.e., $F_\ell(x_{ij}) = 0$), then the* τ-approximation disclosure *of cell value x_{ij} degenerates to its* τ-downward disclosure. ∎

The following examples may give intuitive understandings for these types of privacy disclosure. For example, there are existence disclosure at cell (Tim, T_1) and τ-upward disclosure at cell (Bob, T_1) for $\tau = 1.9$ in the table of Figure 1. If column T_1 in the table records the patient-treatment information, then a snooper can easily infer that Bob receives at least twice $(9+6-13=2)$ and Tim at least once $(9+5-13=1)$ the treatment for specific disease. If the specific disease is highly sensitive, such as AIDS, their respective privacy is violated and they may suffer from discrimination from the society. There is τ-downward disclosure at cell (Dan, T_2) for $\tau = 2.1$. If column T_2 in the table represents certain academic examination results, then a snooper knows the upper bound of Dan's score (i.e., less than 2.1) and can further decipher whether Dan's academic performance is satisfactory or not. Another example in Figure 1 is, there is τ-approximation disclosure at cell (Bob, T_1) for $\tau = 4.1$. If column T_1 in the table stands for salary scales, then Bob's salary scale is known between 2 to 6 and the actual salary scale of Bob can be revealed with high probability.

2.2 Modeling of Error Distribution of Marginal Values

As aforementioned, a released contingency table may contain inaccurate data because errors could be incurred when generating, editing and reviewing the table. The following relationship holds true for any cell value x_{ij} where e_{ij} denotes the error and r_{ij} the correct value of x_{ij}:

$$r_{ij} = x_{ij} + e_{ij} \tag{1}$$

We use μ_{ij} and σ_{ij}^2 to denote the mean and variance of the error e_{ij} of cell value x_{ij}. Regardless the distribution of error e_{ij} of cell value x_{ij}, both its mean μ_{ij} and variance σ_{ij}^2 are statistically obtainable from empirical data when generating the tables.

To estimate the probability of privacy disclosure for a cell value based on its Fréchet bounds, we need to know the error distribution of marginal values. Assume that the error of any cell value is independent from others. Let

$$e_{i+} = \sum_{j=1}^{n} e_{ij}, \qquad e_{+j} = \sum_{i=1}^{m} e_{ij}, \qquad e_{++} = \sum_{i=1}^{m}\sum_{j=1}^{n} e_{ij}$$

where e_{i+} denotes the error of marginal value x_{i+}, e_{+j} the error of marginal value x_{+j}, and e_{++} the error of marginal value x_{++}. Based on Lyapunov's central limit theorem, if the size of a table is big enough (say 30×30), the errors of marginal values follow normal distribution as follows:

$$e_{i+} \sim N\left(\mu_{i+}, \sigma_{i+}^2\right), \qquad e_{+j} \sim N\left(\mu_{+j}, \sigma_{+j}^2\right), \qquad e_{++} \sim N\left(\mu_{++}, \sigma_{++}^2\right)$$

where

$$\begin{cases} \mu_{i+} = \sum_{j=1}^{n} \mu_{ij} \\ \sigma_{i+}^2 = \sum_{j=1}^{n} \sigma_{ij}^2 \end{cases}, \qquad \begin{cases} \mu_{+j} = \sum_{i=1}^{m} \mu_{ij} \\ \sigma_{+j}^2 = \sum_{i=1}^{m} \sigma_{ij}^2 \end{cases}, \qquad \begin{cases} \mu_{++} = \sum_{i=1}^{m} \sum_{j=1}^{n} \mu_{ij} \\ \sigma_{++}^2 = \sum_{i=1}^{m} \sum_{j=1}^{n} \sigma_{ij}^2 \end{cases}$$

3 Estimation of Privacy Disclosure Probability

With the above definitions of privacy disclosure and modeling of error distributions, in what follows, we investigate the probability of a cell value suffering from certain type of privacy disclosure.

Let

$$r_{i+} = \sum_{i=1}^{m} r_{ij}, \qquad r_{+j} = \sum_{j=1}^{n} r_{+j}, \qquad r_{++} = \sum_{i=1}^{m} \sum_{j=1}^{n} r_{ij}$$

where r_{i+}, r_{+j}, and r_{++} represent the correct marginal values. From Equation (1) we have

$$r_{i+} + r_{+j} - r_{++} = x_{i+} + x_{+j} - x_{++} + e_{i+} + e_{+j} - e_{++} \tag{2}$$

Let $\hat{F}_\ell(r_{ij}) = r_{i+} + r_{+j} - r_{++}, \hat{F}_\ell(x_{ij}) = x_{i+} + x_{+j} - x_{++}, \hat{e}_{ij} = e_{i+} + e_{+j} - e_{++}$, Equation (2) can be transformed as

$$\hat{F}_\ell(r_{ij}) = \hat{F}_\ell(x_{ij}) + \hat{e}_{ij}.$$

Let $\hat{\mu}_{ij}$ and $\hat{\sigma}_{ij}^2$ denote the mean and variance of $\hat{e}_{ij}$. Given the assumption that all errors of cell values are mutually independent we have

$$\hat{\mu}_{ij} = \mu_{i+} + \mu_{+j} - \mu_{++} \tag{3}$$

and from $\hat{e}_{ij} = e_{i+} + e_{+j} - e_{++} = e_{ij} - \sum_{\substack{s=1 \\ s \neq i}}^{m} \sum_{\substack{t=1 \\ t \neq j}}^{n} e_{st}$, we have

$$\hat{\sigma}_{ij}^2 = \sigma_{ij}^2 + \sum_{\substack{s=1 \\ s \neq i}}^{m} \sum_{\substack{t=1 \\ t \neq j}}^{n} \sigma_{st}^2 \doteq \sigma_{++}^2 - \sigma_{i+}^2 - \sigma_{+j}^2 \tag{4}$$

Let $P_e(r_{ij})$, $P_u^\tau(r_{ij})$, $P_d^\tau(r_{ij})$, and $P_a^\tau(r_{ij})$ respectively denote the probability of correct value r_{ij} suffering from existence disclosure, τ-upward disclosure, τ-downward disclosure, and τ-approximation disclosure. We have

$$P_e(r_{ij}) = 1 - \Phi\left(\frac{-\hat{F}_\ell(x_{ij}) - \hat{\mu}_{ij}}{\hat{\sigma}_{ij}}\right), \quad P_u^\tau(r_{ij}) = 1 - \Phi\left(\frac{\tau - \hat{F}_\ell(x_{ij}) - \hat{\mu}_{ij}}{\hat{\sigma}_{ij}}\right)$$

$$P_d^\tau(r_{ij}) \doteq \Phi(A) + \Phi(B) - \Phi(A) \cdot \Phi(B)$$
$$P_a^\tau(r_{ij}) \doteq P_d^\tau(r_{ij}) \cdot \big(1 - P_e(r_{ij})\big) + \big(\Phi(C) + \Phi(D) - \Phi(C) \cdot \Phi(D)\big) \cdot P_e(r_{ij})$$

where $\Phi(x)$ is the cumulative density function (cdf for short) of standard normal distribution[4], and A, B, C, and D are defined as follows

$$A = \frac{\tau - x_{i+} - \mu_{i+}}{\sigma_{i+}}, \qquad B = \frac{\tau - x_{+j} - \mu_{+j}}{\sigma_{+j}}$$
$$C = \frac{\tau - x_{++} + x_{i+} - \mu_{++} + \mu_{i+}}{\sqrt{\sigma_{++}^2 - \sigma_{i+}^2}}, \qquad D = \frac{\tau - x_{++} + x_{+j} - \mu_{++} + \mu_{+j}}{\sqrt{\sigma_{++}^2 - \sigma_{+j}^2}}.$$

Detailed calculations about the above disclosure probabilities are given in Appendix A.

4 Control of Privacy Disclosure

As analyzed in the previous section, given the marginal values and their error distributions of a table, the table owner can estimate the probability of privacy disclosure, so can a knowledgeable data snooper if he/she has acquired the error distributions by certain means besides the released marginal values. Therefore, if a table owner can somehow re-organize or camouflage the data before releasing them to the public such that the probability of privacy disclosure is not greater than a predefined probability threshold p_τ, then the table owner can safely assume that the released table is privacy-preserved without worrying about attacks from either naive or knowledgeable snoopers. In our scenario, we set $p_\tau = 0.5$.[5]

Definition 8 (Unsafe Cell Value). *A cell value is assumed* unsafe *if and only if the probability that it suffers from any type of privacy disclosure is greater than* 0.5. ∎

With the above understanding, we propose two methods in the next for re-organizing or camouflaging a table before releasing it. These two methods, known as *marginal value merge* and *marginal value vibration*, ensure that there is not any unsafe cell value in a released table regardless whether or not the original table is error-free. The compromise of using both methods is the acceptably potential loss or inaccuracy of the information released.

[4] The cdf $\Phi(x)$ is defined as $\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} \exp\left(-\frac{u^2}{2}\right) du$ where $x \in \mathbb{R}$. For a random variable X following normal distribution, i.e., $X \sim N(\mu, \sigma^2)$, given x, the probability that $X \leqslant x$ is $P(X \leqslant x) = \Phi\left(\frac{x-\mu}{\sigma}\right)$.

[5] See footnote 3 in subsection 1.2 for a further explanation.

4.1 Marginal Value Merge

Similar to the methods for data sanitization [5,6], the main idea of the method of marginal value merge (MVM for short) is to re-organize a table by merging several marginal values together into a group and releasing their sum instead of individual marginal values. In other words, several rows or columns of a table are merged together as one row or column, thus the marginal value of this merged row or column is the sum of the marginal values of the rows or columns before merge. Taking Figure 1 as example, we may put columns T_1 and T_2 into a group and release the merged marginal value 11 instead of two individual marginal values 9 and 2 to the public. The marginal values merged into a group do not necessarily appear adjacent or consecutive in the original table. This is because usually a released table has been normalized to at least 1NF without any constraint for appearing order of rows and columns. Regardless whether or not a table contains inaccurate data, the MVM method applies to the circumstance in which users may not have to know the details of each marginal value in a group.

Technically, with this MVM method we first estimate the disclosure probability for each cell value and identify those unsafe cell values such as r_{ij}. Then we merge several rows (or columns) together. For example, we merge marginal value x_{i+} (to which the unsafe cell value r_{ij} contributes) with another non-zero marginal value x_{h+} and release the new marginal value $x_{v+} = x_{i+} + x_{h+}$. To minimize the information loss, it is preferred to maximize the number of marginal values after merge as shown by the following optimization:

$$Z = \max\{\text{number of marginal values}\}$$

subject to

$$P(r_{ij}) \leqslant 0.5$$

where r_{ij} $(2 \leqslant i+j < m+n)$ is the correct cell value after merge and $P(r_{ij})$ is the probability of certain type of privacy disclosure for cell value r_{ij}. Theoretically, the optimized solutions exist among $2^{m \times n}$ choices of grouping though, practically the search space is much smaller because many of the grouping are meaningless. Taking Figure 1 as example, if T_1 and T_2 stand for weight and height of a person respectively, it does not make sense to merge them together into a group. Therefore, given the tolerable information loss, we can always find a locally optimal solution from an array of feasible solutions for this optimization. This is quite intuitive as demonstrated by an extreme case in which all marginal values are merged into one group and the marginal value x_{++} is the only element released, leading to a nearly useless table.

4.2 Marginal Value Vibration

The major idea of the method of marginal value vibration (MVV for short) is to camouflage a table by adding vibrations to some marginal values before releasing

them, i.e., by increasing or decreasing some marginal values. This idea can be well depicted by the following optimization:

$$\delta^2 = \min\left\{\sum_{i=1}^{m}\delta_{i+}^2 + \sum_{j=1}^{n}\delta_{+j}^2\right\}$$

subject to

$$P(r_{ij}) \leqslant 0.5 \quad \text{and} \quad \sum_{i=1}^{m}(x_{i+}+\delta_{i+}) = \sum_{j=1}^{n}(x_{+j}+\delta_{+j})$$

where $P(r_{ij})$ is the probability of certain type of privacy disclosure for a cell value r_{ij}, δ_{i+} the vibration for marginal value x_{i+}, and δ_{+j} the vibration for marginal value x_{+j}. We can obtain the optimized vibrations $\delta_{1+}, \delta_{2+}, \ldots, \delta_{m+}$ and $\delta_{+1}, \delta_{+2}, \ldots, \delta_{+n}$ by using simulated annealing [12] to solve the above optimization problem.

The MVV method is applicable to the circumstances where inaccurate marginal values in a released table are acceptable to the public. For error-free tables, the MVV method means to intentionally introduce noises to a table for the purpose of data camouflage so as to preserve data privacy.

5 Experiments

In what follows, we conduct experiments to verify (1) the estimation of privacy disclosure probability, and (2) the effectiveness of both methods for risk control. We synthesize a 30×30 contingency table such that (1) the cell values in row 2 and column 10 follow uniform distribution ranging from 15 to 25, and (2) the cell values in other rows and columns follow uniform distribution[6] ranging from 500 to 1500. An example of such a table could be the stock values of certain commercial or industrial entities during different periods of time, i.e., the cell values in a row stand for the stock values of a commercial or industrial entity in 30 periods of time; whereas the cell values in a column stand for the stock values of all commercial or industrial entities in one period of time. The cell values of row 2 may stand for the data from a low-stock value entity, whereas the cell values in column 10 may stand for a special period of time during which the stock values of all entities are very low due to some reasons such as economic crisis.

We set the mean and variance of the error of a marginal value as 0.01 and 100 times respectively of the marginal value, i.e., $\mu_{i+} = 0.01x_{i+}$, $\sigma_{i+}^2 = 100x_{i+}$ and

[6] We assume uniform distribution as example for all cell values in our experiments. It does not matter what distribution the cell values follow. This is because from our theoretically analysis we can see that the estimation of disclosure probability bears upon the mean values and variances of cell value errors rather than the distribution of cell values or the distribution of errors.

Table 1. Number n of unsafe cell values

τ	480	485	490	495	500
n	0	0	1	1	1

Table 2. Privacy disclosure probability of unsafe cell value $r_{2,10}$

τ	480	485	490	495	500
$P_d^{\tau}(r_{2,10})$	0.4901	0.4999	0.5097	0.5195	0.5293

$\mu_{+j} = 0.01x_{+j}$, $\sigma_{+j}^2 = 100x_{+j}$[7]. All calculations of formulas and optimizations are carried out by Matlab R2008a and Java. Experiments are run on a laptop PC with Duo Core CPU 2GHz clock rate and 2G RAM running Microsoft Windows XP Professional Version 2002 with SP3.

5.1 Estimation of Privacy Disclosure Probability

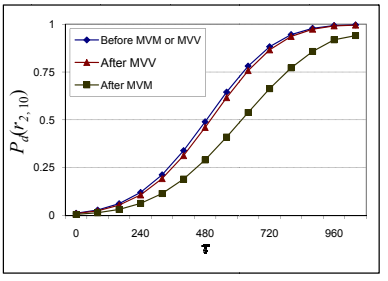
This set of experiments examine the probabilities of privacy disclosure. Based on the estimation formulas derived in Section 3, we calculate the disclosure probability for each cell value with Matlab and identify all unsafe cell values with over 0.5 probability for any type of privacy disclosure. We focus on $\tau = 500$ which is the critical value to differentiate cell values in row 2 or column 10 from other cell values which are not less than 500. In our experiments, we vary τ value from 480 to 500. The experimental results show that (1) there is not any unsafe cell value suffering from existence disclosure. This is because existence disclosure happens on a cell value only if this value is much more greater than all others. All cell values in the table follow uniform distributions hence there is not such an exceptionally great cell value; (2) there is not any unsafe cell value suffering from τ-upward disclosure due to no existence disclosure (refer to Definitions 4 and 5); and (3) when $\tau > 485$, there is only one unsafe cell value $r_{2,10}$ found with over 0.5 probability for τ-downward disclosure (or τ-approximation disclosure equivalently[8]) as shown in Tables 1 and 2.

5.2 Control of Privacy Disclosure

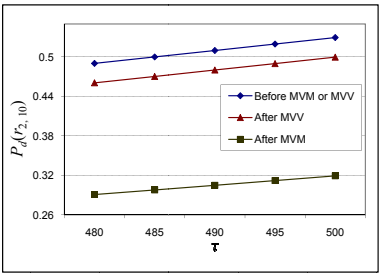
After data re-organizing by the MVM method and data camouflaging by the MVV method, all unsafe cell values are resolved. The experimental results have also shown that after grouping and merging of marginal values with the MVM method for $\tau = 500$, we get 59 marginal values among which 58 are the same as

[7] The choice of the error variance of a marginal value is reasonable. This is because taking marginal value x_{i+} as example, $\sigma_{i+} = 10\sqrt{x_{i+}}$ is less than x_{i+} ($x_{i+} \geqslant 29 \times 500 + 15 > 100$ in our scenario). Thus in the worst case the released marginal value x_{i+} with error is $x_{i+} - \sigma_{i+}$ which is still greater than zero as required for a contingency table.

[8] By Corollary 1, the τ-approximation disclosure of a cell value degenerates to its τ-downward disclosure if there is no existence disclosure for this cell value.



(a) A comparison for $0 \leqslant \tau \leqslant 1000$



(b) A close-up comparison for $480 \leqslant \tau \leqslant 500$

Fig. 2. Comparisons of privacy disclosure probability for cell value $r_{2,10}$

before and only one is new (merged by two previous marginal values). This is because in our experiments the grouping and merging are operated on marginal values x_{2+} and x_{h+} where $h \neq 2$. Moreover, this optimization can be finished within seconds of running time under the system configuration mentioned earlier in this section.

Figure 2 gives a comparison of τ-downward disclosure risk for cell value $r_{2,10}$ under different conditions. Figure 2(a) reveals the fact that cell value $r_{2,10}$ is more likely upper-bounded by τ when τ is close to 1000. However this is not a privacy to preserve because most cell values range from 500 to 1500; whereas cell values in row 2 or column 10 are even smaller, ranging from 15 to 25. Moreover, it comes to the total vibrations $\delta^2 = 2048$ and $|\delta|/x_{++} = 5.3 \times 10^{-5}$, resulting that the curve after operated by the MVV method is very close to the curve before any operation. In other words, while the data privacy is preserved by the MVV method, the vibrations are negligible as compared with the total sum of all cell values hence the data quality is guaranteed for practical usage. Figure 2(b) is a close-up view of the comparison for $480 \leqslant \tau \leqslant 500$ from which we can see that the disclosure probability is not greater than 0.5 after the operation by either the MVM or the MVV method. This optimization can be finished within minutes of running time under the same system configuration.

5.3 Experimental Conclusion and Finding

In summary, we can draw the following conclusions from our experiments: (1) the estimation of privacy disclosure probability can well identify all unsafe cell values; (2) both methods can effectively resolve unsafe cell values; (3) the MVM method leads to information loss however, it does not decrease the utility of a table in some circumstances if the merged marginal values are not of significant interests to users; (4) the MVV method brings with inaccuracy to the released marginal values of a table however, it does not lose any useful information of a table; (5) a table owner has an option from two methods depending on his/her concern on accuracy or information completion of a table.

Moreover, an interesting finding from the experiments is the limitation of the classical analysis of Fréchet bounds. That is, the classical analysis of Fréchet bounds can tell no more than that *any* cell value in row 2 or column 10 may suffer from τ-downward disclosure where $\tau = \min\{x_{2+}, x_{+10}\}$. This is even less than what one can easily infer with just *intuition* that cell value $x_{2,10}$ could be unsafe with potentially higher probability of privacy disclosure as compared with other cell values in row 2 and column 10 given that either marginal value x_{2+} or x_{+10} is much smaller than other marginal values. However, our analysis which is also based on Fréchet bounds can locate cell value $x_{2,10}$ under a smaller τ value with quantitative probability of disclosure. Thus we claim that our analysis and experiments shed a new light for the classical analysis of Fréchet bounds.

6 Conclusion

We have proposed pioneering study on the estimation of privacy disclosure probability and control methods for 2D contingency tables containing inaccurate data. In conclusion, we have made the following contributions. First, we have estimated the probability for each type of privacy disclosure pattern based on the error distribution of marginal values. Second, we have proposed two methods for the control of privacy disclosure so as to preserve private information which could be inferred from released contingency tables. Third, we have conducted experiments which have verified the correctness of our theoretical analysis and the effectiveness of both methods proposed for privacy disclosure control. Fourth, our analysis and methods are also applicable to error-free tables.

For further study, a straightforward direction is to investigate the disclosure control methods for specific contingency tables such as sparse tables; whereas a worthwhile endeavor could be on the quantitative analysis of the relationship between data utility and privacy-preserving of contingency tables.

References

1. Beck, L.L.: A security mechanism for statistical databases. ACM Transactions on Database Systems 5(3), 316–338 (1980)
2. Brodsky, A., Farkas, C., Jajodia, S.: Secure databases: constraints, inference channels, and monitoring disclosures. IEEE Transactions on Knowledge and Data Engineering 12(6), 900–919 (2000)
3. Chen, K., Liu, L.: Privacy preserving data classification with rotation perturbation. In: Proceedings of the 5th IEEE International Conference on Data Mining (ICDM 2005), Houston, TX, USA, November 27-30, pp. 589–592 (2005)
4. Chowdhury, S.D., Duncan, G.T., Krishnan, R., Roehrig, S.F., Mukherjee, S.: Disclosure detection in multivariate categorical databases: auditing confidentiality protection through two new matrix operators. Management Science 45(12), 1710–1723 (1999)
5. Ciriani, V., di Vimercati, S.D.C., Foresti, S., Samarati, P.: K-anonymity. Security in Decentralized Data Management, 323–353 (2007)

6. Cox, L.H.: On properties of multi-dimensional statistical tables. Journal of Statistical Planning and Inference 117(23), 251–273 (2003)
7. Dobra, A., Fienberg, S.E.: Bounds for cell entries in contingency tables induced by fixed marginal totals with applications to disclosure limitation. Statistical Journal of the United States 18(1), 363–371 (2001)
8. Domingo-Ferrer, J., Mateo-Sanz, J.M.: Practical data-oriented microaggregation for statistical disclosure control. IEEE Transactions on Knowledge and Data Engineering 14(1), 189–201 (2002)
9. Farkas, C., Jajodia, S.: The inference problem: a survey. SIGKDD Explorations 4(2), 6–11 (2002)
10. Huang, Z., Du, W., Chen, B.: Deriving private information from randomized data. In: Proceedings of the 2005 ACM SIGMOD International Conference on Management of Data, Baltimore, MD, USA, June 14-16, pp. 37–48 (2005)
11. Kargupta, H., Datta, S., Wang, Q., Sivakumar, K.: On the privacy preserving properties of random data perturbation techniques. In: Proceedings of the 3rd IEEE International Conference on Data Mining (ICDM 2003), Melbourne, FL, USA, December 19-22, pp. 99–106 (2003)
12. LeFevre, K., DeWitt, D.J., Ramakrishnan, R.: Incognito: efficient full-domain k-anonymity. In: Proceedings of the 2005 ACM SIGMOD International Conference on Management of Data, Baltimore, MD, USA, June 14-16, pp. 49–60 (2005)
13. Li, N., Li, T., Venkatasubramanian, S.: t-closeness: privacy beyond k-anonymity and ℓ-diversity. In: Proceedings of the 23rd IEEE International Conference on Data Engineering, Istanbul, Turkey, April 15-20, pp. 106–115 (2007)
14. Liu, K., Kargupta, H., Ryan, J.: Random projection-based multiplicative data perturbation for privacy preserving distributed data mining. IEEE Transactions on Knowledge and Data Engineering 18(1), 92–106 (2006)
15. Lu, H., Li, Y.: Disclosure analysis and control in statistical databases. In: Jajodia, S., Lopez, J. (eds.) ESORICS 2008. LNCS, vol. 5283, pp. 146–160. Springer, Heidelberg (2008)
16. Lu, H., Li, Y.: Practical inference control for data cubes. IEEE Transactions on Dependable Secure Computing 5(2), 87–98 (2008)
17. Lu, H., Li, Y., Wu, X.: On the disclosure risk in dynamic two-dimensional contingency tables (extended abstract). In: Proceedings of the 2nd International Conference on Information System Security (ICISS 2006), Kolkata, India, December 17-21, pp. 349–352 (2006)
18. Lui, S.M., Qiu, L.: Individual privacy and organizational privacy in business analytics. In: Proceedings of the 40th Hawaii International Conference on System Sciences (HICSS 2007), Waikoloa, Big Island, Hawaii, USA, January 3-6, p. 216b (2007)
19. Machanavajjhala, A., Gehrke, J., Kifer, D., Venkitasubramaniam, M.: ℓ-diversity: privacy beyond k-anonymity. In: Proceedins of the 22nd International Conference on Data Engineering (ICDE 2006), Atlanta, GA, USA, April 3-8, pp. 24–35 (2006)
20. Muralidhar, K., Sarathy, R.: A general aditive data perturbation method for database security. Management Science 45(10), 1399–1415 (1999)
21. Schlörer, J.: Security of statistical databases: multidimensional transformation. ACM Transactions on Database Systems 6(1), 95–112 (1981)
22. Schlörer, J.: Information loss in partitioned statistical databases. Computer Journal 26(3), 218–223 (1983)
23. Sweene, L.: Achieving k-anonymity privacy protection using generalization and suppression. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems 10(5), 571–588 (2002)

24. Wang, L., Jajodia, S., Wijesekera, D.: Securing OLAP data cubes against privacy breaches. In: Proceedings of the IEEE Symposium on Security and Privacy (S&P 2004), Berkeley, CA, USA, May 9-12, pp. 161–175 (2004)
25. Wang, L., Li, Y., Wijesekera, D., Jajodia, S.: Precisely answering multi-dimensional range queries without privacy breaches. In: Snekkenes, E., Gollmann, D. (eds.) ESORICS 2003. LNCS, vol. 2808, pp. 100–115. Springer, Heidelberg (2003)

A Estimation of Privacy Disclosure Probability

A.1 Estimation of Existence Disclosure Probability

By the definition of existence disclosure (i.e., $F_\ell(r_{ij}) > 0$, see Definition 4) we have

$$P_e(r_{ij}) = P\big(F_\ell(r_{ij}) > 0\big) = P\left(\hat{F}_\ell(r_{ij}) > 0\right) = P\left(\hat{e}_{ij} + \hat{F}_\ell(x_{ij}) > 0\right)$$
$$= 1 - P\left(\hat{e}_{ij} \leqslant -\hat{F}_\ell(x_{ij})\right) = 1 - \Phi\left(\frac{-\hat{F}_\ell(x_{ij}) - \hat{\mu}_{ij}}{\hat{\sigma}_{ij}}\right)$$

where $\Phi(x)$ is the cumulative density function (cdf for short) of standard normal distribution and $\hat{\mu}_{ij}$ and $\hat{\sigma}_{ij}$ are defined by Equations (3) and (4). This result covers the special case of an error-free table for which $\hat{\mu}_{ij} \to 0$ and $\hat{\sigma}_{ij} \to 0^+$ hold because from $F_\ell(r_{ij}) > 0$ or equivalently $\hat{F}_\ell(r_{ij}) > 0$ (Definition 4) we have

$$\lim_R P_e(r_{ij}) = 1 - \lim_R \Phi\left(\frac{-\hat{F}_\ell(x_{ij}) - \hat{\mu}_{ij}}{\hat{\sigma}_{ij}}\right) = 1 - \Phi(-\infty) = 1 - 0 = 1$$

where R denotes $\{\hat{\mu}_{ij} \to 0,\ \hat{\sigma}_{ij} \to 0^+\}$.

A.2 Estimation of τ-Upward Disclosure Probability

Similarly, by the definition of τ-upward disclosure (Definition 5) we have

$$P_u^\tau(r_{ij}) = P\big(F_\ell(r_{ij}) > \tau\big) = P\left(\hat{F}_\ell(r_{ij}) > \tau\right) = P\left(\hat{e}_{ij} + \hat{F}_\ell(x_{ij}) > \tau\right)$$
$$= 1 - P\left(\hat{e}_{ij} \leqslant \tau - \hat{F}_\ell(x_{ij})\right) = 1 - \Phi\left(\frac{\tau - \hat{F}_\ell(x_{ij}) - \hat{\mu}_{ij}}{\hat{\sigma}_{ij}}\right)$$

where $\hat{\mu}_{ij}$ and $\hat{\sigma}_{ij}$ are defined by Equations (3) and (4). Similar to the previous analysis, for error-free tables, from $F_\ell(r_{ij}) > \tau$ or equivalently $\tau - \hat{F}_\ell(x_{ij}) < 0$ we have

$$\lim_R P_e(r_{ij}) = 1 - \lim_R \Phi\left(\frac{\tau - \hat{F}_\ell(x_{ij}) - \hat{\mu}_{ij}}{\hat{\sigma}_{ij}}\right) = 1 - \Phi(-\infty) = 1 - 0 = 1$$

where R denotes $\{\hat{\mu}_{ij} \to 0,\ \hat{\sigma}_{ij} \to 0^+\}$.

A.3 Estimation of τ-Downward Disclosure Probability

By the definition of τ-downward disclosure (Definition 6) we have

$$\begin{aligned} P_d^{\tau}(r_{ij}) &= P\big(F_u(r_{ij}) < \tau\big) = P\big(\min\{r_{i+},\ r_{+j}\} < \tau\big) \\ &= P\big(\min\{e_{i+} + x_{i+},\ e_{+j} + x_{+j}\} < \tau\big) \end{aligned}$$

Marginal value errors e_{i+} and e_{+j} are not mutually independent because there is a shared element e_{ij}. Given that the size of a table is big enough (e.g., 30×30) and the contribution of the shared element e_{ij} to either e_{i+} or e_{+j} is quite small (e.g., less than 5%), the above calculation can proceed as follows:

$$\begin{aligned} P_d^{\tau}(r_{ij}) &\doteq 1 - P\big(e_{i+} + x_{i+} \geqslant \tau\big) \cdot P\big(e_{+j} + x_{+j} \geqslant \tau\big) \\ &= 1 - \big(1 - P\big(e_{i+} < \tau - x_{i+}\big)\big) \cdot \big(1 - P\big(e_{+j} < \tau - x_{+j}\big)\big) \\ &= \Phi(A) + \Phi(B) - \Phi(A) \cdot \Phi(B) \end{aligned} \tag{5}$$

where $A = \frac{\tau - x_{i+} - \mu_{i+}}{\sigma_{i+}}$ and $B = \frac{\tau - x_{+j} - \mu_{+j}}{\sigma_{+j}}$.

Now consider the special case of an error-free table for which $\mu_{i+} \to 0$, $\mu_{+j} \to 0$, $\sigma_{i+} \to 0^+$, and $\sigma_{+j} \to 0^+$ hold. Let R denote $\{\mu_{i+} \to 0,\ \sigma_{i+} \to 0^+,\ \mu_{+j} \to 0,\ \sigma_{+j} \to 0^+\}$. By definition we have $F_u(r_{ij}) = \min\{e_{i+} + x_{i+}, e_{+j} + x_{+j}\} < \tau$ which contains the following two situations:

(1) $e_{i+} + x_{i+} < \tau$ and $e_{+j} + x_{+j} < \tau$

The above equations are equivalent to $x_{i+} < \tau$ and $x_{+j} < \tau$ under the assumption of error-free tables (i.e., $e_{i+} \to 0$ and $e_{+j} \to 0$). Therefore, from Equation (5) we have

$$\lim_{R} P_d^{\tau}(r_{ij}) \doteq \Phi(+\infty) + \Phi(+\infty) - \Phi(+\infty) \times \Phi(+\infty) = 1 + 1 - 1 \times 1 = 1$$

(2) $e_{i+} + x_{i+} < \tau$ and $e_{+j} + x_{+j} \geqslant \tau$, or $e_{i+} + x_{i+} \geqslant \tau$ and $e_{+j} + x_{+j} < \tau$

Without loss of generality, we consider the first case which is equivalent to $x_{i+} < \tau$ and $x_{+j} \geqslant \tau$ under error-free assumption. From Equation (5) we have

$$\lim_{R} P_d^{\tau}(r_{ij}) \doteq \Phi(+\infty) + \Phi\left(\frac{\tau - x_{+j} - \mu_{+j}}{\sigma_{+j}}\right) - \Phi(+\infty) \times \Phi\left(\frac{\tau - x_{+j} - \mu_{+j}}{\sigma_{+j}}\right) = 1$$

A.4 Estimation of τ-Approximation Disclosure Probability

By the definition of τ-approximation disclosure (Definition 7) we have

$$\begin{aligned} P_a^{\tau}(r_{ij}) &= P\big(F_u(r_{ij}) - F_{\ell}(r_{ij}) < \tau\big) \\ &= P\big(F_u(r_{ij}) - \max\{\hat{F}_{\ell}(r_{ij}),\, 0\} < \tau\big) \\ &\doteq P\big(F_u(r_{ij}) < \tau\big) \cdot P\big(\hat{F}_{\ell}(r_{ij}) < 0\big) \\ &\quad + P\big(F_u(r_{ij}) - \hat{F}_{\ell}(r_{ij}) < \tau\big) \cdot P\big(\hat{F}_{\ell}(r_{ij}) \geqslant 0\big) \\ &= P_d^{\tau}(r_{ij}) \cdot \big(1 - P_e(r_{ij})\big) + P_d^{\tau}\big(r_{ij} - \hat{F}_{\ell}(r_{ij})\big) \cdot P_e(r_{ij}) \end{aligned} \tag{6}$$

where

$$
\begin{aligned}
& P_d^{\tau}\big(r_{ij} - \hat{F}_{\ell}(r_{ij})\big) \\
&= P\big(F_u(r_{ij}) - \hat{F}_{\ell}(r_{ij}) < \tau\big) \\
&= P\big(\min\{r_{i+} - \hat{F}_{\ell}(r_{ij}),\ r_{+j} - \hat{F}_{\ell}(r_{ij})\} < \tau\big) \\
&= P\big(\min\{e_{i+} + x_{i+} - \hat{F}_{\ell}(x_{ij}) - \hat{e}_{ij},\ e_{+j} + x_{+j} - \hat{F}_{\ell}(x_{ij}) - \hat{e}_{ij}\} < \tau\big) \\
&= P\big(\min\{e_{++} - e_{+j} + x_{++} - x_{+j},\ e_{++} - e_{i+} + x_{++} - x_{i+}\} < \tau\big) \\
&\doteq 1 - P\big(e_{++} - e_{+j} + x_{++} - x_{+j} \geqslant \tau\big) \cdot P\big(e_{++} - e_{i+} + x_{++} - x_{i+} \geqslant \tau\big) \\
&= 1 - P\big(e_{++} - e_{+j} \geqslant \tau - x_{++} + x_{+j}\big) \cdot P\big(e_{++} - e_{i+} \geqslant \tau - x_{++} + x_{i+}\big) \\
&= \Phi(A) + \Phi(B) - \Phi(A) \cdot \Phi(B)
\end{aligned}
$$

where

$$
A = \frac{\tau - x_{++} + x_{i+} - \mu_{++} + \mu_{i+}}{\sqrt{\sigma_{++}^2 - \sigma_{i+}^2}}, \qquad B = \frac{\tau - x_{++} + x_{+j} - \mu_{++} + \mu_{+j}}{\sqrt{\sigma_{++}^2 - \sigma_{+j}^2}}
$$

Again, consider the special case of a table without errors for which the following relationships hold:

$$
\mu_{i+} \to 0, \quad \mu_{+j} \to 0, \quad \mu_{ij} \to 0, \quad \sqrt{\sigma_{++}^2 - \sigma_{i+}^2} \to 0^+, \quad \sqrt{\sigma_{++}^2 - \sigma_{+j}^2} \to 0^+
$$

Let R denote

$$
\left\{\mu_{i+} \to 0,\ \mu_{+j} \to 0,\ \mu_{ij} \to 0,\ \sqrt{\sigma_{++}^2 - \sigma_{i+}^2} \to 0^+,\ \sqrt{\sigma_{++}^2 - \sigma_{+j}^2} \to 0^+\right\}.
$$

By the definition of τ-approximation disclosure (Definition 7), we have $F_u(r_{ij}) - F_{\ell}(r_{ij}) < \tau$ or equivalently $\min\{r_{i+}, r_{+j}\} - \max\{r_{i+} + r_{+j} - r_{++}, 0\} < \tau$ for which there are two cases as follows:

(1) $r_{i+} + r_{+j} - r_{++} \leqslant 0$ or equivalently $F_{\ell}(r_{ij}) = 0$. Based on Corollary 1, the τ-approximation disclosure of cell value x_{ij} degenerates to its τ-downward disclosure. Thus we have

$$
\lim_{R} P_a^{\tau}(r_{ij}) = \lim_{S} P_d^{\tau}(r_{ij}) = 1
$$

where S denotes $\{\mu_{i+} \to 0,\ \sigma_{i+} \to 0^+,\ \mu_{+j} \to 0,\ \sigma_{+j} \to 0^+\}$.

(2) $r_{i+} + r_{+j} - r_{++} > 0$ or equivalently $F_{\ell}(r_{ij}) > 0$. In this case, the relationship $\min\{r_{i+}, r_{+j}\} - \max\{r_{i+} + r_{+j} - r_{++}, 0\} < \tau$ can be transformed to $\min\{r_{++} - r_{i+},\ r_{++} - r_{+j}\} < \tau$. With error-free assumption and $F_{\ell}(r_{ij}) > 0$, we have $P_e(r_{ij}) = 1$. Therefore, from Equation (6) we have

$$
\lim_{R} P_a^{\tau}(r_{ij}) = \lim_{R} P_d^{\tau}\big(r_{ij} - \hat{F}_{\ell}(r_{ij})\big) = 1
$$

A Tool for Analyzing and Fixing Infeasible RCTA Instances*

Jordi Castro** and José A. González

Department of Statistics and Operations Research,
Universitat Politècnica de Catalunya,
Jordi Girona 1–3, 08034 Barcelona, Catalonia
jordi.castro@upc.edu,
jose.a.gonzalez@upc.edu
http://www-eio.upc.es/~jcastro

Abstract. Minimum-distance controlled tabular adjustment methods (CTA), and its restricted variants (RCTA), is a recent perturbative approach for tabular data protection. Given a table to be protected, the purpose of RCTA is to find the closest table that guarantees protection levels for the sensitive cells. This is achieved by adding slight adjustments to the remaining cells, possibly excluding a subset of them (usually, the total cells) which preserve their original values. If either protection levels are large, or the bounds for cell deviations are tight, or too many cell values have to be preserved, the resulting mixed integer linear problem may be reported as infeasible. This work describes a tool developed for analyzing infeasible instances. The tool is based on a general elastic programming approach, which considers an artificial problem obtained by relaxing constraints and bounds through the addition of extra elastic variables. The tool allows selecting the subset of constraints and bounds to be relaxed, such that an elastic filter method can be applied for isolating a subset of infeasible table relations, protection levels, and cell bounds. Some computational experiments are reported using real-world instances.

Keywords: statistical disclosure control, controlled tabular adjustment, mixed integer linear programming, infeasibility in optimization, elastic constraints, elastic filter.

1 Introduction

Minimum-distance controlled tabular adjustment methods (CTA) were suggested in [1,9] as an alternative to the difficult cell suppression problem (CSP) [2,11]. In some instances, the quality of CTA solutions has shown to be higher than that of

* Supported by grants MTM2009-08747 of the Spanish Ministry of Science and Innovation, SGR-2009-1122 of the Government of Catalonia, and Eurostat framework contract 22100.2006.002-226.532.

** Corresponding author.

J. Domingo-Ferrer and E. Magkos (Eds.): PSD 2010, LNCS 6344, pp. 17–28, 2010.

solutions provided by CSP [4]. Moreover, quality criteria can be easily added to CTA [6]. A variant of CTA, where only a subset of the cells are allowed to be modified (e.g., total cells), is named restricted controlled tabular adjustment (RCTA).

In 2008, RCTA was included in a solution scheme for the protection of structural business statistics released by Eurostat. The resulting RCTA package [5,10] was developed by the authors under the Eurostat framework contract 22100.2006.002-226.532 in collaboration with Statistics Germany and Statistics Netherlands. This package is linked to two state-of-the-art commercial solvers, CPLEX and Xpress, and it can be used as an stand-alone package or a callable library. It offers the user control about the most instrumental parameters for the mixed integer linear optimization problem (MILP) to be solved. In 2009–2010 this package was extended for the protection of animal production statistics of the European Union, again released by Eurostat. One of these extensions was a tool for analyzing infeasible RCTA instances. Infeasibilities in the MILP optimization model for RCTA may arise by many factors—indeed, by interactions of them: (i) protection levels of sensitive cells may be too large, such that the remaining cells can not accommodate to them (i.e., can not be sufficiently perturbed); (ii) the bounds of non-sensitive cells can be too tight (thus limiting the allowed perturbations); (iii) a particular case of previous point (ii) is when bounds are zero, i.e., some cell values have to be preserved in the released table (usually total cells).

Detecting the cause of infeasibility in MILP optimization is much harder than in LP optimization. Indeed, procedures as, for instance, finding an irreducible infeasible set (IIS) are available in some state-of-the-art solvers only for LP problems, but not for MILP problems (an IIS is a minimal set of constraints and bounds which is infeasible, but it becomes feasible if any constraint or bound is removed). In addition, IIS is in general very time consuming for medium-large instances. For this reason we have considered a general elastic programming approach, which is efficient even for moderately large instances. Briefly, the elastic programming approach computes a minimal relaxation of the constraints and bounds of the problem. The recent monograph [7]—and the many references herein—surveys the state-of-the-art in detecting infeasibility in optimization.

The paper is organized as follows. Section 2 reviews the RCTA MILP formulation. Section 3 shows the methodology underlying the infeasibility repair tool developed. Section 4 describes some of the features of the infeasibility repair tool developed, illustrated by an example. Finally, Section 5 reports computational results with some real-world RCTA instances.

2 The RCTA Problem

Given (i) a set of cells $a_i, i = 1, \ldots, n$, that satisfy some linear relations $Aa = b$ (a being the vector of a_i's); (ii) a lower and upper bound for each cell $i = 1, \ldots, n$, respectively l_{a_i} and u_{a_i}, which are considered to be known by any attacker; (iii) a set $\mathcal{S} = \{i_1, i_2, \ldots, i_s\} \subseteq \{1, \ldots, n\}$ of indices of sensitive cells; (iv) and a lower and upper protection level for each sensitive cell $i \in \mathcal{S}$, respectively lpl_i and upl_i,

such that the released values satisfy either $x_i \geq a_i + upl_i$ or $x_i \leq a_i - lpl_i$; the purpose of CTA is to find the closest safe values $x_i, i = 1, \ldots, n$, according to some distance L, that makes the released table safe. This involves the solution of the following optimization problem:

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

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

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

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

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

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

If the problem has negative protection levels (i.e., $lpl_i < 0$ or $upl_i < 0$ for at least one cell i), the optimization model (3) is no longer valid (let us name it the "classical" model). Problems with negative protection levels can be useful for the sequential protection of correlated tables (indeed, this feature was needed for the protection of real-world data and it was added to the RCTA package in a recent extension [5]). For problems with negative protection levels the following alternative model may be used [3]:

$$\begin{array}{ll} \min\limits_{z^+,z^-,y} & \sum\limits_{i=1}^{n} w_i(z_i^+ + z_i^-) \\ \text{subject to} & A(z^+ - z^-) = 0 \\ & l_z \leq z^+ - z^- \leq u_z \\ & z_i^+ - z_i^- \geq upl_i y_i + l_{z_i}(1 - y_i) \quad i \in \mathcal{S} \\ & z_i^+ - z_i^- \leq -lpl_i(1 - y_i) + u_{z_i} y_i \quad i \in \mathcal{S} \\ & (z^+, z^-) \geq 0 \\ & y_i \in \{0, 1\} \quad i \in \mathcal{S}. \end{array} \tag{4}$$

The main difference between (4) and (3) is that (z^+, z^-) are not related to upper and lower protection deviations in (4), but they are just auxiliary variables to model the L_1 distance. As a result, model (4) is valid for any kind of instance, with either positive or negative protection levels. However, it is less efficient than the classical model (3), and then, for problems with only positive protection levels, the classical model is in general a better option [3].

3 The Elastic Programming Approach for Analyzing Infeasibility

Elastic constraints (and bounds) are constraints (and bounds) that can be relaxed (i.e., violated, stretched) by a certain amount. This amount is represented by one or two artificial variables for each relaxed constraint and bound. The elastic constraints for general inequality and equality constraints are:

Nonelastic constraints	Elastic constraints
$A_1 x \geq b^1$	$A_1 x + s^1 \geq b^1$
$A_2 x \leq b^2$	$A_2 x - s^2 \leq b^2$
$A_3 x = b^3$	$A_3 x + s^3 - s^4 = b^3,$

where all the artificial variables s^1, s^2, s^3, s^4 are nonnegative. Once the artificial variables have been added to all the constraints and bounds (or a subset of them), the elastic problem to be solved is to minimize a function of the artificial variables (according to some objective) subject to the elastic constraints and bounds, and to the remaining constraints and bounds that were not relaxed (if any). Some of the different objectives that can be used are: (1) minimize the sum of artificial variables, i.e., $||s^1||_1 + ||s^2||_1 + ||s^3||_1 + ||s^4||_1$; (2) minimize the Euclidean distance of the artificial variables , i.e., $||s^1||_2^2 + ||s^2||_2^2 + ||s^3||_2^2 + ||s^4||_2^2$; (3) minimize the number of relaxed constraints. Objectives (2) and (3) give rise respectively to a quadratic and a MILP problem, even if the original problem was neither quadratic nor MILP. Objective (1) has been our choice for RCTA.

Applying the above elastic programming approach to the RCTA model (3) the resulting problem is

$$
\begin{aligned}
f^* = \min_{s^i, i=1,\ldots,10} \; & \sum_{i=1}^{10} \sum_{j=1}^{n_i} c_j^i s_j^i \\
\text{s. to } & A(z^+ - z^-) + s^1 - s^2 = 0 \\
& z_i^+ - s_i^3 \leq u_{z_i} && i \notin \mathcal{S} \\
& z_i^+ + s_i^4 \geq 0 && i \notin \mathcal{S} \\
& z_i^- - s_i^5 \leq -l_{z_i} && i \notin \mathcal{S} \\
& z_i^- + s_i^6 \geq 0 && i \notin \mathcal{S} \\
& z_i^+ - upl_i \; y_i + s_i^7 \geq 0 && i \in \mathcal{S} \\
& z_i^+ - u_{zi} \; y_i - s_i^8 \leq 0 && i \in \mathcal{S} \\
& z_i^- + lpl_i \; y_i + s_i^9 \geq lpl_i && i \in \mathcal{S} \\
& z_i^- + l_{zi} \; y_i - s_i^{10} \leq -l_{zi} && i \in \mathcal{S} \\
& y_i \in \{0, 1\} && i \in \mathcal{S} \\
& s^i \geq 0 && i = 1, \ldots, 10,
\end{aligned}
\tag{5}
$$

where c^i denotes a penalty vector for each vector of artificial variables s^i, n_i denotes the dimension of each vector $s^i, c^i, i = 1, \ldots, 10$, and f^* is the optimal objective function value obtained. Similarly, the elastic version of the RCTA model (4) for problems with negative protection levels is

$$
\begin{aligned}
f^* = \min_{s^i, i=1,\ldots,10} \; & \sum_{i=1}^{6} \sum_{j=1}^{n_i} c_j^i s_j^i \\
\text{s. to } & A(z^+ - z^-) + s^1 - s^2 = 0 \\
& z^+ - z^- - s^3 \leq u_z \\
& z^+ - z^- + s^4 \geq l_z \\
& z_i^+ - z_i^- + (l_{z_i} - upl_i) y_i + s_i^5 \geq l_{z_i} && i \in \mathcal{S} \\
& z_i^+ - z_i^- + (-lpl_i - u_{z_i}) y_i - s_i^6 \leq -lpl_i && i \in \mathcal{S} \\
& (z^+, z^-) \geq 0 \\
& y_i \in \{0, 1\} && i \in \mathcal{S} \\
& s^i \geq 0 && i = 1, \ldots, 6.
\end{aligned}
\tag{6}
$$

If all the constraints and variables are relaxed, the solution of either problem (5) or (6) will provide an optimal solution. If only a subset of constraints and bounds are relaxed, then (5) or (6) may still result in an infeasible problem. In this case, the subset of relaxed constraints and bounds should be augmented with some additional constraints and bounds. Once a feasible solution to either (5) or (6) is available, a second optimization problem is solved. The purpose of this second phase is to optimize the objective function in terms of the cell deviations, not the artificial variables, such that the solution provided makes sense for RCTA. In this second phase it is imposed as an additional constraint that the sum of artificial variables is less or equal than f^*, the solution of (5) or (6). We know that this problem is feasible, since at least one solution exists (the one reported by (5) or (6)). Therefore, this second optimization is made up of the objective function of (3) (or (4)), the constraints of (5) or (6), and the extra constraint

$$
\sum_{i=1}^{t} \sum_{j=1}^{n_i} c_j^i s_j^i \leq (1 + \delta) f^*,
$$

where t is either 10 or 6 (depending on whether we used (5) or (6) in the first phase), and $\delta \geq 0$ is a small parameter (e.g., $\delta = 0.001$) to slightly relax the right-hand-side, thus avoiding infeasibility issues. The second phase may be started from the optimal solution of the first problem.

By selecting and iteratively updating the subset of elastic constraints and bounds it would be possible to isolate the cause of infeasibility, i.e., it could be obtained a subset of constraints such that, if not elasticized, the resulting RCTA instance is infeasible. The elastic filter method [8] is an automatic procedure for generating such a subset of constraints. It iteratively solves a sequence of elastic problems, de-elasticizing at each iteration the constraints with a positive artificial variable in the solution. When the elastic problem becomes infeasible, the set of de-elasticized constraints provides the infeasible subset of constraints (i.e., if they are not relaxed, the resulting RCTA model is infeasible). The main inconvenience of this approach is that for large infeasible RCTA instances, each iteration may take a long execution time. This elastic filter approach has not been implemented in the repair tool described in the next section; however, it can be manually applied by the end-user by providing specific subsets of constraints to be relaxed (as shown below, this is one of the features of the infeasibility repair tool).

4 The Infeasibility Repair Tool

The procedure described in the previous section has been implemented and added to a package for RCTA. The resulting tool is named the "infeasibility repair tool". The repair tool has two different working modes. In the first one, it relaxes all the constraints and bounds. In the second mode, the user may select a subset of table constraints $A(z^+ - z^-) = 0$, constraints imposing protection levels for sensitive cells, and bounds on cells deviations. This information is provided by the user in a file with the format shown if Figure 1.

When a cell (either sensitive or not) is included in the second section of Figure 1, its upper bound is relaxed, but not its lower bound. Note that relaxing lower bounds (which are usually zero in most tables) would provide solutions

```
nr, number of table constraints allowed to be relaxed (may be 0)
table constraint number 1
...
table constraint number nr
nx, number of cells allowed to relax their bounds (may be 0)
cell number 1
...
cell number nx
ns, number of sensitive cells allowed to relax their protection levels (may be 0)
cell number 1
...
cell number ns
```

Fig. 1. Format of file for selecting the subset of constraints and bounds to be relaxed

with negative values, which are meaningless; on the other hand, increasing the lower bound would restrict the problem, instead of relaxing it. Therefore, lower bounds are kept fixed. However, the relaxation could be performed for positive lower bounds. In turn, when a sensitive cell is included in the third section of Figure 1, either the constraints

$$upl_i \; y_i \leq z_i^+ \leq u_{z_i} \; y_i \;\; i \in \mathcal{S}$$
$$lpl_i(1 - y_i) \leq z_i^- \leq -l_{z_i}(1 - y_i) \; i \in \mathcal{S},$$

of (3) or, if negative protection levels are present, the constraints

$$z_i^+ - z_i^- \geq upl_i y_i + l_{z_i}(1 - y_i) \; i \in \mathcal{S}$$
$$z_i^+ - z_i^- \leq -lpl_i(1 - y_i) + u_{z_i} y_i \; i \in \mathcal{S}$$

of (4) may be relaxed, which in practice means that both protection levels can be violated. If there exists a solution for the relaxed problem, the tool writes an output file with information about the infeasible cells and infeasible linear table relations.

4.1 Example

The table shown in Figure 2(a), with upper bounds in Figure 2(b), is reported as infeasible by the RCTA package. The file describing this instance in the standard "csplib" format is reported in the Appendix A. If the program is run with default parameters, the following output is obtained:

```
Problem reported as infeasible: optimization terminated
(and not by time limit) with no feasible CTA table
Total CPU time: 0.05
```

If the repair tool is applied, the resulting output is:

```
Repair infeasibility procedure successfully finished, see information file.
Total CPU time: 0.1
```

The information file is:

```
Constraints.
Const. num.     left-hand side  right-hand side
0 infeasibilities detected.

Cells.
0 infeasibilities detected among the variables.

Sensitive cells.
Cell 0 (25.996) under UPL (30)
```

meaning that all the table constraints are satisfied (i.e., the table is additive), all the cells remain between bounds, and there is one sensitive cell that could not fulfill its upper protection level of 30. Note that the value appearing in the

$\mathbf{300}_{\mathbf{40}}^{\mathbf{30}}$	8	5	5	5	$\mathbf{38}_{\mathbf{10}}^{\mathbf{4}}$	361
8	$\mathbf{68}_{\mathbf{10}}^{\mathbf{6}}$	40	76	29	31	252
11	33	20	60	35	44	203
7	28	$\mathbf{36}_{\mathbf{9}}^{\mathbf{3}}$	41	22	63	197
326	137	101	182	91	176	1013

(a)

345	15	10	16	13	44
12	78	46	87	33	36
18	38	23	69	40	51
15	32	41	47	25	72

(b)

326	0	0	5	5	25	361
0	74	40	76	29	33	252
0	35	22	60	35	51	203
0	28	39	41	22	67	197
326	137	101	182	91	176	1013

(c)

Fig. 2. (a) Original infeasible table, with primaries in boldface, lower protections levels as subscripts, and upper protection levels as superscripts; (b) upper bounds for cells, table margins are fixed; (c) Adjusted, nonsafe table after repair infeasibility procedure, with unprotected cell marked in boldface

file refers to the deviation from the initial cell value of 300, so the value for the first cell would be 325.996 (rounded to 326 in Figure 2(c), for convenience), suggesting that if the protection level was 26 instead of 30 the table would have been satisfactorily protected.

If, instead, one is interested in preserving the table linear relations and variable bounds, and only relaxing a subset of the sensitive cells (e.g., sensitive cells with values 38, 68 and 36), the infeasibility repair tool should be fed with the following file:

```
0
0
3
5
8
23
```

Note that the above file matches the format of Figure 1, and that, according to Appendix A, cells 5, 8 and 23 are those with values 38, 68 and 36. Indeed, by removing the cell with value 300 from the file—the one whose protection levels were relaxed in the previous run—we are manually applying the elastic filter method described in Section 3. Running the infeasibility repair tool the following output message is obtained:

```
Repair infeasibility procedure reported relaxed problem is infeasible.
Total CPU time: 0.04
```

It means that only relaxing the protection levels of the three selected cells is not possible to obtain a feasible solution. In this case, the protection levels of

sensitive cell of value 300 have to be relaxed, otherwise the problem becomes infeasible.

5 Computational Results

Most of our available real-world instances—from data provided by Eurostat, and processed by Statistics Germany and Statistics Netherlands—are feasible. Then, for the only purpose of testing, the infeasibility tool was initially applied to a set of feasible real-world instances. We note that the procedure based on elastic programming is equally valid for feasible than for infeasible problems: the only difference is that in the feasible case the sum of elastic variables in the first optimization problem will be zero, and that the solution of the second phase will be a valid RCTA solution. The instances considered are related to structural business statistics, for different NACE sector (C, D and E), and to animal production statistics of the European Union. These instances can be considered difficult, since they have a complex structure. The dimensions of these instances, and the results obtained with the RCTA package, with and without the infeasibility repair tool, are reported in Table 1. Problem names starting with "sbs" and "aps" correspond, respectively, to structural business statistics and animal production statistics instances. Columns n, s and m provide the number of cells, sensitive cells and linear relations of the table. Columns "objective", and "CPU" show the final value of the objective functions, and CPU time in seconds obtained with CPLEX-11, with and without the infeasibility repair tool. For executions with the infeasibility repair tool, the objective function corresponds to the solution of the second optimization problem (when it finished, the objective function of the first optimization—the sum of elastic variables—was zero for feasible instances). The required optimality gap was of at most 5% for all the executions. A time limit of one day of CPU (86400 seconds) was set. All the runs have been performed on a Linux Dell Precision T5400 workstation with 16GB of memory and four Intel Xeon E5440 2.83 GHz processors, without exploitation of parallelism capabilities.

For testing the infeasibility repair tool on a non-trivial infeasible case, the instance sbs-E was modified. Results for the new instance, named sbs-E-infeas, are reported in the last line of Table 1. The problem is reported as infeasible in 0 seconds. If very large upper bounds are considered for cell deviations, then the instance is feasible (with an objective function of 106643) but with two unprotected cells due to numerical issues related to feasiblity tolerances and the large bounds considered. After applying the infeasibility repair tool a (infeasible) solution of objective 89931 in the second optimization problem is reported; the objective of the first optimization problem (i.e., the sum of elastic variables, or sum of infeasibilities) was 709546. In this case, infeasibility is being caused by a single constraint. From Table 1 it is clear that the elastic models are much more computationally expensive than the standard RCTA models. We also mention that CPLEX showed to be more robust than Xpress for the solution of the elastic formulations. In particular, Xpress could not solve any of the "sbs" instances with the repair tool. The smaller "aps" instances could be solved by both solvers.

Table 1. Results with CPLEX-11, with and without applying the infeasiblity repair tool, for some real data of structural business statistics and animal production statistics (provided by Eurostat, and processed by Statistics Netherlands and Statistics Germany)

				Without repair		With repair	
Problem	n	s	m	objective	CPU	objective	CPU
sbs-E	1430	382	991	107955	4	106257	297
sbs-C	4212	1135	2580	314282	52	313888	2225
sbs-D_a	28288	7142	13360	414294 (28.3%)	(1)	(2)	
sbs-D_b	28288	7131	13360	444455 (13.5%)	(1)	(2)	
aps-0102	87	5	35	7.20	0.01	7.20	0.03
aps-0203	87	5	35	67.42	0.01	67.42	0.03
aps-0304	87	5	35	12.07	0.02	12.07	0.02
aps-0405	87	5	35	60.77	0.02	60.77	0.02
sbs-E-infeas	1430	382	991	(3)	0	89931	53

(1) Time limit reached with suboptimal solution (gap in brackets).
(2) Time limit reached with no repair tool solution.
(3) Problem reported as infeasible.

6 Conclusions

Detecting what makes a RCTA instance infeasible may be of great help for data owners. However, isolating the source of infeasibility in a MILP is a difficult task. The tool implemented in this work can be used for obtaining a set of constraints such that, if not relaxed, the instance becomes infeasible. The tool is based on adding extra elastic variables to constraints and bounds. The resulting problem is a MILP one, with a higher number of variables than the original RCTA one, and it requires an efficient MILP solver. Our tool was linked to two of them, CPLEX and Xpress, the former seeming to be the most efficient for the elastic model. The tool may also be used not only for real infeasible instances, but also for problematic instances which are reported as infeasible by numerical tolerances. This tool can be seen as another step towards a reliable RCTA package for tabular data protection.

References

1. Castro, J.: Minimum-distance controlled perturbation methods for large-scale tabular data protection. European Journal of Operational Research 171, 39–52 (2006)
2. Castro, J.: A shortest paths heuristic for statistical disclosure control in positive tables. INFORMS Journal on Computing 19, 520–533 (2007)
3. Castro, J.: Extending controlled tabular adjustment for non-additive tabular data with negative protection levels, Research Report DR 2010-01, Dept. of Statistics and Operations Research, Universitat Politècnica de Catalunya (2010) (submitted)

4. Castro, J., Giessing, S.: Testing variants of minimum distance controlled tabular adjustment. In: Monographs of Official Statistics, Eurostat-Office for Official Publications of the European Communities, Luxembourg, pp. 333–343 (2006)
5. Castro, J., González, J.A., Baena, D.: User's and programmer's manual of the RCTA package, Technical Report DR 2009-01, Dept. of Statistics and Operations Research, Universitat Politècnica de Catalunya (2009)
6. Cox, L.H., Kelly, J.P., Patil, R.: Balancing quality and confidentiality for multivariate tabular data. In: Domingo-Ferrer, J., Torra, V. (eds.) PSD 2004. LNCS, vol. 3050, pp. 87–98. Springer, Heidelberg (2004)
7. Chinneck, J.W.: Feasibility and Infeasibility in Optimization: Algorithms and Computational Methods. Springer, Heidelberg (2008)
8. Chinneck, J.W., Dravnieks, E.W.: Locating minimal infeasible constraint sets in linear programs. ORSA Journal on Computing 3, 157–168 (1991)
9. Dandekar, R.A., Cox, L.H.: Synthetic tabular Data: an alternative to complementary cell suppression. Energy Information Administration, U.S. (2002) (manuscript)
10. Giessing, S., Hundepool, A., Castro, J.: Rounding methods for protecting EU-aggregates. In: Eurostat methodologies and working papers. Worksession on statistical data confidentiality, Eurostat-Office for Official Publications of the European Communities, Luxembourg, 255–264 (2009)
11. Kelly, J.P., Golden, B.L., Assad, A.A.: Cell suppression: disclosure protection for sensitive tabular data. Networks 22, 28–55 (1992)

A File of the Example of Figure 2 in "csplib" Format

```
0
34
0 300 1 u 0 345 40 30 0
1 8 1 s 0 15 0 0 0
2 5 1 s 0 10 0 0 0
3 5 1 s 0 16 0 0 0
4 5 1 s 0 13 0 0 0
5 38 1 u 0 44 10 4 0
6 361 1 z 0 0 0 0 0
7 8 1 s 0 12 0 0 0
8 68 1 u 0 78 10 6 0
9 40 1 s 0 46 0 0 0
10 76 1 s 0 87 0 0 0
11 29 1 s 0 33 0 0 0
12 31 1 s 0 36 0 0 0
13 252 1 z 0 0 0 0 0
14 11 1 s 0 18 0 0 0
15 33 1 s 0 38 0 0 0
16 20 1 s 0 23 0 0 0
17 60 1 s 0 69 0 0 0
18 35 1 s 0 40 0 0 0
19 44 1 s 0 51 0 0 0
20 203 1 z 0 0 0 0 0
21 7 1 s 0 15 0 0 0
22 28 1 s 0 32 0 0 0
23 36 1 u 0 41 9 3 0
24 41 1 s 0 47 0 0 0
25 22 1 s 0 25 0 0 0
26 63 1 s 0 72 0 0 0
27 197 1 z 0 0 0 0 0
28 326 1 z 0 0 0 0 0
29 137 1 z 0 0 0 0 0
30 101 1 z 0 0 0 0 0
31 182 1 z 0 0 0 0 0
32 91 1 z 0 0 0 0 0
33 176 1 z 0 0 0 0 0
10
0 7 : 6(-1) 0(1) 1(1) 2(1) 3(1) 4(1) 5(1)
0 7 : 13(-1) 7(1) 8(1) 9(1) 10(1) 11(1) 12(1)
0 7 : 20(-1) 14(1) 15(1) 16(1) 17(1) 18(1) 19(1)
0 7 : 27(-1) 21(1) 22(1) 23(1) 24(1) 25(1) 26(1)
0 5 : 28(-1) 0(1) 7(1) 14(1) 21(1)
0 5 : 29(-1) 1(1) 8(1) 15(1) 22(1)
0 5 : 30(-1) 2(1) 9(1) 16(1) 23(1)
0 5 : 31(-1) 3(1) 10(1) 17(1) 24(1)
0 5 : 32(-1) 4(1) 11(1) 18(1) 25(1)
0 5 : 33(-1) 5(1) 12(1) 19(1) 26(1)
```

Branch-and-Cut versus Cut-and-Branch Algorithms for Cell Suppression

Juan-José Salazar-González

DEIOC - University of La Laguna,
38271 La Laguna, Tenerife, Spain
jjsalaza@ull.es
http://webpages.ull.es/users/jjsalaza

Abstract. This paper discusses several techniques to apply Cell Suppression Methodology to protect private information when publishing tabular data. All techniques are exact algorithms to find optimal suppression patterns, but they can also be used as heuristic approaches to find good suppression patterns. One of the technique is the branch-and-cut algorithm described in Fischetti and Salazar (2000). A variant of this technique is presented in this paper with the name cut-and-branch algorithm. The paper discusses advantages and disadvantages of the cut-and-branch algorithm when compared to the branch-and-cut algorithm, and shows computational results on a set of real world instances. The computer implementation has been done using only free and open-source libraries. The paper concludes with an exact approach to apply Cell Suppression on tabular data where the response variable is discrete (like in a frequency table).

Keywords: Cell Suppression, Tabular Protection, Branch-and-Cut, Integer Programming.

1 Introduction

Data stewardship organizations must ensure that they protect the information provided by respondents, while enabling to release high quality information about the population and the economy. To allow this guarantee experts have developed methodologies to protect private information. When publishing tabular data, one of the most widely accepted methodologies is *Cell Suppression*. It is based on the idea of replacing some cell values by missing values. These cells are called *suppressions*. Clearly the cells containing private information (and named *sensitive* cells) must be suppressed, and are called *primary suppressions*. Typically only primary suppressions are not enough to ensure confidentiality on their hidden values due to the existence of marginal cells in the table. Therefore typically other non-sensitive cells should also be determined and suppressed, and are called *secondary suppressions*. Determining which cells in a given table are sensitive (and therefore should be primary suppressions) must be decided before start determining the secondary suppressions. To take this decision there are

J. Domingo-Ferrer and E. Magkos (Eds.): PSD 2010, LNCS 6344, pp. 29–40, 2010.

several common-sense rules. See e.g. Willenborg and De Waal [4]. This paper concerns only the problem of finding the secondary suppressions, which is called *Cell Suppression Problem* (CSP) and is known to be a complex Combinatorial Optimization problem. A pioneer research article with a method to approach this problem is Cox [1]. See e.g. Salazar [3] for a survey on techniques.

Branch-and-cut is a known methodology in Mathematical Programming to approach combinatorial optimization problems. It needs a (mixed) integer (linear) programming model of the problem to be solved, and it is a combination of two standard techniques in Integer Programming: *cutting-plane* and *branch-and-bound* approaches. See e.g. Wolsey [5] for an introduction to both. Branch-and-cut techniques have been extensively and successfully used in the last decades on problems arising from many applications, including in Statistical Disclosure Limitation to protect tabular data.

In this paper we summarize the branch-and-cut approach given in Fischetti and Salazar [2], and then introduce a new variant which is called *cut-and-branch* approach. We compare the different approaches, both theoretically and computationally. Both approaches solve the same optimization problem, thus the optimal objective values are the same. The suppression patterns from both approaches may be different, but a quality comparison between these optimal solutions is meaningless. Indeed, even the same approach could end with a different suppression pattern (when several optimal solutions exist) by changing internal parameters (e.g., the algorithm to solve each linear program, branching rule, preprocessing, etc.).

Our implementations only use free and open-source software, and they have been tested on a set of real-world instances. The discussion and comparison of these approaches are original contributions of this paper to the literature. The paper concludes with a new algorithm for applying cell suppression on tabular data where the response variable is discrete, like in frequency tables. Although protecting frequency tables is fundamental to data stewardship organizations, the algorithms to apply cell suppression in the literature assumes that the response variable of the table is continuous. For that reason, the algorithm described in this paper for frequency tables is also another contribution to the literature.

2 Background and Notation

This section aims to define the main notation that is used along all the paper. For simplicity we give the notation to protect a table against one attacker.

We assume to have a tabular data to be protected. The table consists of a set of cell values and a set of mathematical equations. Let n be the number of cells and m be the number of equations. Let $I = \{1, \ldots, n\}$ be the index set for cells and $J = \{1, \ldots, m\}$ the index set for equations. Let a_i be the nominal value in cell i of the table, for all $i \in I$. Note that a_i is typically obtained by adding the values of one variable (called *response variable*) for all respondents within the features characterizing cell i. As assumed in the literature, the response variable is assumed to be continuous (i.e., a fractional numbers are allowed) and

therefore the table is a so-called *magnitude table*. We will keep this assumption during all this article, except in Section 6 where we adapt the algorithm to work on the alternative assumption in which the response variable is discrete, as in the so-called *counting tables*.

The set of cell values is represented by an array $a = [a_i : i \in I]$. The set of mathematical equations is represented by the linear system $\sum_{i \in I} m_{ij} y_i = b_j$ for all $j \in J$, where m_{ij} and b_j are known numbers, and where y_i are mathematical variables representing cell values. This system of equations defines the table structure (which may be k-dimensional, linked, hierarchical, etc). Note that a_i, m_{ij} and b_j are known by the data stewardship organization, while y_i represent a generic cell value. Note also that the given table a satisfies that equations, i.e. $\sum_{i \in I} m_{ij} a_i = b_j$ for all $j \in J$.

For each cell $i \in I$, let lb_i and ub_i be the external knowledge associated to the cell value i when it is not disclosure. This means that the interval $[lb_i, ub_i]$ represents the a-priori information that an attacker knows on the cell i before analyzing the output table released by the data stewardship organization. Of course, $lb_i \leq a_i \leq ub_i$ for all $i \in I$. An example of potential values are $lb_i = 0$ and $ub_i = +\infty$ for all $i \in I$.

Let P be the subset of cells containing sensitive values. Then at least the cells in P need to be suppressed. Let S be the subset of secondary suppressions. Finding S is the major aim of the Cell Suppression methodology. The scope is to find S such that an attacker cannot get a "too accurate approximation" to any sensitive cells by only considering the values of the cells in $I \setminus \{P \cup S\}$, the external knowledge and the structure of the table.

To better define "too accurate approximation" the data stewardship organization must also set upper and lower protection levels (upl_p and lpl_p respectively) for each sensitive cell $p \in P$. The purpose of these protection levels is the following. From the released output, an attacker will compute the maximum value $\overline{y}_p$ and minimum value $\underline{y}_p$ for each sensitive cell $p \in P$. To this end the attacker will search (in a clever way) through all potential tables congruent with the values of the cells in $I \setminus \{P \cup S\}$, the external knowledge and the structure of the table. To be more precise, the attacker solves the following optimization problems:

$$\left.\begin{array}{ll} \underline{y}_p := \min y_p & \\ \sum_{i \in I} m_{ij} y_i = b_j & \text{for all } j \in J \\ lb_i \leq y_i \leq ub_i & \text{for all } i \in P \cup S \\ y_i = a_i & \text{for all } i \notin P \cup S \end{array}\right\} \tag{1}$$

and

$$\left.\begin{array}{ll} \overline{y}_p := \max y_p & \\ \sum_{i \in I} m_{ij} y_i = b_j & \text{for all } j \in J, \\ lb_i \leq y_i \leq ub_i & \text{for all } i \in P \cup S \\ y_i = a_i & \text{for all } i \notin P \cup S \end{array}\right\} \tag{2}$$

When the assumption is that the response variable of the table is continuous then problems (1) and (2) also include the constraint $y_i \in \mathbb{R}$ for all $i \in I$. When

the assumption is that the response variable is integer (as considered in Section 6) then problems (1) and (2) also include the constraint $y_i \in \mathbb{Z}$ for all $i \in I$.

A set of secondary suppression S *protects* the table when $[\underline{y}_p, \overline{y}_p] \subseteq [lpl_p, upl_p]$ for all $p \in P$. In other words,

- Lower protection level requirement: $\underline{y}_p \leq lpl_p$
- Upper protection level requirement: $\overline{y}_p \geq upl_p$

for all sensitive cell $p \in P$. For brevity we also say that S is a *protected solution*.

Other requirements can also be imposed in the definition of protection, like a $\overline{y}_p - \underline{y}_p \geq spl_p$ for a given parameter spl_p. A different requirement may concern the midpoint of the protected interval $[\overline{y}_p, \underline{y}_p]$ by imposing a constraint like $\overline{y}_p + \underline{y}_p \geq tpl_p$ for a given parameter tpl_p. However, for simplicity, this paper considers only upper and lower protection requirements.

Among all protected solutions, the data stewardship organization prefers the one *minimizing the loss of information*. To have a measure of loss of information of a set S of suppressions, a weight w_i is set by the data stewardship organization to each cell $i \in I$. Then the loss of information of S is defined by $\sum_{i \in S} w_i$. For example, when $w_i = 1$ then the aim of the CSP is to find a protected solution S with the minimum cardinality.

3 Branch-and-Cut Algorithm for Cell Suppression

This section summarizes the exact approach proposed in Fischetti and Salazar [2]. The scheme described in this section is referred in Sections 4 and 6.

Each set of secondary suppressions S can be characterized by a decision variable x_i associated with each cell $i \in I \setminus P$, where $x_i = 1$ if $i \in S$ and $x_i = 0$ otherwise. For simplicity of notation we also use $x_i = 1$ for $i \in P$. Then, for a given set S defined by an array x, the problems (1) and (2) are now:

$$\left.\begin{array}{ll} \underline{y}_p := \min y_p & \\ \sum_{i \in I} m_{ij} y_i = b_j & \text{for all } j \in J \\ a_i - (a_i - lb_i)x_i \leq y_i \leq a_i + (ub_i - a_i)x_i & \text{for all } i \in I \end{array}\right\} \tag{3}$$

and

$$\left.\begin{array}{ll} \overline{y}_p := \max y_p & \\ \sum_{i \in I} m_{ij} y_i = b_j & \text{for all } j \in J \\ a_i - (a_i - lb_i)x_i \leq y_i \leq a_i + (ub_i - a_i)x_i & \text{for all } i \in I \end{array}\right\} \tag{4}$$

The assumption that the response variable of the table is continuous allows to apply Duality Theory and replace (3) and (4) by:

$$\left.\begin{array}{l} \underline{y}_p := a_p - \min \sum_{i \in I} (ub_i - a_i)x_i\alpha_i + (a_i - lb_i)x_i\beta_i \\ \alpha_i - \beta_i + \sum_{j \in J} m_{ij}\gamma_j = \begin{cases} -1 & \text{if } i = p \\ 0 & \text{otherwise} \end{cases} \\ \alpha_i \geq 0, \beta_i \geq 0, \gamma_j \text{ unrestricted in sign } \ \forall i \in I \end{array}\right\} \tag{5}$$

and

$$\left.\begin{array}{c} \overline{y}_p := a_p + \min \sum_{i\in I}(ub_i - a_i)x_i\alpha_i + (a_i - lb_i)x_i\beta_i \\ \alpha_i - \beta_i + \sum_{j\in J} m_{ij}\gamma_j = \begin{cases} +1 & \text{if } i = p \\ 0 & \text{otherwise} \end{cases} \\ \alpha_i \geq 0, \beta_i \geq 0, \gamma_j \text{ unrestricted in sign } \ \forall i \in I. \end{array}\right\} \tag{6}$$

The advantage of this reformulation is that now the protection level requirements can be imposed directly on the x variables through a set of linear constraints. For each sensitive cell $p \in P$ and for each array $\gamma = [\gamma_j : j \in J]$ of real numbers, the following inequalities are necessary for x being a protected solution:

– Lower protection level requirement:

$$\sum_{i\in I} \big((ub_i - a_i)\max\{-\delta_i, 0\} + (a_i - lb_i)\max\{\delta_i, 0\}\big)x_i \geq a_i - lpl_i$$

where

$$\delta_i := \sum_{j\in J} m_{ij}\gamma_j + \begin{cases} 1 & \text{if } i = p \\ 0 & \text{otherwise} \end{cases}$$

for all $i \in I$.

– Upper protection level requirement:

$$\sum_{i\in I} \big((ub_i - a_i)\max\{-\delta_i, 0\} + (a_i - lb_i)\max\{\delta_i, 0\}\big)x_i \geq upl_i - a_i$$

where

$$\delta_i := \sum_{j\in J} m_{ij}\gamma_j - \begin{cases} 1 & \text{if } i = p \\ 0 & \text{otherwise} \end{cases}$$

for all $i \in I$.

For simplicity of notation each of these inequalities is represented in this paper by

$$\sum_{i\in I} c_i x_i \geq c_0.$$

Since $c_i \geq 0$ and $x_i \in \{0, 1\}$ for all $i \in I$, one can strengthen this constraint by replacing c_i with $\min\{c_i, c_0\}$. Because the inequality makes sense only when $c_0 > 0$, we can also divide all the coefficients by c_0 and works only with inequalities of type:

$$\sum_{i\in I} d_i x_i \geq 1. \tag{7}$$

These stronger inequalities, derived from the upper and lower protection level requirements as indicated above, are called *capacity constraint*.

Although there is an infinity number of constraints (7), only a finite number of them are enough to impose all of them. What is more, only some of them may be useful to protect a given table. The problem of checking whether a given

solution x^* violates a constraint in (7) or not is known as *separation problem* of (7). A procedure solve the separation problem is called *separation procedure*. A separation procedure for inequalities (7) consists of solving the linear programs (3) and (4) defined by the given x^*. If a protection level requirement is violated then the dual solution γ^* determines a useful constraint (7). Otherwise x^* satisfies all inequalities in (7) without needing the explicit evaluation of each one. Observe that this procedure works also when x^* is not an integer array. For that reason, a branch-and-cut approach is the natural framework to solve the CSP. We now briefly summarize how this branch-and-cut approach works:

1. Initial heuristic: Let x' be a feasible CSP solution and z' be its objective value.
2. Create a list $\mathcal{L}$ of linear problems to be solved with the following linear problem:

$$\begin{aligned} z = \min \sum_{i \in I} w_i x_i & \\ x_i = 1 & \quad \text{for all } i \in P \\ 0 \le x_i \le 1 & \quad \text{for all } i \in I \setminus P. \end{aligned}$$

3. If $\mathcal{L} = \emptyset$ then x' is an optimal CSP solution. Otherwise, extract a linear program from $\mathcal{L}$.
4. Solve the linear program. Let x^* be an (integer or fractional) optimal solution and z^* be its objective value.
5. Primal heuristic: Create (if possible) a feasible CSP solution x" from x^*. If z" is the objective value of x" and $z" < z'$ then update x' and z'.
6. Bounding: If $z^* \ge z'$ go to Step 3.
7. Solve (3) and (4) defined by x^* to check whether it is a protected solution or not.
8. Cutting: If x^* is not protected, build violated capacity constraints (7) by using the optimal dual solution γ^* of (3) and/or (4). Add these constraints to the current linear program. Go to Step 3.
9. If x^* is protected and integer, it is a feasible CSP solution. Update x' and z' if $z^* < z'$. Go to Step 3.
10. Branching: If x^* is protected and fractional, select a variable x_i such that x_i^* is not integer, and add two new linear programs to $\mathcal{L}$. One linear program is a copy of the current linear program and the constraint $x_i = 0$. The other linear program is also a copy of the current linear program and the constraint $x_i = 1$. Go to Step 3.

Note that there are two optimization problems involved in the above algorithm. One is the so-called *master problem*, which is an integer program on the x variables. Another is the so-called *subproblem*, which consists of the linear programs (3) and (4) on the y variables. While the linear programs of a subproblem can be solved by using a black-box linear programming solver, the master problem is solved through an (implicit) enumerative approach which solve a linear program at each iteration.

4 Cut-and-Branch Algorithm for Cell Suppression

This section introduces a new approach to solve the CSP. This problem consists of finding a vector x minimizing $\sum_{i \in I} w_i x_i$ and such that

- x must be an integer vector, and
- x must represent a protected solution.

The algorithm described in Section 3 looks for these two properties within two loops:

- an external loop aims integrability by exploring an enumerative list of linear programs;
- an internal loop imposes protection by adding violated inequalities to the current linear program.

The internal loop is defined by Steps 4 to 9 in the algorithm described in Section 3. The external loop is defined by Steps 3 and 10.

As an alternative approach, one could reverse the order of the loops. The internal loop ensures integrability of x. The external loop checks the protection of integer vectors x, and must generate constraints violated by x when this integer array does not correspond to a protected pattern. Since the order of the loops has been reversed, this alternative approach is called *cut-and-branch*.

Note that in a cut-and-branch approach a primal heuristic (Step 5) that potentially build a feasible CSP solution from an unprotected integer vector x^* is crucial. Otherwise the initial heuristic CSP solution (Step 1) is not improved unless the algorithm ends with a proof of optimality. In a branch-and-cut approach a primal heuristic may be less relevant because Step 9 may improve the CSP solution generated in Step 1. Another important observation is that the integer program solved at each iteration of the cut-and-branch approach is a relaxed problem of CSP. Hence, at each iteration the integer program must be solved to optimality in order to ensure that the final integer solution is optimal for the CSP.

This branch-and-cut approach is based on eliminating non-protected integer solutions x^* by using the family of inequalities (7). Different families of inequalities may lead to different cut-and-branch implementations. Let us illustrate this claim with two examples.

Consider an integer solution x^* which corresponds to a subset S^* not satisfying all protection level requirements. Let $z^* = \sum_{i \in I} w_i x_i^*$. A simple way of eliminating the infeasible solution x^* is by adding the inequality:

$$\sum_{i \in I} w_i x_i \geq z^* + \epsilon$$

where $\epsilon > 0$ is a small but fixed parameter. When w_i (for all $i \in I$) are integer numbers then one could set $\epsilon = 1$. Unfortunately it is known in Mathematical Programming that this type of cutting planes are very weak, and typically an algorithm based on this type of cuts would not work on medium-size instances.

Another alternative way of eliminating the infeasible solution x^* arises by observing that any subset of S^* cannot be a protected solution. In other words, any protected solution contains at least a cell in $I \setminus S^*$. This constraint can be mathematically written by the inequality:

$$\sum_{i \in I \setminus S^*} x_i \geq 1. \tag{8}$$

The family of constraints (8) for all S^* unprotected is sufficient to guarantee the exactness of the algorithm when $w_i \geq 0$. Indeed, each constraint eliminates all infeasible patterns which are subset of un unprotected pattern S^*.

Inequalities similar to (8) appear when solving combinatorial optimization problems from many other applications. They are called *cover inequalities*; see e.g. Wolsey [5]. A cover inequality is stronger when the subset $I \setminus S^*$ is minimal, i.e., when adding a cell to S^* creates a larger set which is a protected pattern. The process of reducing some coefficients in the left-hand-side of inequality (8) is called *lifting procedure* and requires solving the linear programs (3) and (4).

An advantage of using (8) is the simplicity of the separation procedure. It does not need optimal dual solutions of the linear programs (3) and/or (4). This advantage is further exploited in Section 6. A disadvantage, when the cover inequalities are compared to the capacity constraints, is that the cover inequality (8) is not valid when x^* is a non-integer vector. Another disadvantage is that only one inequality (8) is associated with a non-protected integer vector x^*. This disadvantage, however, may be avoid if one uses a lifting procedure since using different sequences to lift the coefficients may induce different lifted inequalities. Still the computational effort of lifting could drastically decrease the effectiveness of the separation procedure. Preliminary computational results confirmed this claim. Even more, also applying both the separation procedures of (7) and (8) did not improved in practice over the cut-and-branch algorithm with only (7). For that reason, in Section 5 we use (7) and not (8). Still, Section 6 points out a situation where constraints (8) are relevant.

5 Computational Results

We have implemented the two approaches introduced in Sections 3 and 4: branch-and-cut and cut-and-branch. The implementation was done on a computer Dell Precision T5400 with Intel Xeon X5460 3.16GHz, and using JAVA programming language and Eclipse [6], a free and open source software (FOSS) for JAVA developments. To solve mathematical programming models we have used GLPK 4.43 [7], which is also FOSS. These options have been selected for the immediate portability of the implementations to different computer platforms and for ensuring the FOSS feature of the final code. These are well-appreciated features in software to protect tables for data stewardship organizations.

It is known that GLPK is far from being competitive in efficiency with other mathematical programming solvers. However, the scope of this section is not to show the performance of the fastest implementation of each algorithm. The

scope of this section is to compare performance of different implementations done under the same conditions. We believe that, by properly scaling the computational times of each run, one can extract similar conclusions if another computer, programming language and mathematical programming solver were used.

The initial heuristic procedure (Step 1) is the one proposed in ([2]), used as starting feasible CSP solution for both the branch-and-cut and the cut-and-branch algorithms. We did not implemented any ad-hoc primal heuristic procedure (Step 5), but we activated the feasibility pump procedure available in GLPK. Note that this procedure is a primal heuristic procedure for the branch-and-cut implementation but not for the cut-and-branch implementation. Indeed, when our branch-and-cut implementation ends with the time limit, the only feasible CSP solution is the one generated by the initial heuristic.

To compare our implementations we are using a collection of real-world tables created by the Incoming Tax Department of the Spanish Ministry of Finance ("Agencia Tributaria, Ministerio de Economía y Hacienda"). This collection contains 157 tables extracted from the 2008 IRPF taxes. The number of cells is between 138 and 570, and the number of equations is between 42 and 310. Protecting these real-world tables was the original motivation of the research contained in this paper. Unfortunately, for confidentiality issues, this collection of instances is not publicly available. There is another collection of instances which is publicly available through the website `http://webpages.ull.es/users/casc`. However, our implementations based on a FOSS mathematical programming tool were unable to deal with most of these instances due to the larger number of cells and equations. For that reason the analysis in this paper is based on running our implementations only on the collection motivating this research.

Over the 157 tables, there are 9 tables without primary suppressions, thus they do not have an associated CSP. Among the remaining 148 tables, there are 15 tables that could not be solved to optimality by the branch-and-cut implementation within a time limit of 1 hour. Table 1 gives the average (av) and the standard deviation (sd) over the 133 tables solved to optimality by the branch-and-cut implementation:

$|I|$: Number n of cells in the table.
$|J|$: Number m of equations in the table.
$|P|$: Number of sensitive cells (i.e., the number of primary suppressions).
z': Objective value of the initial heuristic CSP solution (Step 1).
time': Number of seconds to compute the initial heuristic CSP solution.
sep: Number of solutions x^* that have been checked, and potentially some capacity constraints (7) have been generated (Step 7).
cuts: Number of violated capacity constraints (7) generated.
nodes: Number of calls to the branching procedure (Step 10).
time: Number of seconds required by the branch-and-cut implementation.
z_0: Optimal objective value of the first linear program solved (Step 2).
z_1: Optimal objective value of the last linear program solved before branching (i.e., the lower bound at the root node).
opt: Objective value of the optimal CSP solution.

Table 1. Average (av) and standard deviation (sd) on 133 tables

	instance			heuristic		branch-and-cut						
	$\|I\|$	$\|J\|$	$\|P\|$	z'	time'	sep	cuts	nodes	time	z_0	z_1	opt
av	485.3	258.3	73.5	25032.7	6.0	2539.8	5428.6	431.8	179.2	13316.0	22361.6	22431.0
sd	86.7	54.1	63.2	72956.1	5.6	8880.5	19027.7	1299.8	566.2	54718.1	71187.9	71191.6

The branch-and-cut implementation was able to solve 107 instances without branching. This means that in most of the cases, the external loop was not necessary to achieve integrability. Over the 133 tables solved to optimality, the gap before branching is 0.1%. Over the 148 tables with primary suppressions, the gap before branching is 0.5%. These percentages show a good quality of the lower bound z^* from the linear program before branching, which is a fundamental feature to the success of a branch-and-cut implementation. Of course, it also shows good quality of the upper bound z' provided by the initial heuristic (Step 1). Over the 15 tables where the branch-and-cut implementation was not able to conclude optimality within 1 hour, the gap before branching is 6.5%. The quality of this lower bound is mainly due to the capacity constraints (7), but it also due to the activation of the additional inequalities that are automatically generated inside GLPK. These are mainly Gomory inequalities, and have contributed to close the gap in 5% on average.

Over the 133 tables solved to optimality, the average time to protect a table is 2 minutes, and the worse case is 10 minutes. This is a very satisfactory behavior of an exact approach to solve CSP for a data stewardship organization desiring a free and open source implementation to protect a table. There are 15 tables which could not be solved by this implementation within 1 hour, but even in these few cases the quality of the best feasible CSP solution was quite satisfactory.

Over the 148 tables, 117 CSP instances are solved by cut-and-branch to optimality before the time limit of 1 hour. The cut-and-branch implementation was also faster than the branch-and-cut implementation on 62 tables. This is a very relevant observation. Even more, there are 2 instances solved to optimality by the cut-and-branch implementation and not solved by the branch-and-cut implementation. Details on these two instances are given in Table 2. The meaning of the column coincides with the given for Table 1, except that now they represent individual values and not average values. In addition we find the following columns:

z": Objective value of the best feasible CSP solution when the branch-and-cut ended with the time limit.
iter: Number of steps of the external loop by the cut-and-branch implementation, i.e., the number of integer programs solved.
cuts': Number of capacity constraints (7) generated by the cut-and-branch approach.
time": Number of seconds required by the cut-and-branch implementation.

Table 2. Details on two tables

instance				heuristic		branch-and-cut						cut-and-branch			
id	$\|I\|$	$\|J\|$	$\|P\|$	z'	time'	sep	cuts	nodes	time	z_1	z"	iter	cuts'	time"	opt
738	425	232	74	5770	7	17831	2874	6452	3600	5157.6	5634	92	1245	493	5597
746	553	299	40	3517	4	42896	120228	10028	3600	2987.8	3431	38	682	134	3266

We are reporting results where the cut-and-branch implementation uses capacity constraints and not cover inequalities. The reason for that is because, when using both families of inequalities, in 3 instances the implementation reduced the total time in more than 10 seconds and in 22 instances the implementation increased the total time in more than 10 seconds. On average the implementation saved 34 seconds by only using capacity constraints (7)). For that reason (and because in our collection the response variable may assume any continuous value) we deactivated (8) in our cut-and-branch implementation.

Over the 148 tables, both the branch-and-cut and the cut-and-branch implementations solved to optimality 105 instances. The branch-and-cut was faster than the cut-and-branch on 22 instances and the time reduction was 64%. The cut-and-branch was faster than the branch-and-cut on 59 instances and the time reduction was 69%. For the remaining 24 instances the time difference was smaller than one second.

Over the 117 instances solved to optimality by the cut-and-branch implementation, the average number of iterations of the external loop was 38.9, the average number of capacity cuts generated in the whole approach was 293, and the average number of seconds to end with optimality proof was 121 seconds. Over the 31 instances not solved by the cut-and-branch implementation, the average number of iterations of the external loop was 273.3, the average number of capacity cuts generated in the whole approach was 1414.2, and the average gap between the initial CSP solution and the last unprotected solution was 14%.

6 An Exact Algorithm for Protecting Counting Tables

In this section we replace the assumption that the response variable defining the tabular data may assume any floating-point number. Now instead we assume that a cell value may only be an integer number. This hypothesis occurs when dealing with, for example, counting or frequency tables, i.e., tables where each cell displays the number of respondents within the cell features. Even more, in many magnitude tables it is meaningless that a cell contains a floating-point number with any fractional part. On the contrary, a cell value may only have a fractional part with one digit. This section describe an algorithm that is necessary to deal with this type of non-continuous variables.

The major impact of the new assumption is that models (1) and (2) now needs the extra requirements that y_i is integer for all $i \in I$. Then, the models are not (continuous) linear programs and therefore the Duality Theory is not applicable. More precisely, given an (integer or fractional) array $x^* \in [0,1]^I$,

capacity constraints (7) are necessary but may not be sufficient to ensure all protection level requirements. However, inequality (8) is still valid whenever $x^* \in \{0,1\}^I$ is unprotected. For that reason, the cut-and-branch approach in Section 4 is an exact algorithm to protect a counting table if it uses (8). Note that one cannot get dual optimal solutions γ from the integer programs:

$$\left.\begin{array}{cc} \underline{y}_p := \min y_p & \\ \sum_{i \in I} m_{ij} y_i = b_j & \text{for all } j \in J \\ a_i - (a_i - lb_i)x_i \leq y_i \leq a_i + (ub_i - a_i)x_i & \text{for all } i \in I \\ y_i \in \mathbb{Z} & \text{for all } i \in I \end{array}\right\} \quad (9)$$

and

$$\left.\begin{array}{cc} \overline{y}_p := \max y_p & \\ \sum_{i \in I} m_{ij} y_i = b_j & \text{for all } j \in J \\ a_i - (a_i - lb_i)x_i \leq y_i \leq a_i + (ub_i - a_i)x_i & \text{for all } i \in I \\ y_i \in \mathbb{Z} & \text{for all } i \in I \end{array}\right\} \quad (10)$$

Still, it is possible to slightly modify the branch-and-cut algorithm in Section 3 to also deal with counting tables. Of course the heuristic procedures in Step 1 and 5 need to be modified to deal with the new assumption. Also Step 8 needs a modification: When x^* is integer and protected according to the linear programming objective values, one must solve the integer programs (9) and (10). (Note that in Step 7 only the linear programming relaxation of (9) and (10) were solved.) If x^* is unprotected with the optimal objective values of the integer programs then add constraint (8) to the current linear program and go to Step 3; otherwise, apply Step 8.

Acknowledgements

The author thanks David Pérez and Diego Porras for writing in JAVA the algorithms described in this paper, and "Agencia Tributaria" of the Spanish Government for motivating and supporting this research. This work has also been supported by "Ministerio de Ciencia e Innovación" (MTM2009-14039-C06-01).

References

1. Cox, L.: Suppression Methodology and Statistical Disclosure Control. Journal of the American Statistical Association 75(370), 377–385 (1980)
2. Fischetti, M., Salazar, J.J.: Solving the Cell Suppression Problem on Tabular Data with Linear Constraints. Management Science 47(7), 1008–1026 (2001)
3. Salazar, J.J.: Statistical confidentiality: Optimization techniques to protect tables. Computers & Operations Research 35, 1638–1651 (2008)
4. Willenborg, L.C.R.J., De Waal, T.: Elements of Statistical Disclosure Control. Lecture Notes in Statistics, vol. 155. Springer, Heidelberg (2001)
5. Wolsey, L.: Integer Programming. Series in Discrete Mathematics and Optimization. Wiley, Chichester (1998)
6. http://www.eclipse.org/
7. http://www.gnu.org/software/glpk/

Data Swapping for Protecting Census Tables

Natalie Shlomo[1], Caroline Tudor[2], and Paul Groom[2]

[1] Southampton Statistical Sciences Research Institute, University of Southampton, Highfield, Southampton SO17 1BJ, United Kingdom
N.Shlomo@soton.ac.uk
[2] Office for National Statistics, Segensworth Road, Titchfield, Fareham, PO15 5RR, United Kingdom
{Caroline.Tudor,Paul.Groom}@ons.gsi.gov.uk

Abstract. The pre-tabular statistical disclosure control (SDC) method of data swapping is the preferred method for protecting Census tabular data in some National Statistical Institutes, including the United States and Great Britain. A pre-tabular SDC method has the advantage that it only needs to be carried out once on the microdata and all tables released (under the conditions of the output strategies, eg. fixed categories of variables, minimum cell size and population thresholds) are considered protected. In this paper, we propose a method for targeted data swapping. The method involves a probability proportional to size selection strategy of high risk households for data swapping. The selected households are then paired with other households having the same control variables. In addition, the distance between paired households is determined by the level of risk with respect to the geographical hierarchies. The strategy is compared to a random data swapping strategy in terms of the disclosure risk and data utility.

Keywords: Targeted Data Swap, Random Data Swap, Disclosure risk, Data Utility, R-U confidentiality map.

1 Introduction

Protecting tables containing Census counts is more difficult than protecting tabular data from a survey sample since sampling a priori introduces ambiguity into the frequency counts. More invasive statistical disclosure control (SDC) methods are needed to protect against disclosure risks in a Census context where tables include whole population counts and this impacts negatively on the utility of the data. It is well known that Census data have errors due to data processing, coverage adjustments, imputations for non-response and edit and imputation procedures, although much effort is devoted to minimizing these errors. When assessing disclosure risk, it is essential to take into account these errors and the protection that is already inherent in the data. A quantitative measure of disclosure risk should consider for example the amount of imputation and adjust parameters of the SDC methods to be inversely proportional to the imputation rate. This ensures that the data is not overly protected causing unnecessary loss of information. It should be noted

J. Domingo-Ferrer and E. Magkos (Eds.): PSD 2010, LNCS 6344, pp. 41–51, 2010.

that once Census results are disseminated, they are typically perceived and used by the user community as accurate counts.

The main disclosure risk in a Census context comes from small counts in the tables, i.e. ones and twos, since these can lead to re-identification. Indeed, the amount and placement of the zeros in the table determines whether new information can be learnt about an individual or a group of individuals. Therefore, SDC methods for Census tabular data should not only protect small cells in the tables but also introduce ambiguity and uncertainty into the zero values.

SDC methods for protecting Census tables that are typically implemented at National Statistical Institutes (NSI) include pre-tabular methods, post-tabular methods and combinations of both. In this paper we focus on a pre-tabular method which is implemented on the microdata prior to the tabulation of the data. The most commonly used method is data swapping between a pair of households matching on some control variables (Willenborg and de Waal, 2001). This method has been used for protecting Census tables at the United States Bureau of the Census and the Office for National Statistics (ONS) in the United Kingdom. Data swapping can be seen as a special case of a more general pre-tabular method based on a Post-Randomization Method (PRAM) (Gouweleeuw, Kooiman, Willenborg and De Wolf, 1998). This method adds "noise" to categorical variables by changing values of categories for a small number of records according to a prescribed probability matrix and a stochastic process based on the outcome of a random multinomial draw. PRAM can also be carried out in such a way as to ensure marginal distributions and because it is a stochastic perturbation, users can make use of the probability transition matrix to adjust their statistical analysis. This method however has yet to be implemented for a large scale Census. In practice, NSIs prefer data swapping since the method is easy to implement and marginal distributions are preserved exactly on higher aggregations of the data. It should be noted that NSIs do not typically release parameters of the SDC methods, i.e. swapping rates or probability transition matrices, in order to minimize the risk of deciphering the perturbation process.

In this paper, we propose a data swapping strategy that is targeted to high risk households. In addition, the distance between paired households for carrying out the swap is determined by the geographical level that is most at risk as defined by unique cells on margins of key variables (Young, Martin and Skinner, 2009). The targeted data swapping strategy is compared to a random data swapping strategy through quantitative disclosure risk and data utility measures according to the disclosure risk–data utility framework as described in Willenborg and De Waal (2001), Duncan, Keller-McNulty, and Stokes (2001) and Shlomo (2007). The data utility is assessed by analyzing the impact of the data swapping strategies on chi-square tests for independence as well as measuring distortions to cell counts for specified Census tables. Disclosure risk is assessed by the proportion of unique cells that are not perturbed in the tables. The analysis will be demonstrated on a real data set extracted from the UK 2001 Census.

Section 2 outlines the data swapping methods that are assessed and Section 3 details the data and Census tables that are used in the analysis. Section 4 presents the results of the comparison between the swapping methods followed by a discussion and conclusions in Section 5.

2 Data Swapping Methods

The most common pre-tabular method of SDC for Census tables is data swapping on the microdata prior to tabulation where values of variables are exchanged between pairs of households. In order to minimize bias, pairs of households are determined within strata defined by control variables, such as a large geographical area, household size and the age sex distribution of the individuals in the households. Data swapping can be targeted to high risk households found in small cells of tables as described in Section 2.1 thereby ensuring that households that are most at risk for disclosure are more likely to be swapped.

In a Census context, geography variables are often swapped between households for the following reasons:

- Given household characteristics, other Census variables are likely to be independent of geography and therefore it is assumed that less bias will be induced. In addition, because of the conditional independence assumption, swapping geography will not necessarily result in inconsistent and illogical records. In contrast, swapping a variable such as age would result in many inconsistencies with other Census variables, such as marital status and education level.
- At a higher geographical level and within control strata, the marginal distributions are preserved.
- The level of protection increases by swapping variables which are highly "matchable" such as geography.
- There is some protection for disclosure risk arising from differencing two tables with nested geographies since data swapping introduces ambiguity into the true cell counts and in particular the zero counts.

2.1 Targeted Data Swap Strategy

Targeted data swapping is based on an allocation of a $p\%$ sample of households where p is the swapping rate to be determined by the NSI. Typically the data swapping is carried out within blocks of large geographical areas, eg., Estimation Area or Census Delivery Group Area. Within these large areas are hierarchies of geographies. For the UK 2001 Census data, there are three layers of nested geographies: Local Authority (LA), Wards and Output Areas (OA).

Census tables contain counts of individuals so to identify high risk households we need to first identify high risk individuals. High risk individuals are defined on the basis of frequency counts of univariate distributions on a set of key variables that are typically used to span Census tables at different levels of geography. A cell of size one on the univariate distribution means that there will be a unique individual on one of the margins of the table. Uniques on the margin of a table increase the risk of attribute disclosure since individuals can be identified on the basis of some of the variables spanning the table, and once identified, a new attribute can be learnt. High risk is defined through a score that is calculated for each individual as follows:

- Calculate frequency counts for M selected key variables each having k_m, $(m = 1,...,M)$ categories at the geographical level g: $N^g_{k_m}$ (not including individuals that have been imputed to account for the Census under-coverage).
- For every individual with values of categories $k = (k_1, k_2, ..., k_M)$, calculate a score at each level of geography g by taking the average of the reciprocals of the counts: $HR^g_k = (\sum_{m=1}^{M} 1 / N^g_{k_m}) / M$.
- A threshold is set for each level of geography and those scores above the thresholds determine high risk individuals.
- High risk households are defined as any household having at least one high risk individual.

The overall sample size in a Delivery Group area is calculated by multiplying the swap rate by the number of non-imputed households. This sample is then allocated across the lower level geographies (eg., OAs). We propose using two proportional allocations according to:

(1) the inverse number of (non-imputed) households in the OA, i.e. the larger the OA the less swapping required,
(2) the percentage of high risk households in the OA.

The final sample size for each OA is taken as the average sample size across the two proportional allocations provided that the final sample size is not over 20% of the number of (non-imputed) households in the OA. The random sample of households is drawn within each OA using a probability proportional to size (pps) design according to the above allocation. The size variable for the pps sampling is calculated so that a high value is given to high risk households and a low value is given to low risk households. This ensures that a disproportionate number of households at high risk will be selected in the sample and at the same time, guarantees that households of low risk will have a small but positive chance of being selected in the sample in order to introduce some randomness into the data swapping.

After the sample is selected, each of the households must be paired with another household in order to swap the geographical variables. The paired household must match on a set of control variables, eg. household size, age group and sex distribution, ethnicity indicator, 'hard to count' index. For the targeted swapping strategy we introduce the notion of distance swapping as defined in Young, Martin and Skinner, 2009. The idea is to pair households for swapping at a distance that is consistent with the geographical level of disclosure risk. Similar to the method carried out for defining high risk individuals, we first calculate the univariate distribution frequencies of individuals for the key variables at each geographical level. If there is a unique individual on any of the categories of the key variables at a geographical level g, the individual is flagged for that geographical level. The household geographical disclosure risk level is then defined as the highest geographical risk level from among all individuals in the household. For example, if there is an individual in a household that is flagged as being unique on one of the categories of the key variables at the ward level and another individual in the same household that is flagged at the LA level, the entire household is flagged at the LA

level of disclosure risk. The geographical level of disclosure risk is used in the algorithm for swapping as described below:

For the selected household to be swapped, we first check the level of geographical disclosure risk and choose a paired household at the appropriate geography having the same control variables. For example, if the level of geographical disclosure risk is flagged at LA, then the household must be swapped with a similar household having the same control variables in a different LA but within the large Delivery Group Area. If the level of geographical disclosure risk is flagged at ward, then the household must be swapped with another household having the same control variables in a different ward but within the same LA. If the level of geographical disclosure risk is flagged at OA, then the household must be swapped with another household having the same control variables in a different OA but within the same ward. Therefore, selected sampled households are swapped with other households having the same control variables but only at a distance that is appropriate to the geographical level of risk with respect to the uniqueness on marginal distributions of the key variables. The advantage of 'localized' data swapping is that we minimize the distance between pairs of households depending on the geographical level of risk and therefore at higher aggregations of geography we expect less distortion.

The search for a paired household in the swapping algorithm is carried out through several iterations. In the first attempt, the sampled household must match the paired household on a full set of control variables, eg. 'hard to count' index, household size, ethnicity indicator, sex and broad age distribution in the household. In subsequent attempts to search for a paired household, control variables undergo gradual collapsing to allow a better chance of finding a pair for the sampled household. Note that no household can be paired twice. Once a household is selected, all geographical variables are swapped between the two households.

2.2 Random Data Swap Strategy

We compare the targeted data swapping strategy in Section 2.1 with a random data swapping strategy. The same swapping rates are used for both strategies. The difference between the strategies is that households are selected for swapping using a simple random sample without replacement design in each OA, i.e. all households have equal chance of being selected for swapping. The sampled household is then paired with another household having the same control variables using the iterative procedure described above but no attempt is made to target high risk households or control the distance between swapped households.

3 Data

For this analysis, targeted and random data swapping strategies described in Section 2 were carried out on households from an extract of the 2001 UK Census containing two LAs at the following swapping rates: 2% and 5%. The extract included 327,718 individuals in 124,979 households. In the two LAs there were 35 wards and 1,111 OAs.

We define the following Census tables of individuals at the lower level of geography OA, where the number of categories are given in parenthesis:

(1) Religion(9) × Age-Sex(6) × OA

(2) Travel to Work(12) × Age-Sex(12) × OA

(3) Ethnicity (17) × Sex(2) × OA

(4) Country of Birth (17) × Sex (2) × OA

(5) Economic Activity (9) × Sex (2) × Long-Term Illness (2) × OA

(6) Health status (5) × Age-Sex (14) × OA

The characteristics of the five tables are presented in Table 1. As can be seen, the tables have different average cell sizes and distributions of small cells. We also produce the same Census tables (1) to (6) at the ward level geography.

Table 1. Characteristics of Census tables at the OA geography

	Table 1	Table 2	Table 3	Table 4	Table 5	Table 6
Number of Individuals	327718	240,797	327,718	327,718	238,727	325,594
Number of internal cells	59,994	159,984	37,774	37,774	39,996	77,770
Average cell size	5.46	1.51	8.68	8.68	5.97	4.19
Number of zeros	34,546 (57.6%)	103,361 (64.6%)	23,939 (63.4%)	19,723 (52.2%)	12,697 (31.7%)	40,363 (51.9%)
Number of 1s	5,298 (8.8%)	20,793 (13.0%)	5,468 (14.5%)	7,329 (19.4%)	6,634 (16.6%)	11,260 (14.5%)
Number of 2s	2,771 (4.6%)	10,304 (6.4%)	2,607 (6.9%)	3,767 (10.0%)	4,511 (11.3%)	6,183 (8.0%)

4 Analysis

To compare the two data swapping strategies described in Section 2, we assess disclosure risk in terms of the proportion of unswapped unique cells in the Census tables described in Section 3, and data utility in terms of distortions to distributions and statistical inference.

4.1 Disclosure Risk

Disclosure risk arises from small cells in tables (or small cells appearing in potential slithers of differenced tables). In addition, the number and placement of zero cells can lead to identification and attribute disclosure when many tables are disseminated from one database.

Pre-tabular methods of disclosure control, and in particular data swapping, will not inhibit small cells from appearing in tables and therefore a quantitative disclosure risk measure is needed which reflects whether the small cells in tables are true values.

The quantitative disclosure risk measure for assessing the impact of data swapping is the proportion of cells of size one that have not been perturbed. This is calculated by counting the number of cells that have both an original and perturbed count of one divided by the number of cells with an original count of one. Let T^O represent the original table and let $T^O(c)$ be the cell frequency c in table T^O. Similarly, T^P represents the swapped table. The risk measure is defined as:

$$DR = \frac{\sum_c I(T^O(c) = 1, T^P(c) = 1)}{\sum_c I(T^O(c) = 1)},$$

where I is the indicator function receiving a value of 1 if it is true and 0 otherwise. Note that we ignore those individuals that have been imputed to adjust for the Census under-coverage since these are not considered at risk. In Table 2 we present the *DR* proportions for the Census tables described in Section 3. We also present the *DR* in Table 3 for smaller Census tables defined by the main marginal variable crossed with the OA geography.

In Tables 2 and 3, we see that the higher swapping rate protects more unique cells than the lower swapping rate. The random swapping has higher proportions of unique cells that are unperturbed than the targeted swapping. These results are as expected. The overall disclosure risk in some Census tables is high, even for the 5% data swapping, with over 80% of unique cells unperturbed. In addition, there are two clear patterns in Tables 2 and 3. All Census tables have the highest disclosure risk at the 2% random swap and the lowest disclosure risk at the 5% targeted swap. For some tables, the 2% targeted swap provides lower disclosure risk than the 5% random swap, for example in Tables (1), (3) and (4). The reason for this pattern is that the marginal

Table 2. Proportion of unperturbed unique cells (*DR)* in the Census tables

Table	Random 2%	Random 5%	Target 2%	Target 5%
(1)	0.939	0.853	0.749	0.650
(2)	0.932	0.837	0.912	0.822
(3)	0.944	0.848	0.549	0.457
(4)	0.924	0.831	0.723	0.629
(5)	0.910	0.805	0.894	0.779
(6)	0.929	0.828	0.925	0.819

Table 3. Proportion of unperturbed unique cells *(DR)* on the margins of the Census tables

Table	Random 2%	Random 5%	Target 2%	Target 5%
Religion×OA	0.954	0.851	0.595	0.495
Travel to work ×OA	0.899	0.826	0.912	0.800
Ethnicity ×OA	0.934	0.832	0.421	0.347
Country of birth ×OA	0.916	0.803	0.590	0.518
Economic activity×OA	0.823	0.668	0.816	0.584
Health status ×OA	0.811	0.684	0.821	0.737

distributions of religion, ethnicity and country of birth (as well as age-sex distribution) were used to define high risk households which according to the pps sampling had more chance of being selected into the sample for swapping. It is clear that the disclosure risk based on variables that are used to define high risk households would be reduced considerably under the targeted data swapping approach.

4.2 Data Utility

Data utility measures used in this analysis are based on a distance metric to measure the distortion to distributions and the impact on a measure of association based on a statistical test for independence between categorical variables.

Some useful distance metrics were presented in Gomatam and Karr (2003). The distance metric that is used in this analysis is the average absolute distance per cell of a Census table calculated as: $AD(T^O, T^P) = \sum_c |T^P(c) - T^O(c)| / n_T$ where n_T is the number of cells in the Census table.

Table 4 presents results of the distance metric *AD* for the Census tables defined in Section 3 at the OA geography and Table 5 the same distance metric *AD* for the Census tables defined at the ward geography. The aim is to show that at higher aggregations of geography, the targeted data swapping strategy obtains less distortion, i.e. smaller distance metrics, compared to the random data swapping at a given swapping rate because of the 'localized' search for the household pair.

In Table 4 the highest *AD* metric representing the most distortion in cell counts according to the OA geographical level is obtained under the targeted 5% swap and the lowest *AD* metric is obtained under the random 2% swap. The random swapping strategy has lower *AD* metrics than the targeted swapping strategy which means that more bias is introduced into the Census tables at the OA geography due to the

Table 4. Average absolute distance per cell *(AD)* for Census tables with OA geography

Tables (OA)	Random 2%	Random 5%	Target 2%	Target 5%
(1)	0.391	0.732	0.499	0.841
(2)	0.208	0.427	0.238	0.455
(3)	0.266	0.523	0.665	0.916
(4)	0.261	0.496	0.486	0.713
(5)	0.294	0.577	0.338	0.621
(6)	0.272	0.526	0.290	0.555

Table 5. Average absolute distance per cell *(AD)* for Census tables with ward geography

Tables (wards)	Random 2%	Random 5%	Target 2%	Target 5%
(1)	1.627	2.754	1.227	1.547
(2)	1.141	2.034	0.559	0.678
(3)	1.273	2.260	1.708	2.219
(4)	1.567	2.677	1.133	1.528
(5)	2.055	3.468	1.081	1.366
(6)	1.664	2.781	0.822	1.035

targeted selection of households to swap. For most of the Census tables, the targeted 2% swap has less distortion to the cell counts compared to the random 5% swap, with the exception of Census table (3) involving the variable ethnicity. Ethnicity in particular was used for the targeted data swapping strategy for defining high risk and also as an indicator in the control variables for selecting paired households. This likely induced more bias into the table.

Table 5, however, presents a different picture for the Census tables at the aggregated ward geography level. Obviously, the disclosure risk is considerably less when aggregating to the ward level with less possibility of unique cells. The random data swapping shows more distortions per cell than the targeted data swapping for each of the swapping rates. This clearly demonstrates that taking into account the geographical level of risk when pairing households for swapping as implemented in the targeted data swapping strategy ensures much less bias at higher aggregations of geographies.

A very important statistical tool that is frequently carried out on contingency tables is the Chi-Square test for independence based on the Pearson Chi-Squared Statistic χ^2 which tests the null hypothesis that the criteria of classification, when applied to a population, are independent. The Pearson Statistic for a two-dimensional table is defined as: $\chi^2 = \sum_i \sum_j (o_{ij} - e_{ij})^2 / e_{ij}$ where under the null hypothesis of independence: $e_{ij} = (n_{i.} \times n_{.j}) / n$, $n_{i.}$ is the marginal row total and $n_{.j}$ is the marginal column total.

In order to assess the impact of the SDC methods on tests for independence, the Pearson statistic obtained from a perturbed contingency table is compared to the Pearson statistic obtained from the original contingency table. In particular, we focus on the measure of association, Cramer's V defined as: $CV = \sqrt{\frac{\chi^2 / n}{\min(R-1),(C-1)}}$. The utility measure is the percent relative difference: $RCV\ (T^O, T^P) = 100 \times \frac{CV\ (T^P) - CV\ (T^O)}{CV\ (T^O)}$

Table 6 presents results of the percent relative difference in the Cramer's V Statistic *(RCV)* based on the different data swapping strategies and swapping rates for each of the Census tables in Section 3 according to the OA geography.

Table 6. Percent difference in Cramer's V *(RCV)* for Census tables with OA geography

Tables	Random 2%	Random 5%	Target 2%	Target 5%
(1)	-0.660	-1.228	-1.090	-2.345
(2)	-0.641	-1.525	-0.689	-1.213
(3)	-0.927	-1.614	-1.710	-2.567
(4)	-0.601	-1.380	0.347	-0.884
(5)	-0.920	-1.635	-1.046	-2.065
(6)	-0.573	-0.986	-0.479	-0.941

The values in Table 6 are all generally negative which means that the swapped tables provide a measure of association that is always smaller than the measure of association based on the original table. This implies that data swapping of geographical variables attenuates the distributions in the tables and they lean more towards independence. Table 6 shows mixed results between the random and targeted swapping strategies at the same level of swapping rate. Tables (1), (3), (5) have higher *RCV* under the targeted data swapping while the other tables have lower *RCV*. Again, the reason for this pattern is due to the use of key variables to define high risk households which had a disproportionate chance of being selected for swapping. For those variables, the targeted swapping strategy induces more bias. In general, the 2% targeted swapping have lower values of *RCV* compared to the 5% random swapping, although this is not the case for Census table (3) where the 2% targeted swap has a higher *RCV* than the 5% random swap. Similar results are obtained at the ward level geography.

4.3 R-U Confidentiality Map

In this section, an R-U Confidentiality Map (Duncan, et al., 2001) is presented for the different data swapping strategies on each of the Census tables at the OA geography defined in Section 3. Figure 1 presents the empirical R-U confidentiality map based on the disclosure risk measure *DR* on the y-axis and the distance metric *AD* on the x-axis (note that the x-axis is reversed since a high *AD* represents low utility).

The lower left hand quadrant in Figure 1 represents low utility-low disclosure risk and the upper right hand quadrant high utility-high disclosure risk. In many cases, the 2% targeted data swapping has a lower disclosure risk than the 5% random data swapping and in general higher utility, i.e. smaller distance metrics *AD*. A line on the frontier of the data points is drawn in Figure 1 representing the points with the highest utility at each given disclosure risk. Three of the points are based on the 2% targeted

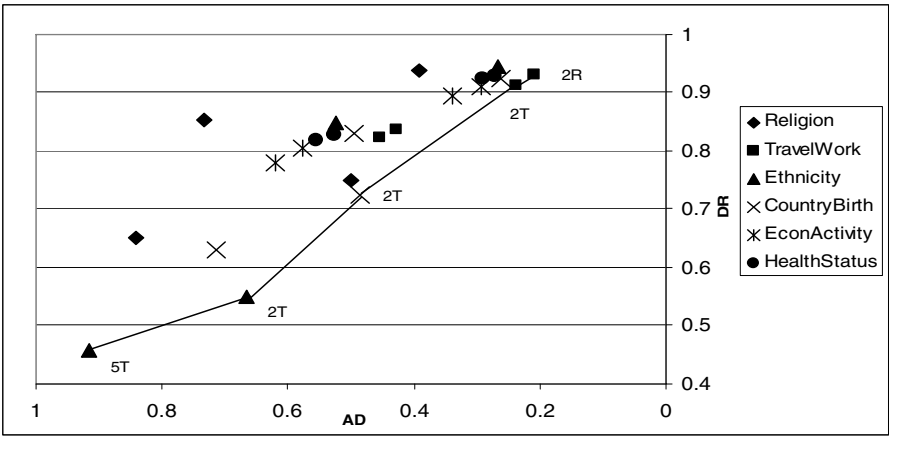


Fig. 1. R-U confidentiality map with *DR* (proportion of unperturbed unique cells) on the y-axis and the *AD* (distance metric) on the x-axis for all Census tables

swap and this would be the preferred option for this analysis based on the swapping rates and swapping strategies studied.

5 Discussion

In general, data swapping as a sole SDC method for protecting Census tables results in high probabilities that small cells in tables are true values. The method should be used in combination with other SDC methods, for example implementing a comprehensive and strict output design strategy with fixed categories of variables, population thresholds, etc. or some small cell masking.

We propose a targeted data swapping strategy which lowers the disclosure risk for a given swapping rate compared to random data swapping, especially for Census tables involving key variables that are used to define high risk households that are targeted for swapping. Higher swapping rates raise the level of protection but also cause more loss of utility. The results from the analysis show that the proposed targeted data swapping lowers disclosure risk approximately equal to that of a random data swapping at double the swapping rate whilst having generally higher utility. The analysis also showed that there are considerable gains using the targeted data swapping strategy compared to a random data swapping strategy, especially when aggregating lower levels of geography.

In any perturbative SDC method that is used to protect statistical data there are hidden non-transparent effects to the data which impacts on the ability to carry out statistical analysis. While the Census tables have the advantage that they are consistent and additive, this is undermined by the inability to obtain confidence intervals that take into account the perturbation. NSIs need to provide information and guidance to users in order to inform them of the impact of SDC methods and how to analyze disclosure controlled statistical data. Quality measures should be disseminated with the release of the Census tables to allow users to try and correct inferences using measurement error models.

References

1. Duncan, G., Keller-McNulty, S., Stokes, S.: Disclosure Risk vs. Data Utility: the R-U Confidentiality Map. Technical Report LA-UR-01-6428. Statistical Sciences Group. Los Alamos National Laboratory, Los Alamos (2001)
2. Gomatam, S., Karr, A.: Distortion Measures for Categorical Data Swapping. Technical Report Number 131, National Institute of Statistical Sciences (2003)
3. Gouweleeuw, J., Kooiman, P., Willenborg, L.C.R.J., De Wolf, P.P.: Post Randomisation for Statistical Disclosure Control: Theory and Implementation. Journal of Official Statistics 14, 463–478 (1998)
4. Shlomo, N.: Statistical Disclosure Control Methods for Census Frequency Tables. International Statistical Review 75(2), 199–217 (2007)
5. Young, C., Martin, D., Skinner, C.J.: Geographical Intelligent Disclosure Control for Flexible Aggregation of Census Data. International Journal of Geographical Information Science 23(4), 457–482 (2009)
6. Willenborg, L., de Waal, T.: Elements of Statistical Disclosure Control. Lecture Notes in Statistics, vol. 155. Springer, New York (2001)

Eliminating Small Cells from Census Counts Tables: Some Considerations on Transition Probabilities

Sarah Giessing[1] and Jörg Höhne[2]

[1] Federal Statistical Office of Germany,
65180 Wiesbaden, Germany
Sarah.Giessing@destatis.de
[2] State Statistical Institute Berlin-Brandenburg,
Alt-Friedrichsfelde 60
10315 Berlin, Germany
Joerg.hoehne@statistik-bbb.de

Abstract. The software SAFE has been developed at the State Statistical Institute Berlin-Brandenburg and has been in regular use there for several years now. It involves an algorithm that yields a controlled cell frequency perturbation. When a micro-data set has been protected by this method, any table which can be computed on the basis of this micro-data set will not contain any small cells, e.g. cells with frequency counts 1 or 2. We compare empirically observed transition probabilities resulting from this pre-tabular method to transition matrices in the context of variants of micro-data key based post-tabular random perturbation methods suggested in the literature, e.g. [8] and [4].

1 Introduction

In preparation for the German census 2011 we have started a comparative study of several perturbation methods for census frequency counts. The German Census will partly be register based, and partly be the outcome of a sample survey. This leads of course to limitations in the amount of detail of tables that can sensibly be released, as compared to a full census. Nevertheless, a huge amount of tabular output is going to be published. Publication of tables will to a major extent be pre-planned, but there will also be some flexible, user demand driven release of tabular data.

Given the size of the publication, and other complexities (like non-nested hierarchies that are foreseen for some classification variables like “age”) non-perturbative methods like cell suppression do not seem to be a good choice: one of the issues to be raised here is that with cell suppression, there would be a considerable disclosure risk due to incomplete coordination of cell suppression patterns across tables. Perturbation methods also have the advantage that they introduce ambiguity into the zero cells which helps to avoid attribute disclosure when (nearly) all members of a population group score on only one (sensitive) category of a variable.

In this paper, we investigate into basically three alternative methods. The software SAFE is in regular use at the State Statistical Institute Berlin-Brandenburg. SAFE is an implementation of an algorithm that yields a controlled cell frequency perturbation. When a micro-data set has been protected by this method, any table which can

J. Domingo-Ferrer and E. Magkos (Eds.): PSD 2010, LNCS 6344, pp. 52–65, 2010

be computed on the basis of this micro-data set will not contain any small cells, e.g. cells with frequency counts one or two. These small frequencies are the main concern for disclosure risk in Census counts tables, since they give information on the uniqueness or rareness of certain attributes or attribute combinations of individuals. Because SAFE is a pre-tabular method, all tables computed from the perturbed micro-data set protected by SAFE are fully consistent and additive.

In comparison to SAFE, we intend to study two post-tabular perturbation methods which both are based on the use of microdata keys. This technique can ensure full, or at least approximate, consistency of perturbations across different tables. Across table consistency has two aspects: On one hand, inconsistencies may be irritating to users. More severe from the disclosure control point of view is that inconsistency may lead to disclosure risk. For example, an average taken over eventually inconsistently perturbed values of logically identical cells (taken from different tables) should not be an unbiased estimate of the original cell value.

Each of the two post-tabular methods involves two steps. The first step yields fully or approximately consistently perturbed, but non-additive tables. Non-additivity is a potential nuisance for users, and may also be a source of disclosure risk. Therefore, in a second step, table additivity should be restored. This can be achieved by statistical methods such as the iterative proportional fitting algorithm. In this paper we discuss using linear optimization techniques for this step.

In order to avoid a perception of disclosure risk, and to provide a "visible" kind of protection, we require both methods to provide, like SAFE, perturbed data without small cells (i.e. without counts of one and two).

This paper reports on findings of the first phase of the study when implementing the methodologies. It is organized in seven sections: In section 2, we outline the methodological approach of SAFE. Technical issues of constructing suitable probability transition matrices for random perturbation methods are discussed in section 3, followed by a discussion of issues that came up when implementing the selection procedure for the random noise in section 4. In section 5, we suggest an optimization technique to restore table-additivity, e.g. the CTA method of [2], and propose a measure of information loss on the cell level for SAFE results in section 6. We conclude the paper with a brief summary section 7.

2 Methodological Background of SAFE

In this section we briefly describe the methodological approach of SAFE, as far as it is relevant for an application to protect tabulations of population Census counts data. For a more complete description the interested reader is referred to [5], or [6]. Starting point for the method is a microdata file where all variables are recoded to give the highest degree of detail foreseen for any publication. Imagine a variable like age, where perhaps data are collected so that for each person age could be deduced down to the level of age in months, but publications should offer data at most by age in years. Then the variable would be recoded to the level of age in years. We also assume here that the data-set consists of categorical variables only.

The basic idea of the method is to turn this data-set (with, say, N variables at n_i ($i=1,\ldots,N$) categories) into a data-set, in which either none of the records, or at least three records score on each of the $n_1 * n_2 * \ldots * n_N$ theoretical combinations of categories.

With respect to data quality, the method aims to preserve as far as possible cell counts in a pre-defined set of 'controlled' tables. For those tables, the method yields results that are in some sense 'optimal'. If any other table is derived from the perturbed data-set, it will be safe (i.e. it will not contain any ones or twos), but differences between original counts and those computed on basis of the perturbed data set can be much larger than they arise for the controlled tables. The experience is that the method is usually able to achieve a maximum deviation between 4 and 8 for a sensibly defined set of controlled tables.

The program computes a heuristic solution for the problem of minimizing the maximum absolute deviation between true and perturbed cells values in the controlled tables. While the initial mathematical statement of the problem resembles a huge non-linear integer optimization problem which is computationally intractable, an efficient heuristic algorithm has been developed that gives near optimal solutions at reasonable expense of computer resources.

3 Generating Random Noise for Frequency Tables

The Australian Bureau of Statistics ([4],[7]) has developed a concept for a cell perturbation method. They propose that the random noise should have zero-mean and a fixed variance. An alternative cell perturbation method referred to as "Invariant Post-tabular SDL" method was suggested in [8]. In the following two subsections we briefly outline the two alternative concepts and discuss the technical construction of suitable probability transition matrices for a random perturbation eliminating all small frequency counts.

3.1 How to Create Zero-Mean / Fixed Variance Cell Perturbations?

[4] propose to generate for each cell c with non-zero cell count i_c an independent integer value perturbation d_c satisfying the following two criteria:

(a) mean of zero
(b) fixed variance V for all cells c and all frequency counts i

A third criterion, in order to meet the requirement that perturbed cells do not have a count of one or two, would be

(c) $i_c+d_c \notin \{1,2\}$ f.a. i_c, d_c

This means we look for a L x L transition matrix $\mathbf{P}$ [1] containing conditional probabilities: $p_{ij} = p$ (perturbed cell value is j | original cell value is i) with the following properties:

[1] As index j may take a value of zero (when a cell value is changed to zero), in the following we start counting matrix and vector indices at 0, enumerating rows and columns of the L x L matrix by $0,1,2,\ldots,L$-1.

(1) $p_i \nu_i = 0$

(2) $p_i (\nu_i)^2 = \mathrm{V}$

(3) $p_{ij} = 0$ for j in $\{1,2\}$

(4) $\sum_j p_{ij} = 1$

(5) $p_{ij} = 0$; if $j < i - D$ or $j > i + D$,

(6) $p_{00} = 1$ and $p_{0j} = 0$ for $j > 0$, and of course

(7) $0 \le p_{ij} \le 1$

where p_i denote the ith row-vector of matrix $\boldsymbol{P}$ and ν_i a column vector of the noise which is added, if an original value of i is turned into a value of j. I.e. the j^{th} entry of ν_i is (j-i). For example ν_i = (-1,0,1,2,3,…,L-2). (1) is equivalent to (a) and expresses the requirement that the expected value of the noise should be zero. Similarly, (2) is equivalent to (b), expressing the requirement of a constant variance. (3) and (7) are of course necessary for any Transition matrix, (5) states a maximum allowed absolute perturbation of some pre-defined constant D and (6) states that zero frequencies must not change. Note for all rows after row D + 2, condition (3) is always satisfied, when (5) holds. Hence we can facilitate the task of computing suitable transition probabilities by adding a symmetry requirement for all rows after row D + 2:

(8) $p_{i,i-k} = p_{i,i+k}$ for k = 1,..,D, if $i > D + 2$

With (8), condition (1) is always satisfied because the negative and positive deviations balance each other. (2) simplifies into

(2a) $2 \sum_{j=1,\ldots,D} p_{ij} j^2 = V$.

For simplicity, in the following we therefore assume $L - 1 = D + 3$, applying the perturbation probabilities given by row (D + 3) of matrix $\boldsymbol{P}$ to all cell counts $\ge D + 3$.

For every row (or cell count) i (i=1,..,D + 2) conditions (1) to (5) can be rewritten as system of three linear equations

(9) $\boldsymbol{A}_{iD}\, x = b$, where

$\boldsymbol{A}_{iD}$ is a (3 x (min(i,D)+ 1+ D-k)) [2] coefficient matrix and b = (1,0,V)'. The elements of x correspond to the entries of row i in P which are not zero anyway by definition (because of (3) or (5)). The first row of $\boldsymbol{A}_{iD}$ corresponds to condition (4), the second row to (1) (e.g. unbiasedness) and the third row to (2) (fixed variance V).

Consider for example $\boldsymbol{A}_{13} = \begin{Bmatrix} 1 & 1 & 1 \\ -1 & 2 & 3 \\ 1 & 4 & 9 \end{Bmatrix}$. In this simple case, the coefficient matrix is invertible. The last row of the inverse is (-1/2,-1/4,1/4). Hence, in order for p_{13} to be positive, (-1/2,-1/4,1/4) • b = (-1/2+ V/4) must be positive, and hence V must be at least 2. In this case (9) has a unique solution, depending on the choice of V only. If V is exactly 2, p_{13} is zero.

In general, $\boldsymbol{A}_{iD}$ has more columns than rows. So usually, there is no unique solution for (9). But we can use (9) to derive feasibility intervals for x (e.g. for the p_{ij}).

[2] k is the number of elements in $\{1,2\} \cap [i\text{-}D\ ;\ i\text{+}D]$.

A practical approach is to fix V to 2+ε with a small positive value for ε (increasing ε and hence the variance of the perturbation leads to an unnecessary loss of information). The system (9) can be further strengthened by additional constraints, for example to express desirable monotony properties like $p_{ij} \geq p_{i,j+1}$ for $j > i$, or to improve symmetry by bounding the difference between $p_{i,i-1}$ and $p_{i,i+1}$.

We have experimented with $D = 3$, 4 and 5. One of the findings was that for small D and i, the linear programming problem derived from (9) (eventually together with the additional constraints) gives quite small intervals for x. In those cases it is usually reasonable to minimize or maximize one of the variables. In some cases for example we have maximized the variable corresponding to the most right hand tail of the distribution (e.g. $p_{i,i+D}$), when it was very small anyway, or we have maximized a variable at the centre of the distribution, which more or less fixes the remaining variables.

For larger D and i the intervals for x are wider. In those cases we first fixed a value (like 70%) for the centre of the distribution, p_{ii}. Afterwards we fitted each tail of the distribution p_{ij}, $j > i$ and p_{ij}, $j < i$ to the tails of a normal distribution using a heuristic approach outlined in the appendix. Table 1 in the appendix shows the final probability matrices for D=3,4 and 5 and compares them to the transition probabilities observed empirically for the cells of the set of controlled tables after protecting the data by SAFE. Obviously, the SAFE method results in much smaller probabilities that cell values change by less than three.

3.2 How to Combine Invariance and a "No-Small-Cells" Requirement?

The idea of the "Invariant Post-tabular SDL" method ([8]) is to preserve the frequency distribution of the cell counts. But in our setting we require the frequency of perturbed small counts (ones and twos) to be zero. So for the small counts these are aims that clearly exclude each other. The way out we propose here is to relax the goal of invariance. We only seek to preserve the frequency distribution of cell counts above three and the total frequency of all cell counts below four. This can be achieved as follows:

As shown in [8], an invariant matrix $\boldsymbol{R}$ is obtained by multiplying some pre-defined initial transition matrix $\boldsymbol{P}$ (for an example see [8]) with a suitable matrix $\boldsymbol{Q}$. $\boldsymbol{Q}$ is obtained by transposing matrix $\boldsymbol{P}$, multiplying each column j by the relative frequency of count j and then normalizing its rows so that the sum of each row equals one. Finally the diagonal elements of this matrix are increased by the following transformation $\boldsymbol{R}^* = \alpha R + (1-\alpha) I$, where I is the identity matrix of the appropriate size.

We adapt this procedure using a two-stage approach. In the first stage, we compute an invariant matrix $\boldsymbol{R}^*$ such that the first row gives the joint transition probabilities of all counts under four, and the first column gives the probabilities for changing a given count into a count smaller than four. The procedure to obtain $\boldsymbol{R}^*$ is the same as in [8], except that we use a vector of relative frequencies, where the entries corresponding to the ones, twos and threes are added up to one joint entry v_{1-3}. We also replace the first row of the initial transition matrix by a column vector where all entries except for the first two are zero. The first two entries can be computed as follows: Define an initial transition matrix P_0 for the small counts that leads to an unbiased perturbation like

$P_0 = \begin{Bmatrix} 1 & 0 & 0 & 0 & 0 \\ 2/3 & 0 & 0 & 1/3 & 0 \\ 1/3 & 0 & 0 & 2/3 & 0 \\ 0.1 & 0 & 0 & 0.6 & 0.3 \end{Bmatrix}$. Sum up the four rows of this matrix and divide through the total of the matrix entries. Use the three non-zero entries of this row-vector as first three entries of the initial transition matrix. Now, in order to come to a matrix $P_{1\text{-}3}$ of separate transition probabilities for the small counts remaining small counts, we add a second stage to the procedure: after stage one, multiplying the first row of ***R**** with the total frequency of counts under four gives expected numbers $n_{u4,u4}$ and $n_{u4,3+j}$ of cells under four that remain cells under four, and of cells that change into counts $(3 + j)$ for some $j > 0$. Assume without loss of generality a non-zero frequency of threes in the original table. Set the expected number of threes that change into counts $(3+j)$, $n_{3,3+j}$ to $n_{u4,3+j}$ and, correspondingly, $n_{i,3+j}$ to zero for i=1,2 and $j > 0$. Use the probability p_{33} of threes to remain unchanged of the initial transition matrix P_0 to compute an expected number n_{30} of threes that change into zero by subtracting the expected number of threes that do not change or turn into $3 + j$ from the number of threes in the table: $n_{30} = V_3 - \sum_{j\geq 0} n_{3,3+j}$. Then compute transition probabilities p_{1j} and p_{2j} (j=0,3) so that the expected bias of the counts under four that remain counts under four is zero, e.g. $-V_1 p_{10}+2V_1 p_{13}-2V_2 p_{20}+V_2 p_{23}-3p_{30}=0$. It is straightforward to show that adding $\cdot \begin{Bmatrix} -\beta & 0 & 0 & \beta \\ -\beta & 0 & 0 & \beta \end{Bmatrix}$ to rows 2 and 3 of the initial transition matrix $\boldsymbol{P}_0$ solves this problem, when $\beta = n_{30} / (V_1 + V_2)$. Obviously, this definition of p_{ij} (or n_{ij}, resp.) for $i = 1, 2, 3$ and $j = 0,3$ is also consistent with ***R****: Summing the entries of $(V_1, V_2, V_3) * P_{1\text{-}3}$ should equal the first entry of the first row of ***R**** multiplied with the total number of counts under four, $V_{1\text{-}3}$. Because of the properties of the invariant matrix ***R**** the latter equals the total number of counts under four minus the number of counts under four that change into counts $3 + j, j > 0$, e.g. $V_1 + V_2 + V_3 - \sum_{j>0} n_{3,3+j}$ which equals $V_1 + V_2 + n_{30} + n_{33}$ because of the definition of n_{30}. On the other hand, because the row totals of the first two rows of $P_{1\text{-}3}$ are one, summing the entries of $(V_1, V_2, V_3) * P_{1\text{-}3}$ yields $V_1 + V_2 + n_{30} + n_{33}$ as well. Hence, if we replace the first line of ***R**** by the separate transition probabilities for counts under four as explained here (and attach three columns of zeros to the other lines), we get a transition matrix ***R*****, which is almost invariant, except that for counts under four only their total frequency is preserved. An example using real data of a table of the last West German census of 1987 is given in below.

Example 1

For a census table with frequencies $(V_1, V_2, V_3, V_4, V_5,\ldots)$ = (96,32,20,16,15, …) observed for counts (1,2,3,4,5,…), we computed an initial invariant matrix ***R**** (with D=2). Table 1 shows the first four rows and six columns of the matrix of expected frequencies obtained from $(V_{1\text{-}3}., V_4, V_5, V_6, V_7,\ldots) \bullet \boldsymbol{R}^*$

Table 1. Expected frequencies $n_{i,j}$ of counts of i perturbed into counts of j

	0-3	4	5	6	7	8
0-3	144.62403	2.9690406	0.4069246	0	0	0
4	2.9690406	11.81552	1.0893785	0.1260606	0	0
5	0.4069246	1.0893785	12.211631	1.196397	0.0956693	
6	0	0.1260606	1.196397	10.676843	0.9239783	0.0767213

We now set n_{34} and n_{35} to 2.9690406 and 0.4069246 and compute n_{33}=0.6 • 20=12.

Then $n_{30} = V_3 - \sum_{j \geq 0} n_{3,3+j}$ yields

20 - 0.6 • 20 - 2.9690406 - 0.4069246 = 4.6240348 (hence p_{30}=0.2312017375), resulting in $\beta = n_{30} / (V_1 + V_2) = 4.6240348 / 128 = 0.03612527$.

According to the specifications above, $(p_{ij})_{i=1,2;j=0,3}$ are then $\begin{Bmatrix} 2/3-\beta & 1/3+\beta \\ 1/3-\beta & 2/3+\beta \end{Bmatrix} = \begin{Bmatrix} 0.63054139 & 0.36945861 \\ 0.29720806 & 0.70279194 \end{Bmatrix}$.

Table 3 (appendix) shows the first six rows and six columns of the matrix of expected frequencies computed as $(V_1, V_2, V_3, V_4, V_5, \ldots)$ • $\boldsymbol{R}^{**}$. The sum of the first two column totals in table 3 (regarding j=0,3) is 148, e.g. the total observed frequency of the counts under 4 (= 96+32+20) is exactly preserved.

4 Selection of Random Noise

The random mechanism proposed in [4] can be implemented very easily: For our experiments, we used the SAS random number generator which produces pseudo random numbers distributed uniformly over $[0;2^{31}-1]$. We assign such a random key to each record in the microdata file. When computing the tables, also the random keys are aggregated. The result is then transformed back into a random number on this interval by applying the modulo function, e.g. $\mathrm{mod}_{2^{31}-1}$. If the same group of respondents is aggregated into a cell, the resulting random key will always be the same. Cells which are logically identical thus have identical random keys.

Then we simply use a transition matrix computed to give zero-mean / fixed variance noise (as explained in 3.1), compute cumulated probabilities (for each row) and multiply the resulting matrix by $2^{31}-1$. Denoting the entries of this matrix by $\boldsymbol{M}_{ij}$ we change a cell count of i of some cell c into j, if the random key of cell c is between $\boldsymbol{M}_{i,j-1}$ and $\boldsymbol{M}_{ij}$. This will guarantee that the expected values of the perturbed counts are identical to the original counts (unbiasedness) and lead to consistently perturbed data. However, for a given table, the mean perturbation of cells of a given frequency count i is not necessarily zero. This mean will depend on the actual distribution of the corresponding record keys, see the three right-most columns of table 4 in the appendix for an example. In this example the observed difference between a true cell count and the mean of the corresponding perturbed counts varies between -0.82 and 0.78.

4.1 Without Replacement Strategy – Some Practical Issues

According to the selection procedure outlined above, the selection of random noise for a group of cells with the same cell value i is completely independent. [8] point out that this resembles a selection strategy "without replacement". Instead of this, they propose a strategy "with replacement". With such a strategy, the perturbation is carried out using the exact proportions given by the probability matrix. As explained in [8], such a strategy would be implemented by sorting cells in the table that have identical cell count i by their record keys, divide this list into subsets according to the proportions given by row i of the transition matrix, and assign the corresponding perturbed count to each subset[3]. [8] suggest to do the partitioning into subsets in such a way that the number of elements of each subset matches the expected frequencies given by ($V\boldsymbol{R}^*$) (V denote the vector of observed frequencies of the cell values in the table). However, the expected frequencies are fractional numbers. Therefore, we propose this minor modification:

In a first step, round down the expected frequencies to the next integer, add them up (separately for each cell count i) and sample randomly the corresponding number of records from the sorted lists of cells with frequency i. Subdivide each sample according to the integer parts of the corresponding expected frequencies n_{ij} and assign perturbed cell count j to subset j.

In the second step, add up the decimal-parts of the expected frequencies for each diagonal of matrix ($V\boldsymbol{R}^*$), e.g. compute across the different cell frequencies expected numbers N_d of cells which have not been assigned a perturbed frequency so far and should be given a perturbation of d, $d \in \left\{-d^-_{max\chi},...,1,1,...,d^+_{max\chi}\right\}$. Sample the respective number of cells alternating from the left hand and from the right hand side of a list of cells that have not been assigned a perturbed frequency so far, sorted by record keys (but not by original cell count!). The finally remaining cells (those with record keys closest to the median) remain unperturbed. The first two columns of table 4 in the appendix present for the data of example 1 for cells with original count less or equal to twelve the frequencies of original and perturbed counts.

5 How to Restore Table-Additivity?

Non-additivity is a potential nuisance for users, and may also be source of some disclosure risk. As simple example, assume random noise with a maximum perturbation of two has been applied. Assume two cells with original count one are perturbed to count three, and the original total of two is perturbed to zero. Users are informed on the maximum perturbation. Hence they know that both inner cells must have original count one at least. But if any of them were greater then one, the original total would be at least three and could not get a perturbed value of zero.

[3] Of course this may lead to inconsistencies in the perturbations between tables. However, there will still be a certain tendency in the perturbations: a cell with a large random key will be more likely to get a positive perturbation than to get no, or a negative perturbation. Hence, if this cell appears in several tables, and an intruder takes the mean over those cells to estimate the true count, the result will usually overestimate the true count.

This kind of disclosure risk typically arises, when all inner cells are all perturbed in the same direction, each with the maximum possible perturbation, and the total cell is perturbed in the other direction, also with the maximum possible deviation. With perturbations based on transition matrices like the ones discussed in section 3 with usually small probabilities on the tails these events will be relatively rare. However, we should also bear in mind, that this is only the simplest kind of attack. A systematic analysis based on linear optimization techniques and taking into account the aggregate structure of a perturbed non-additive multidimensional table with a published maximum perturbation would probably break other perturbation patterns as well.

Restoring table additivity, as suggested in [4] and [8] is an integral part of the method, ensuring that the protection provided by the perturbations cannot be undone easily. [7] and [8] point out that restoring additivity can be achieved by iterative methods. As an alternative, we suggest to consider using a linear programming based method like Controlled Tabular Adjustement (see f.i. [3],[1]).

For a first experiment, we use the CTA implementation of [2]. The algorithm restores additivity to a table, minimizing an overall distance to the table provided as input. The distance function implemented is a weighted sum of absolute per-cell-distances. Weights are provided by the user of the software. The user can define for each cell upper and lower bounds on the deviations, and can define a set of cells labeled as '*sensitive cells*'. Sensitive cells are forced to change their values. For each sensitive cell, the user defines a '*protection interval*'. The adjusted cell value is not allowed to take a value within the protection interval.

The computational complexity of the problem depends strongly on the number of sensitive cells. In a first experiment, we therefore use a two stage approach: in a first CTA run, we only restore additivity to the table. Although in this step we assign cell weights which will avoid to some extent that the algorithm adjusts cell counts of zero[4] or three, we will usually get an adjusted table with some small cell counts (e.g. ones and twos). In a refinement run, we define these ones and twos as sensitive, and define the corresponding protection interval as the interval (0;3). At the same time, for all cells with counts greater or equal to three we defined a lower bound of at least three. For all cells with zero count, the upper bound is zero. This way, however, we run a certain risk of defining an infeasible problem, especially if we define at the same time rather narrow constraints for the non-sensitive cells.

As yet, we have tested this approach only on a fairly small, 2-dimensional test table (760 cells, c.f. example 1) which has been perturbed using an invariant matrix derived with the methodology outlined in 3.2, to obtain an adjusted table, where the maximum perturbation of cell counts is identical before and after the adjustment. This is certainly encouraging, but it seems unlikely that it is a general result. Before such an approach could be put into practice, much more experience would have to be gained, for example how to avoid infeasibility problems. A lot of experience is also necessary to determine "sustainable" parameters for the initial random perturbation in the sense that the adjustment process can preserve to some extent the properties of the random perturbation (like f.i. the maximum perturbation). In the example, for instance, we observed that after adjustment the percentage of cells with absolute perturbation below two decreased from about 92 % to about 89 %.

[4] Note that we do not allow original zero cell counts to be adjusted.

Because the adjustment cannot simultaneously take into account all tables ever to be released, it introduces inconsistencies in the perturbation. Identical cells, even if they received the same perturbation by the random process, may become adjusted to different values. This fact leads to some risk that some perturbations might be undone, if intruders run an LP-based analysis taking into account the aggregate structure across several tables. But this is not such an easy task, on one hand, and on the other hand, it may not be very successful, because it may happen that only original frequencies can be broken that do not cause disclosure risk.

Of course one might consider using the adjustment methodology without previous random perturbation, only to 'remove' cells with small counts from the table. But as long as this does not – unlike the SAFE method - yield a fully consistent data base, there is then a risk that by averaging cell values over a number of tables a user can recover the original data. With a previous random perturbation, such an approach will only recover the underlying perturbed table, as pointed out in [7].

6 Data Utility – A Cell Level Measure of Information Loss

Probably, many users of census counts data do not use them for complex statistical analyses, but are merely interested in learning simple facts, like 'how many people with properties X live in area Y?'. When those counts are perturbed, they should be informed how reliable each individual cell is. This is especially important, if a perturbation method may produce fairly large perturbations, although only for a very small portion of the cells, which can f.i. be the case for SAFE for cells which do not belong to the set of controlled tables.

A simple information loss measure on the cell level could be given by publishing along with the perturbed counts the absolute value of the perturbation. However, this may be too much information, leading to disclosure risk. Instead, one might publish the absolute value of a perturbed version of the perturbation.

Usually, to inform about data utility, one publishes information on the perturbation on the table level, like the frequency distribution. Therefore, when perturbing the perturbations, it makes sense seeking to preserve these frequencies. E.g. use an invariant matrix of transition probabilities for perturbing the perturbations of the original counts in a table. Generating such a transition matrix is a straightforward application of [8]. The only difference is that, unlike the original counts which are positive numbers, the perturbations take values between $-D$ and D.

7 Summary and Final Remarks

In preparation for a comparative study of several perturbation methods for census tabular frequency data, in this paper we have raised some practical issues regarding the implementation of two alternative approaches explained in literature. In particular, this paper has discussed in some detail how to construct zero-mean / fixed variance transition matrices required to implement the methodology of [4]. We have also extended the idea of an invariant transition matrix suggested in [8] to a situation where the perturbation procedure should eliminate small cells.

As pointed out in [4] and [8], additivity is not preserved by the post-tabular random perturbation method, but can be restored afterwards. We have outlined and started testing an approach based on linear optimization, e.g. CTA methodology.

Leaving a larger scale empirical comparison of the post-tabular methods discussed in the paper with the pre-tabular perturbation method SAFE briefly outlined in section 2 for the future, the paper provides evidence that the post-tabular methods as implemented here tend to result in smaller changes to the data than SAFE. On the other hand, as a pre-tabular method SAFE preserves additivity and consistency and is easier to implement in a flexible OnLine table generation environment. These are important properties and may be worth "less optimal" performance regarding data quality to some degree.

References

1. Castro, J.: Minimum-distance controlled perturbation methods for large-scale tabular data protection. European Journal of Operational Research 171, 39–52 (2006)
2. Castro, J., González, J.A.: A Package for L1 Controlled Tabular Adjustment. Paper Presented at the Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality, Bilbao (December 2-4, 2009), http://www.unece.org/stats/documents/2009.12.confidentiality.htm
3. Dandekar, R.H., Cox, L.: Synthetic Tabular Data – an Alternative to Complementary Cell Suppression (2002) (unpublished) (manuscript)
4. Fraser, B., Wooton, J.: A proposed method for confidentialising tabular output to protect against differencing. In: Monographs of Official Statistics. Work session on Statistical Data Confidentiality, Eurostat-Office for Official Publications of the European Communities, Luxembourg, pp. 299–302 (2006)
5. Höhne, J.: SAFE – ein Verfahren zur Geheimhaltung und Anonymisierung Statistischer Einzelangaben. In: Berliner Statistik - Statistische Monatsschrift 3/2003 (2003a)
6. Höhne, J.: SAFE - a method for statistical disclosure limitation of microdata. Paper Presented at the Joint ECE/Eurostat Worksession on Statistical Confidentiality in Luxembourg (2003b) (December 2007), http://www.unece.org/stats/documents/2003/04/confidentiality/wp.37.e.pdf
7. Leaver, V.: Implementing a method for automatically protecting user-defined Census tables. Paper Presented at the Joint ECE/Eurostat Worksession on Statistical Confidentiality in Bilbao (December 2009), http://www.unece.org/stats/documents/2009.12.confidentiality.htm
8. Shlomo, N., Young, C.: Invariant post-tabular protection of census frequency counts. In: Domingo-Ferrer, J., Saygın, Y. (eds.) PSD 2008. LNCS, vol. 5262, pp. 77–89. Springer, Heidelberg (2008)

Appendix

An algorithm to fit the tail of a transition probability distribution p_{ij}, $j > i$ and p_{ij}, $j < i$ to the tails of a normal distribution.

Table 2. Zero mean, Variance 2+ε probability transition matrices for maximum perturbations *D* of 3, 4 and 5 (short: ABS3 to ABS5) vs. empirically observed transition probabilities for SAFE

	0	3	4	5	6	7	8	9	10	11	12	13
						ABS, D=3						
1	0.667	0.332	0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
2	0.334	0.666	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
3	0.125	0.687	0.063	0.063	0.063	0.000	0.000	0.000	0.000	0.000	0.000	0.000
4	0.000	0.601	0.099	0.100	0.100	0.100	0.000	0.000	0.000	0.000	0.000	0.000
5	0.000	0.167	0.167	0.416	0.083	0.083	0.083	0.000	0.000	0.000	0.000	0.000
6	0.000	0.072	0.072	0.072	0.571	0.072	0.072	0.072	0.000	0.000	0.000	0.000
7	0.000	0.000	0.072	0.072	0.072	0.571	0.072	0.072	0.072	0.000	0.000	0.000
8	0.000	0.000	0.000	0.072	0.072	0.072	0.571	0.072	0.072	0.072	0.000	0.000
						ABS, D=4						
1	0.667	0.333	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
2	0.334	0.666	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
3	0.120	0.700	0.082	0.045	0.027	0.026	0.000	0.000	0.000	0.000	0.000	0.000
4	0.064	0.076	0.700	0.068	0.037	0.029	0.026	0.000	0.000	0.000	0.000	0.000
5	0.000	0.143	0.143	0.542	0.043	0.043	0.043	0.043	0.000	0.000	0.000	0.000
6	0.000	0.063	0.063	0.063	0.662	0.038	0.038	0.038	0.038	0.000	0.000	0.000
7	0.000	0.032	0.033	0.034	0.050	0.700	0.050	0.034	0.033	0.032	0.000	0.000
8	0.000	0.000	0.032	0.033	0.034	0.050	0.700	0.050	0.034	0.033	0.032	0.000
						ABS, D=5						
1	0.667	0.333	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
2	0.334	0.666	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
3	0.119	0.700	0.082	0.050	0.028	0.014	0.007	0.000	0.000	0.000	0.000	0.000
4	0.062	0.076	0.704	0.075	0.037	0.020	0.014	0.012	0.000	0.000	0.000	0.000
5	0.025	0.068	0.068	0.700	0.059	0.027	0.019	0.018	0.017	0.000	0.000	0.000
6	0.000	0.057	0.057	0.057	0.700	0.041	0.023	0.021	0.021	0.021	0.000	0.000
7	0.000	0.025	0.035	0.035	0.062	0.700	0.060	0.028	0.020	0.018	0.018	0.000
8	0.000	0.015	0.016	0.019	0.032	0.068	0.700	0.068	0.032	0.019	0.016	0.015
						SAFE						
1	0.680	0.288	0.031	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
2	0.408	0.472	0.073	0.006	0.040	0.001	0.000	0.000	0.000	0.000	0.000	0.000
3	0.208	0.514	0.101	0.015	0.153	0.008	0.000	0.001	0.000	0.000	0.000	0.000
4	0.077	0.440	0.122	0.026	0.262	0.058	0.002	0.013	0.000	0.000	0.000	0.000
5	0.022	0.294	0.112	0.046	0.337	0.111	0.023	0.053	0.002	0.000	0.000	0.000
6	0.004	0.157	0.085	0.051	0.347	0.154	0.052	0.136	0.010	0.000	0.002	0.000
7	0.000	0.037	0.070	0.044	0.294	0.182	0.087	0.198	0.071	0.004	0.013	0.000
8	0.000	0.009	0.015	0.035	0.203	0.164	0.119	0.244	0.123	0.044	0.042	0.002

At first, provisionally fix the other (say, the left-hand) tail of the distribution[5]. This gives a target total probability and target total variance for the right-hand tail (through subtracting the corresponding left hand tail values from differencing one (V, resp.)). Then approximate p_{ij} ($j > i$) by $F_{k+0.5+i} - F_{k-0.5+i}$, where F_x denote the Normal distribution with zero expectation and suitable Variance σ^2 at x, and k denote the starting

Table 3. Expected frequencies $n_{i,j}$ of counts of i perturbed into counts of j for example 1

	0	3	4	5	6
1	60.531974	35.468026			
2	9.510658	22.489342			
3	4.6240348	12	2.9690406	0.4069246	0
4		2.9690406	11.81552	1.0893785	0.1260606
5		0.4069246	1.0893785	12.211631	1.196397
6		0	0.1260606	1.196397	10.676843

Table 4. Data Utility for Example 1[6] cell counts less or equal twelve for different perturbation methods: I. Cell count frequencies. II. Differences between observed means of the perturbed counts and true cell count.

	I: Cell Count Frequencies			II: Differences of Perturbed Count Means and True Count					
Count	Orig	Inv	Inv+CTA	Inv	Inv+CTA	SAFE	ABS3	ABS4	ABS5
1				0.17	0.06	0.00	0.01	-0.01	-0.01
2				0.18	-0.18	0.15	0.36	-0.36	-0.36
3				0.91	0.64	0.36	-0.05	-0.09	-0.09
1-3	155	156	155						
4	19	17	16	0.00	0.26	0.53	0.32	0.21	0.21
5	15	15	15	0.07	0.27	0.00	0.00	0.00	0.20
6	15	16	14	-0.07	-0.27	0.60	0.07	0.13	0.20
7	14	12	17	0.14	0.50	-0.86	0.50	0.50	0.43
8	9	11	11	0.11	0.33	0.22	0.78	0.78	0.78
9	16	15	10	0.00	0.19	0.19	-0.19	0.00	0.00
10	17	19	21	0.06	0.00	-0.71	0.18	0.06	-0.12
11	11	12	14	0.00	-0.09	-0.09	-0.82	-0.55	-0.55
12	10	7	6	-0.30	-0.30	-0.20	-0.30	-0.40	-0.40

[5] In some cases (combinations of D,i) the intervals for the elements of x corresponding to the left hand tail are so narrow that the provisional distribution gained by maximizing or minimizing one of the variables should be considered as final).

[6] In fact, for an extended data-set. Example 1 refers only to the set of inner cells of a two-dimensional table for a particular municipality. The results presented in table 4 also refer to cells in the margins of this table. Note, we apply the methodology separately to inner cells and each set of marginal cells of a table. This should help to exactly preserve the marginal total, e.g. the number of inhabitants of the municipality, considered as highly important.

point of the distribution tail. The starting point k should be selected as to achieve that the approximate $p_{i,D+i}$ is about zero. Usually, the sum of the approximate probabilities (for i=1,…,D) is below the target total probability for the right hand side. We correct this by adding 1/D of the difference to each approximate $p_{i,i+j}$. Now the total tail probability equals the target probability, while the tail variance still does not equal the target tail variance. This can be achieved by setting the variance σ^2 of the normal distribution to a suitable value, which can be established for example by a simple numeric interpolation approach.

The corrected approximate $p_{i,i+j}$ distribution can then be used to derive the target values for a corrected total probability and variance of the left-hand tail. Carry out the procedure described for the right hand tail for the left hand tail now. Finally, feed back the corrected approximate p_{ij} into the system (9) (of section 4) and (by minimizing or maximizing one of the variables) obtain a final distribution which meets the requirements of (9) with sufficient precision. □

Three Ways to Deal with a Set of Linked SBS Tables Using τ-ARGUS

Peter-Paul de Wolf and Anco Hundepool

Statistics, Netherlands
pwof@cbs.nl, ahnl@cbs.nl

Abstract. In Council Regulation no. 2701/98 of the European Committee, a framework is given on an extensive set of tables concerning economic statistics. Some of these tables are linked to each other. Until recently, there existed no practical solution to a consistent protection of that set of tables, save for a rather naive one. In this paper we will show the new way specific sets of linked tables can be protected using the τ-ARGUS software and compare this with two other approaches.

1 Introduction

In section 2 we will describe a set of tables that are related to the SBS regulation of Eurostat. We will describe the way they are linked to each other. In section 3 we will describe the naive way of protecting that set of tables and present the results. A more elaborate way to deal with such a set of linked tables is to use τ-ARGUS to protect the tables in a specific order, and suppression patterns are copied between tables. See [1] for a conceptual framework to protect SBS tables using this approach. Section 4 will discuss the implementation we used. In section 5 we will present the newly implemented way of dealing with particular sets of linked tables, as described in [2], [3] and [7]. Finally, in section 6 we will draw conclusions on the results from the three approaches.

2 The Set of SBS Tables

Community Structural Business Statistics (SBS) are collected within the framework of Council Regulation no. 2701/98 (EC, EURATOM). Definitions and table breakdowns are specified in a series of Commission and Council Regulations. In this paper, we will call 'the set of SBS tables' the set of core tables defined in the first four annexes of the regulation, covering the business economy (annex 1), industry (annex 2), distributive trades (annex 3) and construction (annex 4).

These tables consist of e.g., the total turnover of businesses of a certain activity. The spanning variables for those tables are NACE (Statistical Classification of Economic Activities in the European Community, Rev. 1.1), SC (size class, a classification of businesses according to the number of employees) and NUTS (Nomenclature of Territorial Units for Statistics). The NACE code is a classification of businesses according to different sectors in the economy, like trade,

J. Domingo-Ferrer and E. Magkos (Eds.): PSD 2010, LNCS 6344, pp. 66–73, 2010.

code	number of employees
01	0–1
02	0–9
03	2–9
04	10–19
05	20–49
06	50–249
07	250 or more
10	0–49
30	Total

Fig. 1. Codes of spanning variable Size Class

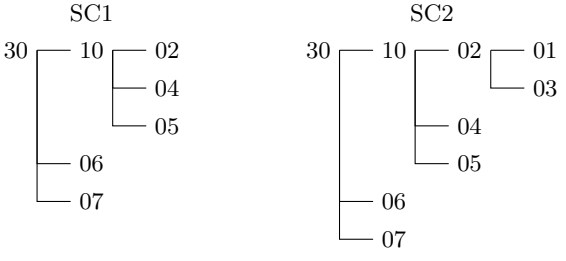


Fig. 2. The two hierarchical structures SC1 and SC2

industry, wholesale, etc. This classification is hierarchical in the sense that the more digits a code has, the more detailed branch it describes. E.g., 2 digit NACE code 45 stands for 'Construction' and 3 digit NACE code 451 stands for 'Site preparation'. For a complete overview of NACE codes and NUTS codes, we refer to the website `http://ec.europa.eu/eurostat/ramon`.

In the NACE classification, several codes are redundant. E.g., NACE 16 is not split into more detailed subcategories. In the code list this means that codes 16, 160 and 1600 are essentially the same and only appear to be able to define NACE on a 3 and 4 digit level. While protecting the tables, we have to take this into account. To simplify things, we have removed these so called 'bogus levels' and hence reduced the number of categories of the NACE variable.

The variable 'size class' is also hierarchical. See figure 1 for the codes of this variable. In figure 2 two slightly different structures are shown. The first structure (SC1) is used in tables with NACE C-F, whereas the second structure (SC2) is used in tables about the other NACE sectors G, H, I and K.

Using these spanning variables, we can divide the 'core set' of tables into three classes of differing NACE-level:

T1 annual enterprise statistics at four-digit NACE code (NACE4)
T2 annual enterprise statistics at three-digit NACE code (NACE3) broken down by size classes

Table 1. Set of core SBS tables

Table name	Classifying variable(s)	Sector(s) of NACE
T1.1	NACE4	C–K, excluding J
T2.1	NACE3 × SC1	C–F
T2.2	NACE3 × SC2	G-H-I-K
T2.3	NACE3 × NUTS2	G
T3.1	NACE2 × NUTS2	C–K, excluding G and J

T3 annual regional statistics at two-digit NACE code (NACE2) broken down by NUTS2 regional classification.

Obviously, the NACE code is the variable that links all classes of tables to each other. For part of the NACE classification (distributive trade, section G), the regional tables is released at NACE3 instead of NACE2.

For each class a number of tables is produced with different response variables according to the SBS regulation. Examples are turnover, value added, etc. In this paper we will concentrate on the variable turnover.

An additional complication is that the tables have different response units. E.g., class T1 reports on Kind of Activity Units (KAU) whereas classes T2 and T3 report on enterprises. In this paper we will consider these units as being the same (as is the pragmatic approach used at Statistics Netherlands).

The complete set of core tables we will consider in the current paper, is given in table 1. The safety rule we used in this paper is the p%-rule with $p = 15$. That is, no respondent in a cell should be able to estimate any other contribution to that cell within 15% of the true value of that contribution. The cost function we used was set equal to the cell value. This means that for each (sub)table, τ-ARGUS tries to find the suppression pattern such that the safety rules are satisfied while at the same time the total cell value that is suppressed is minimized.

Cells that do not satisfy the p%-rule are called primary unsafe. The objective of τ-ARGUS is to find a suppression pattern such that the resulting table is safe, whilst at the same time the amount of information that is lost is minimal. To find a safe table, a safety range is assigned to each primary unsafe cell. As a result, the suppression pattern τ-ARGUS tries to find should be such that for each primary unsafe cell an interval of possible values can be calculated that is at least as large as the safety range. A suppression pattern that satisfies the just stated property related to the safety ranges, is called feasible. See [5] for more details.

Using τ-ARGUS we did not use the singleton-options, since this would unnecessarily make the comparison of the three approaches more difficult. For a discussion of the singleton-options we refer to [4].

3 The Naive Way

The variables that define the tables are NACE (business classification), SC (size class) and NUTS (region). The NACE and the SC variables are hierarchical. All tables mentioned in table 1 can be considered to be subtables of one large

Table 2. Results of the naive approach

	Number of cells				Costs ($\times 10^6$)	
Table name	Total	Empty	Primary	Secondary	Total	Secondary
T1.1	545	4	54	31	6,878	59
T2.1	1,078	66	122	183	8,820	287
T2.2	549	5	21	42	11,789	36
T2.3 and T3.1	1,326	17	100	164	16,194	265

table, spanned by NACE (2-, 3- and 4-digit level), SC and NUTS. Using the terminology of [2] and [7] that large table would be called the covering table. A simple way to deal with the set of core tables would then be to completely protect the covering table. This is what we call the naive way. Since the core set of SBS tables that we consider results in a hierarchical 3-dimensional table, we will use the 'modular' method of τ-ARGUS. This method breaks down the hierarchical table into a large number of non-hierarchical subtables and deals with each subtable separately. This is done in a special order, keeping track of all suppressions of previously protected subtables. See [6] for more details on this method.

The cover table consists of 92,650 cells of which 24,957 are empty and 19,127 are unsafe according to the p%-rule, i.e., primary unsafe. Recall that we have removed the bogus levels from the NACE classification. Most of the empty cells (and most of the primary unsafe cells) are present in parts of the cover table that do not appear in any of the tables in the set of core CBS tables we consider in this paper.

We have used τ-ARGUS to protect this cover table. In the complete cover table 15,713 secondary suppressions were needed. It took over one hour to protect the complete cover table using the modular option within τ-ARGUS. Table 2 shows the results specified to each table in the core set of SBS tables. Tables 2.3 and 3.1 are grouped together, since they both are about the same regional variable and only differ in NACE-detail for one sector. Table 2 shows some general information about each table, the number of secondary suppressions and the total costs associated to the secondary suppressions (the information loss).

4 The Traditional Way

A more elaborate way to deal with the core set of tables, would be to iteratively protect all tables of that core set one at a time, while taking over the protection of a previously protected table. It is essential however, to determine the order in which the tables of the core set are dealt with, appropriately. The way we will deal with the set of tables in this paper is depicted in figure 3 and is essentially taken from [1].

According to this scheme, we would first protect table T1.1, then carry over that suppression pattern to tables T2.1 and T2.2. Next these two tables are protected and the pattern of table T2.2 is carried over to table T2.3 and table

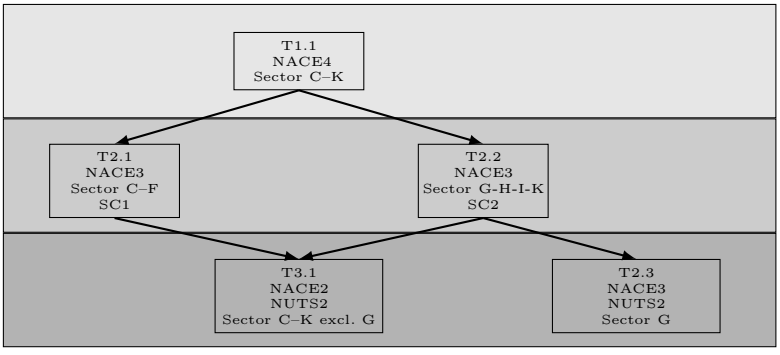


Fig. 3. Links between the tables corresponding to the order

T2.3 is protected. Finally, the patterns of tables T2.1 and T2.2 are carried over to table T3.1 and that latter table is protected. In figure 3 each arrow represents the fact that a suppression patterns is to be carried over. Each gray bar represents a group of tables at the same level of the process.

When carrying over a suppression pattern from table A to table B, the suppressed cells in table A that are also present in table B should be given the status 'unsafe'. The safe cells of the table A that are present in table B, are set to 'protected' in table B. In case a cell has the status 'protected' this means that τ-ARGUS is not allowed to use that cell as a secondary suppression when looking for a feasible suppression pattern. For some instances, setting cells to 'protected' may lead to an infeasible problem, i.e., τ-ARGUS cannot find a feasible solution. This yields that we have to go back to table A to check whether or not additional suppressions are needed. If that is the case, we should not set these cells to 'protected' in table B, but we should set their costs very high. Then τ-ARGUS will *try* not to suppress these cells in table B but would suppress them if really needed to get a feasible solution. Then we have to check if any newly suppressed cells in table B should also be suppressed in table A. If so, this would mean that we have to repeat the whole process until there are no additional suppressions in a previously protected table.

For this traditional approach, τ-ARGUS can be used as well. To carry over a pattern from one table to another, we have to produce a so called 'a priori' or 'history' file containing the information about the suppression pattern that has to be carried over. The thus introduced suppressions are called 'manually unsafe'. For these kind of unsafe cells, τ-ARGUS can not calculate an appropriate safety range automatically. Therefore a safety range has to be imposed by the user. This safety range is called the 'manual safety range' and is often specified as a percentage of the cell value of the manually unsafe cell.

We have used τ-ARGUS to protect the core set of SBS tables in this way. The secondary suppressed cells were considered primary with a manual safety range of 1% of the cell value. This small percentage was used to limit the effect of carrying over secondarily suppressed large cell values. Moreover, it often

Table 3. Results of the traditional approach

	Number of cells				Costs ($\times 10^6$)	
Table name	Total	Empty	Primary	Secondary	Total	Secondary
T1.1	545	4	54	25	6.878	52
T2.1	1,078	66	122	156	8,820	273
T2.2	549	5	21	40	11,789	36
T2.3 and T3.1	1,326	17	100	118	16,194	229

suffices to impose the condition that secondary suppressions can not be recalculated exactly.

Table3 shows the results specified to each table in the core set of SBS tables. Fortunately, it sufficed to carry the patterns over using the 'protected' approach, that is, we did not have to go back to tables that had already been protected earlier in the process. The time to protect the defined subtables was about one minute in total. However, this does not include the manual interaction needed to make a-priori files that carry over the suppression patterns, nor the time needed to think about the order in which the linked tables should be protected.

5 The New Approach

The new approach as implemented in a test version of τ-ARGUS, starts with constructing the same cover table as the one used in the naive approach. However, whilst the naive approach would protect the complete cover table, the new approach protects only the parts of the cover table that belong to any of the tables of the core set. The parts of the cover table that appear in none of the tables in the core set, will not be protected. Essentially this means that in the breakdown of the cover table into non-hierarchical subtables, some of these subtables will not be protected. See [2] and [3] for more details. Table 4 shows the results for this approach. The time needed to find this solutions was about two minutes.

Table 4. Results of the new approach

	Number of cells				Costs ($\times 10^6$)	
Table name	Total	Empty	Primary	Secondary	Total	Secondary
T1.1	545	4	54	25	6.878	52
T2.1	1,078	66	122	156	8,820	273
T2.2	549	5	21	40	11,789	36
T2.3 and T3.1	1,326	17	100	118	16,194	229

6 Conclusions

As has been shown in the previous sections, there are several approaches to choose from, when applying cell suppression as a disclosure control technique to the core set of SBS tables as defined in this paper.

The naive approach is obviously the easiest way to proceed. However, it often leads to over protection. Cells that do not appear in any of the tables of the core set, will be protected as well. This will often lead to additional suppressed cells that do appear in one or more of the core tables. Indeed, table 2 shows that the total information loss in terms of the number of suppressed cells as well as in terms of the sum of the suppressed cell values is larger than for the other two approaches.

The other two approaches are taking care of that problem by protecting only the published tables. The traditional and the new approach both lead to the same suppression patterns. For the instance used in this paper, it turns out that only a limited number of secondary suppressions need to be carried over in the traditional approach. Indeed: only in the process of going from table T1.1 to T2.1 six secondary suppressions needed to be carried over. Moreover, table T3.1 turned out to have only one single primary unsafe cell. However, in general this would not be the case.

For the same reason, no iterative procedure was needed in the traditional approach: assigning the status 'protected' to the safe cells that were carried over, still made it possible for τ-ARGUS to find feasible patterns. Hence, in our instance, the intensity of the manual interaction was not much for the traditional approach.

Another aspect that in theory could influence the intensity of the manual interaction, is the order in which the tables are protected. In our paper we made use of the suggestions made in [1].

The main advantages of the new approach can be summarized as:

- It is not necessary to think about the order in which the linked tables should be protected.
- No additional manual interaction is needed to carry over suppression patterns between tables.
- Manual backtracking will never be necessary.
- The amount of overprotection is limited.

A disadvantage of the new approach is that it can only be used when the set of linked tables can be viewed as part of a so called cover table of up to 4 dimensions. This restricts the number of situations in which the new approach is applicable. However, in practice a cover table of 4 dimensions often suffices. In the situations where the new approach is not applicable, the traditional approach would be a good alternative.

Concluding we would argue that the new approach as discussed in section 5 is to be preferred. Since we only applied the three approaches to a single instance, additional research should be made on other instances to back our conclusions.

References

1. Capobianchi, A., Franconi, L.: Cell suppression in linked tables from structural business statistics using Tau-Argus 3.3.0: a conceptual framework. Paper Presented at the NTTS 2009 Conference, Brussels (February 18-20, 2009)

2. Giessing, S., Wolf, P.P.: How to make the τ-ARGUS Modular Method Applicable to Linked Tables. In: Domingo-Ferrer, J., Saygın, Y. (eds.) PSD 2008. LNCS, vol. 5262, pp. 37–49. Springer, Heidelberg (2008)
3. Giessing, S., de Wolf, P.P.: Adjusting the τ-ARGUS modular approach to deal with linked tables. Data & Knowledge Engineering 68, 1160–1174 (2009)
4. Hundepool, A., van de Wetering, A., Ramaswamy, R., de Wolf, P.P., Giessing, S., Fischetti, M., Salazar, J.J., Castro, J., Lowthian, P.: τ-ARGUS user's manual, version 3.3, http://neon.vb.cbs.nl/casc/tau.htm
5. Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Lenz, R., Longhurst, J., Schulte Nordholt, E., Seri, G., de Wolf, P.P.: Handbook on Statistical Disclosure Control (2009), http://neon.vb.cbs.nl/casc/handbook.htm
6. de Wolf, P.-P.: HiTaS: A Heuristic Approach to Cell Suppression in Hierarchical Tables. In: Domingo-Ferrer, J. (ed.) Inference Control in Statistical Databases. LNCS, vol. 2316, p. 74. Springer, Heidelberg (2002)
7. de Wolf, P.P.: Cell suppression in a special class of linked tables. In: Work Session on Statistical Confidentiallity, Manchester, Luxembourg, December 17-19, pp. 220–226 (2007), Office for Official Publications of the European Communities, ISBN 978-92-79-12055-8

IPUMS-International Statistical Disclosure Controls: 159 Census Microdata Samples in Dissemination, 100+ in Preparation

Robert McCaa*, Steven Ruggles, and Matt Sobek

Minnesota Population Center, 50 Willey Hall
Minneapolis MN 55455 USA
rmccaa@umn.edu

Abstract. In the last decade, a revolution has occurred in access to census microdata for social and behavioral research. More than 325 million person records (55 countries, 159 samples) representing two-thirds of the world's population are now readily available to bona fide researchers from the IPUMS-International website: www.ipums.org/international hosted by the Minnesota Population Center. Confidentialized extracts are disseminated on a restricted access basis at no cost to bona fide researchers. Over the next five years, from the microdata already entrusted by National Statistical Office-owners, the database will encompass more than 80 percent of the world's population (85 countries, ~100 additional datasets) with priority given to samples from the 2010 round of censuses. A profile of the most frequently used samples and variables is described from 64,248 requests for microdata extracts. The development of privacy protection standards by National Statistical Offices, international organizations and academic experts is fundamental to eliciting world-wide cooperation and, thus, to the success of the IPUMS initiative. This paper summarizes the legal, administrative and technical underpinnings of the project, including statistical disclosure controls, as well as the conclusions of a lengthy on-site review by the former Australian Statistician, Mr. Dennis Trewin.

Keywords: Census microdata samples, data privacy, data dissemination, IPUMS-International.

1 Introduction

A revolution occurred in access to population census microdata for social and behavioral research in the first decade of the twenty-first century. The most successful initiative, with the cooperation of some 85 National Statistical Agencies world-wide, is the IPUMS-International project led by the Minnesota Population Center (MPC, Figure 1).

At this writing, datasets for 55 countries—159 anonymized, integrated samples totaling 325,430,447 person records—are available to registered researchers at no cost via the IPUMS-International web-site (Table 1). From the 250-odd datasets already

* Corresponding author.

J. Domingo-Ferrer and E. Magkos (Eds.): PSD 2010, LNCS 6344, pp. 74–84, 2010.

entrusted to the project, the number of countries represented is likely to increase to 85 or more over the next five years, and the number of datasets to some 250. Twenty to thirty samples are integrated into the database each year. 2010 round census data will be assigned the highest priority for integration, as they become available. For each country, an effort is made to construct a series of samples for all censuses for which microdata survive. Of the 159 samples currently in the database, 37 are from the 2000 round compared with 44 for the 1990s, 37 for the 1980s, 27 from the 1970s and only 13 from the 1960s. High precision household samples with a density of five percent or more number 128. Of the 30 lower precision samples, many consist of all the surviving microdata for the respective census. Notable exceptions are the samples for four censuses of Canada, and two each for China, the Netherlands, and the United Kingdom. The Chinese household samples, with a density of only one percent, number over ten million person records each.

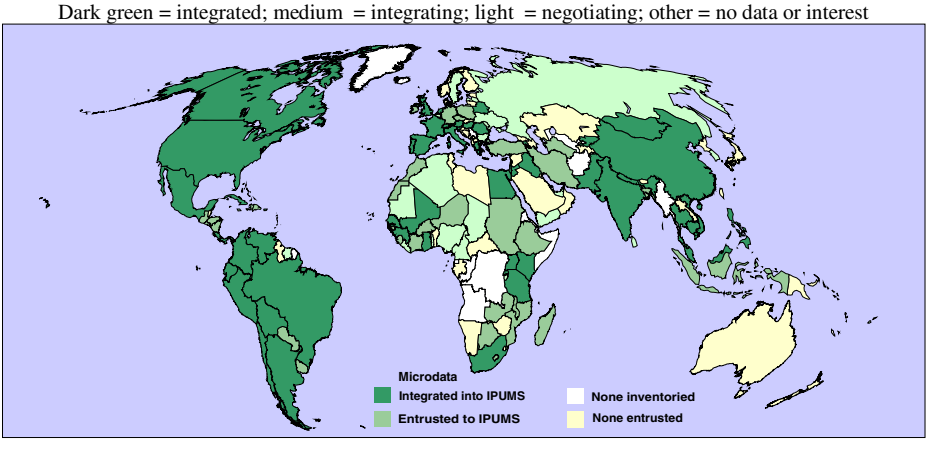


Fig. 1. The IPUMS-International World Map

3,750 researchers, representing 84 countries, are approved for access to the microdata from the IPUMS-International web-site (usage statistics as of July 1, 2010). 64,248 extracts, excluding those by MPC personnel, have been made, totaling 731,531 integrated variables extracted. Ten variables account for one-quarter of the usage: educational attainment, employment status, age, marital status, person weight, relationship to head (or reference person), sex, person number, sample identifier, and class of worker. The microdata of a mere seven countries account for one-half of the extracts: Mexico, USA, Brazil, Colombia, France, Chile and Argentina. The striking preference for American microdata is due to the fact that samples for these countries were among the first integrated into the database. Moreover, these countries have long series of censuses with extant microdata stretching back to the 1960s. Finally, the samples are rich, with at least 50 person variables and 10 household or dwelling variables.

At the PSD2006, we laid out the statistical disclosure controls to protect the privacy of persons, households and other entities developed on the first 47 samples integrated into the IPUMS-International database [1]. In this paper we describe how

the legal, administrative and technical procedures are being implemented to protect privacy and statistical confidentiality and, thus, to facilitate access to this massive trove of data. Restricting access to trusted users is the key to our success. To date, there have been no allegations of misuse of IPUMS-International census microdata extracts. On the contrary, what is most remarkable is the substantial usage by researchers, given the fact that the usability of "public use" microdata is sometimes deemed "limited" [2]. Despite the "PU" in the IPUMS acronym, "RA," "TU," or "SA" might be more appropriate because the data are disseminated as "restricted access," "trusted user," or "scientific access" files [3, 1].

2 Thwarting Intruders

The casual intruder (and casual user) is readily thwarted by the IPUMS-International registration form and policy statements. At 1,100 words, the IPUMS form is considerably shorter than the 5,830 words that constitute the FACEBOOK privacy policy, but, unlike FACEBOOK (where most registrants click "I agree" without reading the small print), the IPUMS registration requires the applicant to agree to each of eight detailed conditions of use. Failure to agree to even one condition results in an automatic rejection of the registration. The successful applicant provides not only personal details, but also must identify institutional affiliation, including name, official email address and phone number, web-link identifying affiliation, name and email of supervisor, the name, title and other pertinent information of any grant used to conduct the research and, most importantly, the name of "an Institutional Review Board (IRB), or Office for Human Subject Protections, Professional Conduct or similar committee." Applications that omit this information are reviewed, but a positive decision is delayed until bona-fides are explained and verified. Perhaps the biggest obstacle for a successful application is the research project description, which is carefully scrutinized to confirm that access to the database is needed for the proposed research. A researcher may possess adequate technical and professional qualifications, but if there is no research need for the microdata, access will be denied. Approximately one-third of completed applications are denied. In complete registrations—those that are begun and but never submitted—go uncounted, but it is likely that their number is not inconsiderable.

The rogue intruder—armed with the appropriate bona-fides but with malevolent intent—faces legal and institutional sanctions as well as substantial technical obstacles. If the violation occurs in the United States, the intruder risks civil prosecution with a maximum fine of US$250,000 and/or three years imprisonment. Elsewhere, since the laws of the country in which the violation occurs would apply, the discretion to prosecute would rest with the National Statistical Authority. The legal counsel of the University of Minnesota is committed to providing vigorous legal assistance. This threat of legal action is probably less a deterrent than institutional and professional sanctions. The IPUMS Case Study in [4, Annex 1.23] describes the sanctions as follows:

1. "sanctions against both the individual and the institution with which the individual is associated (e.g., University, international organization) [would be imposed];

2. "denial of access would immediately be invoked against the individual and his/her institution and would continue until corrective measures were deemed to be sufficient by the University of Minnesota and the National Statistical Office whose data were violated. If the institution where the breach occurred was the recipient of a grant from the National Institutes of Health of the United States, each researcher at the institution could be required to undergo Human Subjects Protection training and re-certification before access was re-instituted for individuals at that institution."

Commercial researchers are prohibited from accessing the data. Some petition for access, but are denied because of the restriction to non-commercial users. None seek access to identify individuals. Instead, commercial users often require population statistics that are not readily obtainable elsewhere, such as to compute weights or expansion factors for specific population sub-groups. There is no interest by commercial or other entities in linking confidentialized population census samples to other sources because much more valuable data are readily available elsewhere. Then too, leaving aside the difficulties of constructing successful links, sample microdata are too ordinary to excite the slightest interest for the purpose of linking.

3 Statistical Disclosure Controls

Threats to privacy and statistical confidentiality by intruders have long provided the rationale for National Statistical Offices to simply deny access to census microdata, regardless of the professional qualifications and scientific needs of would-be researchers. IPUMS International is successful in overcoming these objections because our procedures are designed to thwart intruders, first, by screening to permit trusted researchers to use the data while denying access to potential intruders; second, by erecting strong sanctions against misuse; and, third, by imposing stringent statistical disclosure controls. We endorse the standard of the Office of National Statistics (UK) [5] that statistical controls should be such that it "would take a disproportionate amount of time, effort and expertise for an intruder to identify a statistical unit to others, or to reveal information about that unit not already in the public domain." Population census variables are mundane. Census attributes are relatively crude in comparison to the details available in employment or health surveys. IPUMS-International suppresses variables considered to be sensitive by the official statistical agency, but to date, there has been only one such request: "tribe" for a census from an African country, where ethnic violence is a grave concern.

Census operations produce a considerable amount of data that is less than perfect. Editing and imputation are necessary to produce coherent datasets. Few statistical agencies report the details. The rich samples of the 2001 population census microdata of the United Kingdom make it possible to assess the degree of imputation and of perturbation—the introduction of intentional error to protect confidentiality in the data. The ONS relied upon the Post Randomization Method (PRAM) to produce the 2001 Licenses Individual SAR. De Kort and Wathan [6] compared the Individual Controlled Access Microdata Sample (not perturbed) with the Licensed Individual SAR (perturbed—note that this is the sample integrated into IPUMS-International)

and discovered that the relative frequency of imputation was several times greater than perturbation. Of the twelve variables analyzed, the authors found that for "Social Grade of Reference Person" 15% of attributes were imputed, versus 2% perturbed. For "Age" imputation and perturbation were roughly the same at 1%. The frequency of perturbations was typically less than one percent, whereas imputations for five variables were 5% or more. For researchers inclined to ignore the imputed data, de Kort and Wathan warn that "raw data are not necessarily to be preferred." The ONS-UK is to be lauded for producing flags to indicate imputation for every variable in its samples. Flags empower researchers to gauge the effects of imputation and editing and take appropriate action.

Purdam and Elliott [7] carried the analysis a step farther to assess the effects of perturbation on published analyses. Thanks to the ability to replicate certified samples such as the SARS and the CAMS, replication of research results can be accomplished with a degree of confidence. Their findings are disconcerting for researchers: "disclosure control measures had a significant impact on the usability of the data (analytical completeness) and on the accuracy of the data in relation to the findings reached when the data were used in analyses (analytical validity)."

As in the case of the 2001 SARS, a few statistical agencies entrust samples that have already been subjected to privacy protections. Sometimes these go seriously awry, as in the case of the United States PUMS [8]. Beginning with datasets from 2000 through 2008, serious errors were introduced into the public use files for males and females aged 65 years and over. Due to a programming error, statistical disclosure controls corrupted age attributes so that published distributions differed from those computed from microdata samples by as much as 15%. Three series of microdata samples were corrupted: the 2000 census, the American Community Survey (2003-6), and the Current Population Survey (2004-9) [9]. Despite the uproar in the media only one dataset was corrected, and some researchers fear that the correction may actually make matters worse.

Most statistical offices entrust "raw" microdata to the IPUMS-International project (not truly raw because names and addresses are stripped out thereby anonymizing the data before shipment). In Table 1, these instances are identified by "IPUMS" in the column headed "Confidentiality Protocols". In such cases, we apply a series of straight-forward SDC measures. First, the data are anonymized by suppressing any names, addresses, or precise geographic identifiers. Second, a sample is drawn so that researchers have access to only a minor fraction of the complete dataset. Third, additional disclosure protections are imposed on the sample, variable-by-variable and code-by-code. Finally, a small fraction of households is swapped across geographic boundaries.

Our procedures are summarized in a contract with one of our statistical agency partners, as follows:

> (1) Detailed geographic codes will be suppressed.
> (2) Any geographical unit with fewer than 20,000 individuals will be aggregated to the next highest geographical unit.

Table 1. Microdatasets entrusted, confidentiality protocols and sample densities

Sample density			Country	Confidentiality protocols	Census decade				
10%+	~5%	<=4%			2000s	1990s	1980s	1970s	1960s
Integrated and Disseminating 2002-2010: 55 countries, 159 censuses, 87 million households and 325 million person records									
4			**Argentina**	INDEC	**2001**	**1991**	**1980**	**1970**	1960
1			**Armenia**	SCS	**2001**		1989	1979	1970
4			**Austria**	IPUMS	**2001**	**1991**	**1981**	**1971**	1961
1			**Belarus**	IPUMS		**1999**	1989	1979	1970
3			***Bolivia**	IPUMS	**2001**	**1992**		**1976**	
5			**Brazil**	IBGE	**2001**	**1991**	**1980**	**1970**	**1960p**
2			**Cambodia**	IPUMS	**2008§**	**1998**			1962
		4	**Canada**	STATSCAN	**2001p**	**1991p**-6	**1981p**-6	**1971p**	1961
4		1	***Chile**	IPUMS	**2002**	**1992**	**1982**	**1970**	**1960p**
		2	**China**	NBS	**2000**	**1990**	**1982**		1964
3		2	***Colombia**	IPUMS	**2005**	**1993**	**1985**	**1973**	**1964p**
3	1		***Costa Rica**	IPUMS	**2000**		**1984**	**1973**	**1963**
1			**Cuba**	IPUMS	**2002**		1981	1970	
4		1	***Ecuador**	IPUMS	**2001**	**1990**	**1982**	**1974**	**1962p**
3			**Egypt**	IPUMS	**2006§**	**1996**	**1986**	1976	1964
1	6		**France**	INSEE	**2006§**	**1990,9**	**1982**	**1975**	**1968,2**
2			***Ghana**	IPUMS	**2000**		**1984**	1970	
4			**Greece**	IPUMS	**2001**	**1991**	**1981**	**1971**	1961
2			***Guinea, C.**	IPUMS		**1996**	**1983**		1960
	4		**Hungary**	CSO	**2001**	**1990**	**1980**	**1970**	
		5	**India**	NSSO	**2005m**	**1993,9m**	**1983,7m**		
1			***Iraq**	IPUMS		**1997**	1987	1977	1967
5			**Israel**	CBS	**2008**	**1995**	**1983**	**1972**	**1961**,7
	1		**Italy**	ISTAT	**2001**	**1991**	**1981**	1971	1961
1			**Jordan**	IPUMS	**2004**	**1994**	1979		
	3		**Kenya**	IPUMS	**1999**	**1989**	**1979**	**1969**	
1			**Kyrgyz Rep.**	IPUMS	**2009**	**1999**	1989		
		4	**Malaysia**	IPUMS	**2000**	**1991**	**1980**	**1970**	1960
3			***Mali**	IPUMS	**2008**	**1998**	**1987**	**1976**	
4		3	**Mexico**	INEGI	**2000,5**	**1990,5**	**1980**	**1970**	**1960p**
2			***Mongolia**	IPUMS	**2000**		**1989**	1979	**1956**
1			**Nepal**	CBS	**2001**	1991?	1981	1971	1961
		3	**Netherlands**	CBS	**2001pm**			**1971p**	**1960p**
2			**Palestine**	CBS	**2007§**	**1997**			
3			***Pakistan**	IPUMS		**1998**	**1981**	**1973**	1961
5			***Panama**	IPUMS	**2000**	**1990**	**1980**	**1970**	**1960**

Table 1. (*continued*)

2			**Peru**	IPUMS	**2007**	**1993**	**1981**	1972	1961
3			***Philippines**	IPUMS	**2000**	**1990**	**1980**	**1970**	**1960p**
	3		**Portugal**	INE	**2001**	**1991**	**1981**	1970	1960
	4		**Puerto Rico**	USCB	**2000**	**1990**	**1980**	**1970**	1960
3			**Romania**	IPUMS	**2001**	**1992**		**1977**	1965
2			***Rwanda**	IPUMS	**2002**	**1991**			
2			***Saint Lucia**	IPUMS	**2001**	**1991**	**1980**	1970	1960
3			***Senegal**	IPUMS	**2002**		**1988**	**1976**	
1			**Slovenia**	SORS	**2001**	1991	1981		
6		1	**South Africa**	StatsSA	**2001,7**	**1996-1**	**1985-0**	**1970**	1960
	3		**Spain**	INE	**2001**	**1991**	**1981**	1970	1960
	4		**Switzerland**	IPUMS	**2000**	**1990**	**1980**	**1970**	1960
2			***Tanzania**	IPUMS	**2002**		**1988**	1978	1967
		4	**Thailand**	NSO	**2000**	**1990**	**1980**	**1970**	1960
2			***Uganda**	IPUMS	**2002**	**1991**	1980		1969
		2	**United King.**	ONS	**2001p**	**1991**	**1981**	**1971**	**1966,1**
	6		**USA**	USCB	**2000,5**	**1990**	**1980**	**1970**	**1960**
4			***Venezuela**	IPUMS	**2001**	**1990**	**1981**	**1971**	1961
	2		**Vietnam**	IPUMS	**2009**	**1999**	**1989**	1979	
Europe									
			Albania	-	**2001**	1989	1979	1969	1960
			Bulgaria	-	**2001**	**1992**	**1985**	1975	1965
			Belgium	-	**2001**	**1991**	**1981**	**1970**	1961
	2		**Czech Rep.**	IPUMS	**2001**	**1991**	**1980**	**1970**	1961
			Estonia	-	**2000**	**1989**	1979	1970	1959
4			**Germany §**	FSO	**2001m**	**1991m**	**1981-7**	**1970,1**	1961
8			**Ireland §**	CSO	**2002, 6**	**1991, 6**	**1981, 6**	**1971,9**	
			Latvia	-	**2000**		**1989**	1979	
			Poland	-	**2001**	**1995**	**1988**	**1970,8**	1960
			Russia	-	**2002**		**1989**	1979	1970
			Turkey	TurkSTAT	**2000**	**1990**	**1985, 0**	1975,0	1960
			Ukraine	IPUMS	**2001**		1989	1979	1970
North America and the Caribbean									
1	1	2	***DominicanR.**	IPUMS	**2003**	1993	**1981**	**1970**	**1960p**
1			***El Salvador**	IPUMS	**2007**	**1992**		1971	1961
2		3	***Guatemala**	IPUMS	**2002**	**1994**	**1981**	**1973**	**1964**
3			***Jamaica§**	IPUMS	**2001**	**1991**	**1982**	1970	1960
2			***Haiti**	IPUMS	**2003**		**1982**	**1971**	
3		1	***Honduras**	IPUMS	**2000**		**1988**	**1974**	**1961**
2		1	***Nicaragua §**	IPUMS	**2005**	**1995**		**1971**	1963
South America									

Table 1. (*continued*)

4		1	***Paraguay**	IPUMS	**2002**	**1992**	**1982**	**1972**	**1962**
4			***Uruguay**	IPUMS		**1996**	**1985**	**1975**	**1963**
Africa									
			Benin		**2002**	**1990**		**1979**	
3			***Botswana**	IPUMS	**2001**	**1991**	**1981**	1971	1964
			Burkina Faso		**2006**	**1996**	**1985**	1975	
			Burundi		**2008**	1990?	1979?	1970?	
			Cameroon		**2005**		**`1987**	**1976**	
			Cape Verde	IPUMS	**2000**	**1990**	1980	1970	1960
			Central Afr. R.		**2003**		**1988**	1974	
			Chad		**2008**	**1993**	**1989**		1969
			Côte d'Ivoire		**2009**	**1998**	**1988**	1975	
2			***Ethiopia**	IPUMS	**2007**	**1994**	**1984**		
			Gabon		**2003**	**1993**	1980		1969
			Guinea-Bis.	IPUMS	**2009**	**1991**		1979	
2			**Lesotho**	IPUMS	**2006**	**1996**	**1986**	**1976**	1966
			Liberia		**2008**		1984	**1974**	
1			***Madagascar**	IPUMS		**1993**			
2			***Malawi**	IPUMS	**2008**	**1997**	**1987**	1977	1967
			Mauritania		2001		1988	1977	
2			***Mauritius**	IPUMS	**2000**	**1990**	**1983**	1972	1962
	3		**Morocco**	IPUMS	**2004**	**1994**	**1982**	1971	1960
1			**Mozambique**	IPUMS	**2007**	**1997**	1980		
2			***Niger**	IPUMS	**2001**		**1987**	**1977**	
			Nigeria	NatPopCom	**2006**	**1991**		1973	1963
1			***Sierra L.§**	IPUMS	**2004**		1985	1974	1963
3			***Sudan**	IPUMS	**2008**	**1993**	**1983**	**1973**	
			Togo		2010		**1981**	**1970**	**1958**
2			***Zambia**	IPUMS	**2000**	**1990**	1980	1969	1963
Asia and Oceania									
1		1	***Bangladesh**	IPUMS	**2001**	**1991**	**1981**	1974	1961
5			***Fiji Islands**	IPUMS	**2007**	**1996**	**1986**	**1976**	**1966**
8			**Indonesia §**	BP/IPUMS	**2000, 5**	**1990, 5**	**1980, 5**	**1971,6**	1961
1			Iran §	SCI	**2006**	**1996**	**1986**	**1976**	**1966**
			Korea, Rep.	KOSTAT	**2005, 0**	**1995, 0**	**1985, 0**	**1975**	**1960,6**
			Sri Lanka	DCS	**2001**		**1981**	**1971**	1960
1			**Turkmenistan**	IPUMS		**1995**	1989	1979	1970
			United A. E.		**2005**	1995	1985, 0	1975	1968

bold country = Memorandum of Understanding with Regents of the University of Minnesota;
IPUMS = systematic household sample: every n^{th} household stratified by enumeration district; confidentiality specifications (see text).
Year = census conducted; **Bold year** = microdata survive; §= samples for launch in 2011
*** =** 100% microdata entrusted, where extant; m = microcensus; p = person sample

(3) Any social characteristic (categorical variables such as place of birth, occupation, etc.) with fewer than 250 individuals in the population will be re-coded as missing, suppressed or aggregated.
(4) Continuous variables (such as income, size of rooms, etc.) will be top/bottom coded to prevent identification of individuals or other entities with unique characteristics.

(5) The geographical identifiers of a fraction of households will be recoded to a different geographical unit so that any allegation that an individual or other entity is positively identified is false. Swapping of individuals and households across geographical boundaries (that is, editing the geographical identifiers of a small fraction of individuals and households to one that is false) introduces uncertainty into any attempts at identification.

The thresholds in this contract are those usually authorized by most statistical agencies that entrust "raw" microdata to the IPUMS-International project. Nonetheless, the thresholds may be adjusted at the request of the National Statistical Office-owner. For example, in the case of France, place of residence is limited to 22 regions. The smallest region has a population exceeding 80,000 in the 1990 census (sample n > 4,000). The population count for any identifiable single year of age is >2000. For any identifiable country of citizenship the threshold is also >2000. Recently, INSEE, the French national statistical authority, began a reconsideration of these thresholds, particularly for the historical datasets that are now more than a decade old. A comprehensive assessment is being prepared to develop lower thresholds so that detailed attributes may be made available for several key variables, such as place of residence, country of birth, occupation, and industry.

During the process of confidentializing international microdata at the Minnesota Population Center, all work is performed by senior staff who have taken the appropriate training and signed official statements to protect the data. Once the statistical disclosure controls are in place, junior staff may begin integrating the microdata. Original source microdata, whether "raw" or confidentialized, are encrypted and archived off-line and thus are preserved in case there are questions about errors introduced by the SDCs. To date there have been no queries about the validity of any IPUMS samples. Errors have been discovered—some due to the integration and others in the source microdata, but none attributed to the process of confidentializing samples.

4 An Evaluation of Security

Statistical data privacy is more than simply SDC. All procedures and processes associated with the microdata must be secure and must be perceived as such by the public. With the large stock of international microdata archived at the Minnesota Population Center, protecting these assets is a major concern of the Center, the University, and official statistical agencies, international as well as national—whether associated with the project or not.

The first author of this paper invited Mr. Dennis Trewin to conduct an on-site inspection of the IPUMS-International facilities and procedures [10]. Mr. Trewin is well qualified for the undertaking. As Australian Statistician one of his achievements was the extension of microdata services to researchers while maintaining public trust and abiding by the conditions outlined in the legislation of Australia governing microdata access. He also chaired the Conference of European Statisticians Task Force on Guidelines and Core Principles of Confidentiality and Microdata Access. The Guidelines were adopted by the CES plenary session in June 2006 and published

as Managing Statistical Confidentiality & Microdata Access: Principles and Guidelines of Good Practice [4]. Not surprisingly, Mr. Trewin is noted for his critical acumen and professional probity. The terms of reference for his review was to identify weaknesses and lapses so that IPUMS-International could improve its procedures and thereby provide an additional layer of protection for official statisticians as well as trust for the public. Mr. Trewin's review included attendance at a side-meeting of the IPUMS-International at the 2007 International Statistical Institute in Lisbon as well as bilateral interviews with official statisticians. His report concludes:

> Without question IPUMS International meets the four Core Principles outlined in CES (2007). It is cited in CES (2007) as a Case Study of good practice. This review confirms its status as good practice for Data Repositories. Indeed it is likely to provide the best practice for a Data Repository of international statistical data sets.
> …
> The security of the computing environment used by IPUMS-International is first class and appears to be of the standard of the best statistical offices.
> …
> IPUMS-International is a valuable and trustworthy microdata service. It meets the fundamental principles of good practice with respect to confidentiality and microdata. Consequently, my recommendations are limited.

Mr. Trewin's recommendations to IPUMS-International for enhancing security and data confidentiality are, indeed, "limited". Nonetheless all have been or are being implemented, including his recommendation that "checks should be made of published outputs from time to time to provide some assessment of whether there has been any inappropriate use of microdata (e.g., reference to individual cases)."

5 Conclusion

The goals of IPUMS-International are, first, to recover census microdata that are at risk of loss; second, to archive microdata; and third, to disseminate confidentialized, custom-tailored, integrated extracts to researchers world-wide at no cost. In the first decade of operations, more than 3,700 researchers registered for access, a vast trove of microdata was entrusted to the Minnesota Population Center, and 159 datasets underwent the arduous process of confidentializing the microdata and integrating both data and documentation into the IPUMS-International system. Over the next five years, an additional hundred datasets are likely to be integrated into the IPUMS-International system. Academics and policy makers needing census microdata for research are invited to visit the project website: www.ipums.org/international.

Acknowledgements. Funded in part by the National Science Foundation of the United States, Grant Nos. SES-0433654 and 0851414; National Institutes of Health, Grant Nos. R01HD047283 and R01 HD044154.

References

1. McCaa, R., Ruggles, S., Davern, M., Swenson, T., Mohan Palipudi, K.: IPUMS-International High Precision Population Census Microdata Samples: Balancing the Privacy-Quality Tradeoff by Means of Restricted Access Extracts. In: Domingo-Ferrer, J., Franconi, L. (eds.) PSD 2006. LNCS, vol. 4302, pp. 375–382. Springer, Heidelberg (2006)
2. Museux, J.-M., Peeters, M., Santos, M.-J.: Legal, Political and Methodological Issues in Confidentiality in the European Statistical System. In: Domingo-Ferrer, J., Saygın, Y. (eds.) PSD 2008. LNCS, vol. 5262, pp. 324–334. Springer, Heidelberg (2008)
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, pp. 37–46. Office for Official Publications of the European Communities, Luxembourg (2006)
4. United Nations Economic Commission for Europe. Conference of European Statisticians. Managing Statistical Confidentiality & Microdata Access: Principles and Guidelines of Good Practice. Geneva: United Nations (2007), http://www.unece.org/stats/publications/Managing.statistical.confidentialityandmicrodataaccess.pdf
5. United Kingdom. Office of National Statistics. National Statistics Code of Practice Protocol on Data Access and Confidentiality. HMSO, London (2004)
6. De Kort, S., Wathan, J.: Guide to Imputation and Perturbation in the Samples of Anonymised Records (2009) (unpublished)
7. Purdam, K., Elliot, M.: A case study of the impact of statistical disclosure control on data quality in the individual UK Samples of Anonymised Records. Environment and Planning A 39(5), 1101–1118 (2007)
8. Alexander, J.T., Davern, M., Stevenson, B.: Inaccurate Age and Sex Data in the (United States) Census PUMS Files: Evidence and Implications. NBER Working Paper No. 15703 (2010) (Forthcoming Public Opinion Quarterly)
9. Wolfers, J.: Can You Trust Census Data? Freakonomics blog. New York Times (February 2, 2010), http://freakonomics.blogs.nytimes.com/2010/02/02/can-you-trust-census-data
10. Trewin, D.: A Review of IPUMS-International (2007) (unpublished), http://www.hist.umn.edu/~rmccaa/IPUMSI/trewin_ipums_report.pdf

Uncertainty for Anonymity and 2-Dimensional Range Query Distortion

Spyros Sioutas[1], Emmanouil Magkos[1], Ioannis Karydis[1], and Vassilios S. Verykios[2]

[1] Department of Informatics, Ionian University, Corfu, Greece
{sioutas,emagos,karydis}@ionio.gr
[2] Department of Computer and Communication Engineering, University of Thessaly, Greece
verykios@inf.uth.gr

Abstract. In this work, we study the problem of anonymity-preserving data publishing in moving objects databases. In particular, the trajectory of a mobile user on the plane is no longer a polyline in a two-dimensional space, instead it is a two-dimensional surface: we know that the trajectory of the mobile user is within this surface, but we do not know exactly where. We transform the surface's boundary poly-lines to dual points and we focus on the information distortion introduced by this space translation. We develop a set of efficient spatio-temporal access methods and we experimentally measure the impact of information distortion by comparing the performance results of the same spatio-temporal range queries executed on the original database and on the anonymized one.

Keywords: Uncertainty, Privacy, Anonymity, Moving Objects Databases.

1 Introduction

The technological advances in sensors and wireless communications have made possible the offering of high accuracy in location tracking at a low cost [3], [4], [8]. The increased location accuracy gave rise to a series of location-based applications that exploit positional data to offer high-end services to their subscribers [5]. We consider a population of mobile users who are supported by some telecommunication infrastructure, owned by a telecom operator. Every user periodically transmits through his/her mobile device a location update to some traffic monitoring system residing in a trusted server of the telecom operator. The transformation of the exact user location to a spatiotemporal area is achieved through the use of k-anonymity. The k-anonymity principle for relational data [15], [16] requires that each record in a given dataset is indistinguishable from at least $k-1$ other records with respect to a certain set of identifying variables, collectively known as the "quasi-identifier". The k-anonymity principle requires that the spatiotemporal area that is generated by the trusted server from the exact location of the mobile user is such that the identity of the user cannot be disclosed with a probability that is larger than $1/k$, among $k-1$ other users.

J. Domingo-Ferrer and E. Magkos (Eds.): PSD 2010, LNCS 6344, pp. 85–96, 2010.

This trusted server is requested to efficiently answer Range Queries (RQs) of mobile users moving on the plane.

In our privacy model we assume an attacker who has knowledge of (i) the frequent movement behavior of all the users in the system, computed by the trusted server as part of its functionality, (ii) the anonymized location updates of the users, as received and anonymized by the trusted server, and (iii) the algorithms used by the trusted server to support user privacy. Our solution is capable of ensuring the privacy of the users, even in the case that all this diverse knowledge is at the disposal of the attacker. k-anonymity is essential to protect the privacy of the users, starting from the point of request for a RQ service and continuing for as long as the requested service withstands completion. As part of our framework, we deliver the type of k-trajectory anonymity [6] that identifies $k-1$ users that are close to the requester at the time of request and thus could have issued the request. This includes a minimum circular spatial area A_{min} around the requester, where the participants of the anonymity set should be searched for so that the user is adequately covered up. The proposed framework deals with the event of failure in the provision of k-anonymity, in the case where the number of participants inside this minimum spatial area is less than k-1. In this case, the trusted server postpones the servicing of the user request for a small period of time. After that, if the anonymization process fails again, the requester is protected by blocking the servicing of the request. The proposed privacy framework relies on a user privacy profile that stores the necessary information related to his/her privacy requirements. This includes (i) the preferred value of k (in k-anonymity) for each requested RQ service, (ii) the minimum circular spatial area A_{min}, around the requester, where the participants of the anonymity set should be searched for so that the user is adequately covered up. This threshold defines the minimum extent of the spatial area that must replace the real location of the user, in the anonymized request.

Based on the proposed privacy model we implement a framework that uses the Spatial extensions of MySql 5.x to offer privacy in RQ services. This type of queries focuses on the problem of indexing mobile users in two dimensions and efficiently answering range queries over the users locations. This problem is motivated by a set of real-life applications such as intelligent transportation systems, cellular communications, and meteorology monitoring. There are two basic approaches used when trying to handle this problem; those that deal with discrete and those that deal with continuous movements. In a discrete environment the problem of dealing with a set of moving objects can be considered to be equivalent to a sequence of database snapshots of the object positions/extents taken at time instants $t_1 < t_2 < \ldots$, with each time instant denoting the moment where a change took place. From this point of view, the indexing problems in such environments can be dealt with by suitably extending indexing techniques from the area of temporal [17] or/and spatial databases [7]; in [12] it is elegantly exposed how these indexing techniques can be generalized to handle efficiently queries in a discrete spatiotemporal environment. When considering continuous movements there exists a plethora of efficient data structures [9,11,13,14,18]. The

common thrust behind these indexing structures lies in the idea of abstracting each object's position as a continuous function $f(t)$ of time and updating the database whenever the function parameters change; accordingly an object is modeled as a pair consisted of its extent at a reference time (design parameter) and of its motion vector. One categorization of the aforementioned structures is according to the family of the underlying access method used. In particular, there are approaches based either on R-trees or on Quad-trees. On the other hand, these structures can be also partitioned into (a) those that are based on geometric duality and represent the stored objects in the dual space [11,14] and (b) those that leave the original representation intact by indexing data in their native n-d space [2,13,18]. The *geometric duality transformation* is a tool heavily used in the Computational Geometry literature, which maps hyper-planes in R^n to points and vice-versa.

In this work, we study the problem of anonymity-preserving data publishing in moving objects databases. In particular, the trajectory of a mobile user on the plane is no longer a polyline in a two-dimensional space, instead it is a two-dimensional surface: we know that the trajectory of the mobile user is within this surface, but we do not know exactly where. We transform the surface's boundary poly-lines to dual points [11,13] and we focus on the information distortion introduced by this space translation. We develop a set of efficient spatio-temporal access methods and, we experimentally measure the impact of information distortion by comparing the performance results of the same spatio-temporal range queries executed on the original database and on the anonymized one.

In Section 2 we give a formal description of the problem. In Section 3 we present the method of transforming the trajectory poly-lines to two-dimensional surfaces. In Section 4 we present the duality transformation of surface's boundary poly-lines and we focus on the information distortion introduced by this space translation. Section 5 presents an extended experimental evaluation and section 6 concludes the paper.

2 Definitions and Problem Description

We consider a database that records the position of moving objects in two dimensions on a finite terrain. We assume that objects move with a constant velocity vector starting from a specific location at a specific time instant. Thus, we can calculate the future object position, provided that its motion characteristics remain the same. Velocities are bounded by $[u_{min}, u_{max}]$. Objects update their motion information, when their speed or direction changes. The system is dynamic, i.e. objects may be deleted or new objects may be inserted. Let $P_z(t_0) = [x_0, y_0]$ be the initial position at time t_0 of object z. If object z starts moving at time $t > t_0$, its position will be $P_z(t) = [x(t), y(t)] = [x_0 + u_x(t - t_0), y_0 + u_y(t - t_0)]$, where $U = (u_x, u_y)$ is its velocity vector. For example, in Figure 1 the lines depict the objects trajectories on the (t, y) plane. We would like to answer queries of the form: "Report the objects located inside the rectangle $[x_{1_q}, x_{2_q}] \times [y_{1_q}, y_{2_q}]$ at the time instants between t_{1_q} and t_{2_q} (where $t_{now} \leq t_{1_q} \leq t_{2_q}$), given the current motion information of all objects".

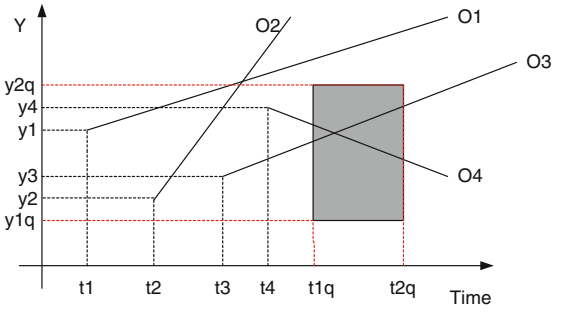


Fig. 1. Trajectories and query in (t, y) plane

3 Trajectory Poly-Lines and Two-Dimensional Surfaces

For every mobile user, we calculate a circular range query with center its current 2-D location and radious a given value R defined by a minimum circular spatial area A_{min}. If this circular spatial area includes at least $k-1$ other neighbours, then the mobile user is adequately covered up. Otherwise, if the number of participants inside this minimum spatial area is less than k-1, the trusted server postpones the servicing of the user request for a small period of time. After that, if the anonymization process fails again, the requester is protected by blocking the servicing of the request. As a result, consecutive circular spatial areas construct a 2-D buffer defined by its upper and lower boundary poly-lines y' and y'' respectively, which anonymize the original trajectory y of mobile user A (see figure 2). By using the Spatial extensions of MySql 5.x we can create each mobile user as 2-dimensional point as follows:

```
CREATE TABLE Points (
name VARCHAR(20) PRIMARY KEY,
location Point NOT NULL,
description VARCHAR(200),
SPATIAL INDEX(location)
);
```

In order to obtain points in a circular area as a counted result ordered by distance from the center of the selection area, we write:

```
SELECT COUNT(name), AsText(location), SQRT(POW( ABS( X(location)
- X(@center)), 2) + POW( ABS(Y(location) - Y(@center)), 2 )) AS distance
FROM Points
WHERE Intersects( location, GeomFromText(@bbox) )
AND SQRT(POW( ABS( X(location) - X(@center)), 2) +
POW( ABS(Y(location) - Y(@center)), 2 )) ≤ @R
ORDER BY distance;
```

If the result returned is less than k-1, the trusted server postpones the servicing of the user request for a small period of time.

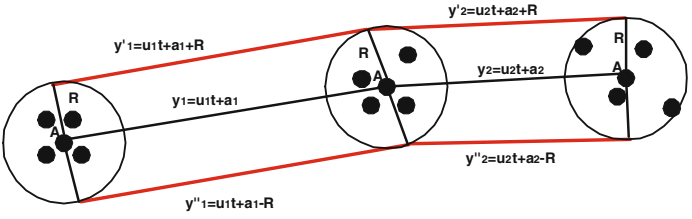


Fig. 2. 2-D Buffer and Boundary Trajectories y' and y'' of mobile user A

So, the RQ service depicted in figure 1, was transformed to the Privacy-Aware RQ service depicted in the figure 3:

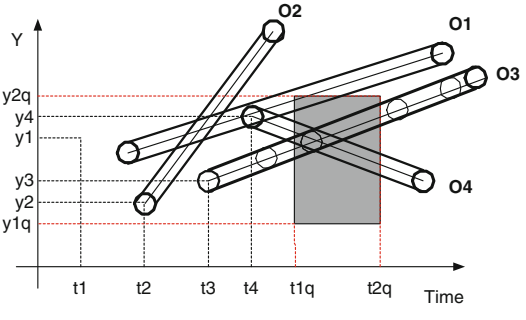


Fig. 3. Boundary Trajectories and query in (t, y) plane

If the whole buffer lies inside the query area (see the buffers of mobile users O_3 or O_4 in figure 3), meaning that both upper and lower boundary poly-lines lie in the query rectangle then the same holds for the original trajectory. If the whole buffer lies outside the query area (see the buffer of mobile user O_2 in figure 3), meaning that both upper and lower boundary poly-lines lie outside the query rectangle then the same holds for the original trajectory. In the worst-case, we face the distortion effect, where one of the two boundary poly-lines lie in the query rectangle (see the buffer of mobile user O_1 in figure 3). In the later case, we have to check what happens with the original trajectory.

4 Duality Transformation of Boundary-Trajectories and Information Distortion

The duality transform, in general, maps a hyper-plane h from R^n to a point in R^n and vice-versa. In this subsection we briefly describe how we can address the

problem at hand in a more intuitive way, by using the duality transform on the 1-d case.

4.1 Hough-X Transform

One duality transform for mapping the line with equation $y(t) = ut + a$ to a point in R^2 is by using the dual plane, where one axis represents the slope u of an objects trajectory (i.e. velocity), whereas the other axis represents its intercept a. Thus we get the dual point (u, a) (this is the so called *Hough-X transform* [11,13]). Accordingly, the 1-d query $[(t_{1_q}, t_{2_q}), (y_{1_q}, y_{2_q})]$ becomes a polygon in the dual space. By using a linear constraint query, the initial query $[(t_{1_q}, t_{2_q}), (y_{1_q}, y_{2_q})]$ in the (t, y) plane is transformed to the following rectangular query $[(u_{min}, u_{max}), (y_{1_q} - t_{1_q} u_{max}, y_{2_q} - t_{2_q} u_{min})]$ in the (u, a) plane. In a similar way, for the upper $(y'(t) = ut + a + R)$ and lower $(y''(t) = ut + a - R)$ boundary trajectories, we get the dual points $(u, a + R)$ and $(u, a - R)$ as well as the final (transformed) rectangular queries become $[(u_{min}, u_{max}), (y_{1_q} - R - t_{1_q} u_{max}, y_{2_q} - R - t_{2_q} u_{min})]$ and $[(u_{min}, u_{max}), (y_{1_q} + R - t_{1_q} u_{max}, y_{2_q} + R - t_{2_q} u_{min})]$ respectively in the (u, a) plane.

4.2 Hough-Y Transform

By rewriting the equation $y = ut + a$ as $t = \frac{1}{u} y - \frac{a}{u}$, we can arrive to a different dual representation (the so called *Hough-Y transform* in [11,13]). The point in the dual plane has coordinates (b, n), where $b = -\frac{a}{u}$ and $n = \frac{1}{u}$. Coordinate b is the point where the line intersects the line $y = 0$ in the primal space. By using this transform horizontal lines cannot be represented. Similarly, the Hough-X transform cannot represent vertical lines. Nevertheless, since in our setting lines have a minimum and maximum slope (velocity is bounded by $[u_{min}, u_{max}]$), both transforms are valid. Similarly, the initial query $[(t_{1_q}, t_{2_q}), (y_{1_q}, y_{2_q})]$ in the (t, y) plane can be transformed to the following rectangular query in the (b, n) plane: $[(t_{1_q} - \frac{y_{2_q}}{u_{min}}, t_{2_q} - \frac{y_{1_q}}{u_{max}}), (\frac{1}{u_{max}}, \frac{1}{u_{min}})]$. In a similar way for the upper $(y'(t) = ut + a + R)$ and lower $(y''(t) = ut + a - R)$ boundary trajectories, we get the transformed rectangular queries $[(t_{1_q} - \frac{y_{2_q} - R}{u_{min}}, t_{2_q} - \frac{y_{1_q} - R}{u_{max}}), (\frac{1}{u_{max}}, \frac{1}{u_{min}})]$ and $[(t_{1_q} - \frac{y_{2_q} + R}{u_{min}}, t_{2_q} - \frac{y_{1_q} + R}{u_{max}}), (\frac{1}{u_{max}}, \frac{1}{u_{min}})]$ respectively in the (b, n) plane.

4.3 The Proposed Algorithm for Privacy-Aware Indexing

Let $S = \{y_1, y_2, \ldots, y_n\}$ be the initial set of original trajectory equations, and $S' = \{y'_1, y''_1, y'_2, y''_2 \ldots, y'_n, y''_n\}$ the set of boundary trajectory equations defined by the buffer.

Algorithm 1. Index-Building

1: Decompose the 2-d motion into two 1-d motions on the (t, x) and (t, y) planes;
2: For each projection, build the corresponding index structure;

Then, let $H_xS = \{(u_1, a_1 + R), (u_1, a_1 - R), \ldots, (u_n, a_i + R), (u_n, a_i - R)\}$ and $H_yS = \{b'_1, b''_1, \ldots, b'_n, b''_n\}$ be the set of dual Hough-X and Hough-Y transforms respectively.

Algorithm1 depicts the procedure for building the index, Algorithm2 presents the procedure for Partitioning the Mobile Users according to their velocity, and Algorithm3 outlines the privacy-aware algorithm for answering the exact 2-d RQ query:

Algorithm 2. Mobile-User-Partitioning

1: Users with small velocity are stored using the Hough-X dual transform;
2: The rest are stored using the Hough-Y dual transform;
3: Motion information about the other projection is also included;

Algorithm 3. Privacy-Aware-RQ Query

1: Decompose the query into two 1-d queries, for the (t, x) and (t, y) projection;
2: For each projection get the dual-simplex query;
3: For each projection calculate a specific criterion c (for details see [11,13]) and choose the one (say p) that minimizes it;
4: For all dual-points of H_xS or H_yS sets, search in projection p the simplex query of the Hough-X or the MBR of the simplex query of the Hough-Y partition. In the latter case, perform a refinement or filtering step "on the fly", by using the whole motion information;
5: if the dual-points of both upper and lower boundary trajectories $((u_i, a_i + R), (u_i, a_i - R)$ or $b'_i, b''_i)$ lie inside the dual-simplex spatial area then the same holds for the dual-point $((u_i, a_i)$ or $b_i)$ of the original trajectory;
6: else if the dual-points of both upper and lower boundary trajectories lie outside the dual-simplex spatial area then the same holds for the dual-point of the original trajectory;
7: else having in mind the value R, search the simplex query of the Hough-X or Hough-Y partition for the dual-points of original trajectories;

In [11,13], $Q_{Hough-X}$ is computed by querying a 2-d partition tree, whereas $Q_{Hough-Y}$ is computed by querying a B^+-tree that indexes the b parameters. Here, we consider the case, where the users are moving with non small velocities u, meaning that the velocity value distribution is skewed (Zipf) towards u_{min} in some range $[u_{min}, u_{max}]$ and as a consequence the $Q_{Hough-Y}$ transformation is used (denote that u_{min} is a positive lower threshold). Moreover, our method will incorporate the Lazy B-tree [10] indexing scheme, since the latter can handle update queries in optimal (constant) number of block-transfers (I/Os). As a result, we get Algorithm 4.

Let K be the number of b_i parameters associated to boundary trajectories of buffers that intersect with the query rectangle. Then, algorithm 4 requires $T(n) = O(Cost(LazyB\_tree) + K)$ I/Os or block transfers. Moreover, and according to notations presented in [1], let say D be the initial Database that

Algorithm 4. Privacy-Aware Indexing of $Q_{Hough-Y}$ partition with Lazy B-tree

1: BEGIN_PSEUDOCODE
2: Decompose the query into two 1-d queries, for the (t, x) and (t, y) projection;
3: For each projection get the dual-simplex query;
4: Search the MBR of the simplex query of the Hough-Y partition and perform a filtering step "on the fly", by using the whole motion information;
5: Let $B \subset H_y S$ be the answer set of dual parameters, which satisfy the query above;
6: $Result = 0$;
7: For all elements of B do
8: Begin_for
9: if ($b_i' \in B$ AND $b_i'' \in B$) then
$Result = Result \cup (idofuser_i, neighbours(idofuser_i, k-1)$;
10: return Result;
11: else if ($b_i' \notin B$ AND $b_i'' \notin B$) then return Result;
12: else begin
13: if $b_i \in MBR$ of Hough-Y simplex partition then
14: $Result = Result \cup (idofuser_i, neighbours(idofuser_i, k-1)$;return Result;
15: else return Result;
16: end
17: End_for
18: END_PSEUDOCODE

stores the N original trajectories and D' the Privacy-Aware Database that stores the $2N$ boundary-trajectories. Let also say, $Q(D)$ and $Q(D')$ be the Query Results obtained consuming $T(D)$ and $T(D')$ block-transfers (I/Os) in D and D' database schemes respectively. We define as $Distortion\_Ratio = \frac{|Q(D)-Q(D')|}{max(Q(D),Q(D'))}$ and as $Competitive\_Ratio = \frac{|T(D)-T(D')|}{max(T(D),T(D'))}$. In the most of the cases, $T(D') > \ldots > T(D)$, thus it is very important to find out, how competitive to the optimal one ($T(D)$) is the privacy-aware method that answers the query in D'. Since, the distortion effect in D' absolutely depends on parameter K, an experimental evaluation of ***Competitive_Ratio vs*** K is also presented in the following section.

5 Experimental Evaluation

This section compares the query performance of our privacy-aware method, when incorporates STRIPES [14] (the best known solution), Lazy B-trees (LBTs) and TPR*-tree, respectively. We deploy spatio-temporal data that contain insertions at a single timestamp 0. In particular, objects' MBRs are taken from the LA spatial dataset (128971 MBRs)[1]. We want to simulate a situation where all objects move in a space with dimensions 100x100 kilometers. For this purpose each axis of the space is normalized to [0,100000]. For the TPR*-tree, each object is associated with a VBR (Velocity Bounded Rectangle) such that (a) the object

[1] Downloaded from the Tiger website `http://www.census.gov/geo/www/tiger/`

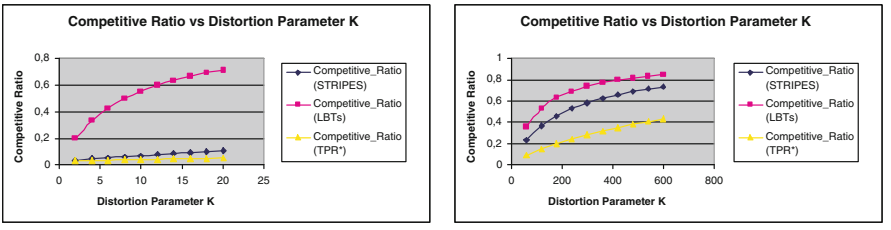


Fig. 4. $q_V len = 5$, $q_T len = 50$ $q_R len = 100$ (top), $q_R len = 1000$ (bottom), $R_{max} = 50$ (top), $R_{max} = 200$ (bottom)

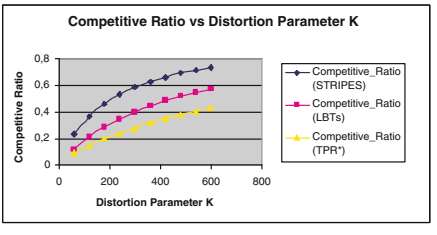


Fig. 5. $q_R len = 2000$, $q_V len = 5$, $q_T len = 50$, $R_{max} = 500$

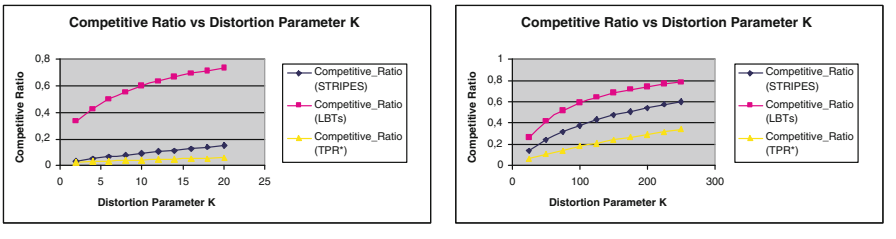


Fig. 6. $q_V len = 10$, $q_T len = 50$, $q_R len = 400$ (top), $q_R len = 1000$ (bottom), $R_{max} = 100$ (top), $R_{max} = 200$ (bottom)

does not change spatial extents during its movement, (b) the velocity value distribution is skewed (Zipf) towards 30 in range [30,50], and (c) the velocity can be either positive or negative with equal probability. As in [2], we will use a small page size so that the number of index nodes simulates realistic situations.

Thus, for all experiments, the page size is 1 Kbyte, the key length is 8 bytes, whereas the pointer length is 4 bytes. Thus, the maximum number of entries ($< x >$ or $< y >$, respectively) in both Lazy B-trees and B$^+$-trees is 1024/(8+4)=85. In the same way, the maximum number of entries (2-d rectangles or $< x1, y1, x2, y2 >$ tuples) in TPR*-tree is 1024/(4*8+4)=27. On the other hand, the STRIPES index maps predicted positions to points in a dual transformed space and indexes this space using a disjoint regular partitioning

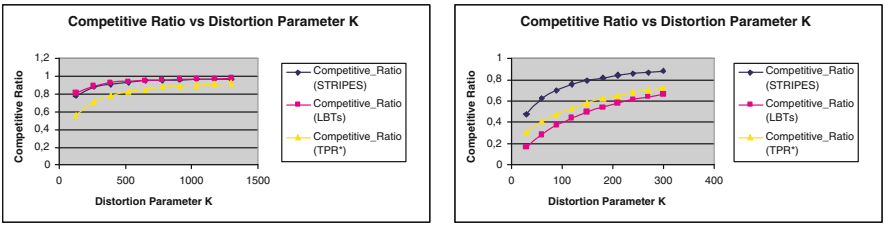


Fig. 7. $q_V len = 5$, $q_T len = 1$, $q_R len = 400$ (top), $q_R len = 1000$ (bottom), $R_{max} = 100$ (top), $R_{max} = 200$ (bottom)

of space. Each of the two dual planes, are equally partitioned into four quads. This partitioning results in a total of $4^2 = 16$ partitions, which we call *grids*. The fanout of each non-leaf node is thus 16. For each dataset, all indexes except for STRIPES have similar sizes. Specifically, for LA, each tree has 4 levels and around 6700 leaves apart from STRIPES index which has a maximum height of seven and consumes about 2.4 times larger disk space. Each query q has three parameters: $q_R len$, $q_V len$, and $q_T len$, such that (a) its MBR q_R is a square, with length $q_R len$, uniformly generated in the data space, (b) its VBR is $q_V = -q_V len/2, q_V len/2, -q_V len/2, q_V len/2$, and (c) its query interval is $q_T = [0, q_T len]$. The query cost is measured as the average number of node accesses in executing a workload of 200 queries with the same parameters. Implementations were carried out in C++ including particular libraries from SECONDARY LEDA v4.1.

5.1 Query Cost Comparison

We measure the Competitive Ratio of LBTs method (this method incorporates two Lazy B-trees that index the appropriate b parameters in each projection respectively, and finally combines the two answers by detecting and filtering all the pair permutations), the TPR*-tree and STRIPES presented in [18] and [14] respectively, using the same query workload, after every 10000 updates. Figures 4 up to 8 show the Competitive Ratio as a function of K (for datasets generated from LA as described above), using workloads with different parameters. Parameter K represents boundary trajectories of buffers that intersect with the query rectangle, and obviously require an additional filtering on the fly process. Obviously, the required number of block transfers depends on the answer's size as well as the size of K.

Figure 4 depicts how competitive to the optimal solution the LBTs method is, in comparison to TPR*-tree and STRIPES. The Ratio degrades as the query rectangle length grows from 100 to 1000. When the query rectangle length or equivalently the query surface becomes extremely large (e.g. 2000), then the STRIPES index becomes more competitive (see Figure 5).

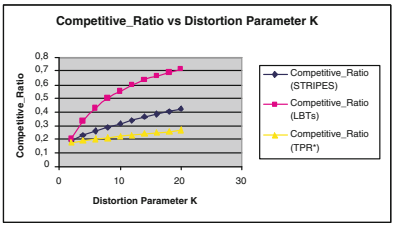


Fig. 8. $q_R len = 400$, $q_V len = 5$, $q_T len = 100$, $R_{max} = 100$

Figure 6 depicts how competitive to the optimal solution the LBTs method is, towards to TPR*-tree and STRIPES, in case the velocity vector grows. The Ratio degrades as the query rectangle length grows from 400 to 1000.

Figure 7 depicts the performance of all methods in case the time interval length degrades to value 1. Even in this case, the LBTs method is more competitive than STRIPES and TPR*-tree. As query rectangle length grows from 400 to 1000, the LBTs method advantage decreases; we remark that STRIPES becomes faster, whereas LBTs method has exactly the same performance with the TPR*-trees.

Figure 8 depicts the efficiency of LBTs solution in comparison to that of TPR*-trees and STRIPES respectively in case the time interval length enlarges to 100.

6 Conclusions

We presented the problem of anonymity preserving data publishing in moving objects databases. In particular, we studied the case where the trajectory of a mobile user on the plane is no longer a polyline in a two-dimensional space, instead it is a two-dimensional surface. By transforming the surface's boundary poly-lines to dual points we experimentally focused on the impact of information distortion introduced by this space translation.

References

1. Abul, O., Bonchi, F., Nanni, M.: Never Walk Alone: Uncertainty for Anonymity in Moving Objects Databases. In: Proceedings 24th IEEE International Conference on Data Engineering, Cancun, pp. 376–385 (2008)
2. Beckmann, N., Kriegel, H.P., Schneider, R., Seeger, B.: The R*-tree: an Efficient and Robust Access Method for Points and Rectangles. In: Proceedings ACM International Conference on Management of Data (SIGMOD), Atlantic City, NJ, pp. 322–331 (1990)
3. Bajaj, R., Ranaweera, S., Agrawal, D.: GPS: location-tracking technology. Computer 35(4), 92–94 (2002)
4. Clarke, R.: Person location person tracking - technologies risks and policy implications. Information Technology and People 14(2), 206–231 (2001)

5. D'Roza, T., Bilchev, G.: An overview of location-based services. BT Technology Journal 21(1), 20–27 (2003)
6. Gkoulalas-Divanis, A., Verykios, V.S., Bozanis, P.: A network aware privacy model for online requests in trajectory data. Data Knowl. Eng. 68(4), 431–452 (2009)
7. Gaede, V., Gunther, O.: Multidimensional Access Methods. ACM Computing Surveys 30(2), 170–231 (1998)
8. Hoh, B., Gruteser, M., Xiong, H., Alrabady A.: Preserving privacy in gps traces via uncertainty-aware path cloaking. In: ACM Conference on Computer and Communications Security, pp. 161–171 (2007)
9. Jensen Christian, S., Lin, D., Ooi, B.C.: Query and Update Efficient B+-Tree Based Indexing of Moving Objects. In: VLDB 2004, pp. 768–779 (2004)
10. Kaporis, A., Makris, C., Sioutas, S., Tsakalidis, A., Tsichlas, K., Zaroliagis, K.: ISB-Tree: a New Indexing Scheme with Efficient Expected Behaviour. In: Deng, X., Du, D.-Z. (eds.) ISAAC 2005. LNCS, vol. 3827, pp. 318–327. Springer, Heidelberg (2005)
11. Kollios, G., Gunopulos, D., Tsotras, V.: On Indexing Mobile Objects. In: Proceedings 18th ACM Symposium on Principles of Database Systems (PODS), Philadelphia, PA, pp. 261–272 (1999)
12. Manolopoulos, Y., Theodoridis, Y., Tsotras, V.: Advanced Database Indexing. Kluwer Academic Publishers, Dordrecht (2000)
13. Papadopoulos, D., Kollios, G., Gunopulos, D., Tsotras, V.J.: Indexing Mobile Objects on the Plane. In: Hameurlain, A., Cicchetti, R., Traunmüller, R. (eds.) DEXA 2002. LNCS, vol. 2453, pp. 693–697. Springer, Heidelberg (2002)
14. Patel, J., Chen, Y., Chakka, V.: STRIPES: an Efficient Index for Predicted Trajectories. In: Proceedings ACM International Conference on Management of Data (SIGMOD), Paris, France, pp. 637–646 (2004)
15. Samarati, P.: Protecting respondents' identities in microdata release. IEEE Transactions on Knowledge and Data Engineering 13(6), 1010–1027 (2001)
16. Sweeney, L.: K-anonymity: a model for protecting privacy. International Journal on Uncertainty, Fuzziness and Knowledge-based Systems 10(5), 557–570 (2002)
17. Salzberg, B., Tsotras, V.J.: A Comparison of Access Methods for Time-Evolving Data. ACM Computing Surveys 31(2), 158–221 (1999)
18. Tao, Y., Papadias, D., Sun, J.: The TPR*-Tree: an Optimized Spatio-Temporal Access Method for Predictive Queries. In: Proceedings 29th International Conference on Very Large Data Bases (VLDB), Berlin, Germany, pp. 790–801 (2003)

PRAM Optimization Using an Evolutionary Algorithm

Jordi Marés and Vicenç Torra

IIIA - Institut d'Investigació en Intelligència Artificial
CSIC - Consejo Superior de Investigaciones Científicas
Campus de Bellaterra, 08193 Bellaterra, Catalonia, Spain
{jmares,vtorra}@iiia.csic.es

Abstract. PRAM (Post Randomization Method) was introduced in 1997 but it is still one of the least used methods in statistical categorical data protection. This fact is because of the difficulty to obtain a good transition matrix in order to obtain a good protection. In this paper, we describe how to obtain a better protection using an evolutionary algorithm with integrated information loss and disclosure risk measures to find the best matrix. We also provide experiments using a real dataset of 1000 records in order to empirically evaluate the application of this technique.

Keywords: Information Privacy and Security, Evolutionary Algorithms, Post Randomization Method, Information Loss, Disclosure Risk.

1 Introduction

As there are continuously more and more public dataset availables for analyses, more reliable protection methods are needed to ensure the privacy of the data. An obvious measure to maintain the privacy of the individuals is to replace or suppress any explicit identifier. However the application of this measure alone may be insufficient. Linking groups of records between different data sets might reveal the identity of individuals and involve an unauthorized disclosure of sensitive information [7].

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. The discussion for this can be found in [3].

In our case, we focus on categorical data which has more limited actions to perform when protecting because arithmetic operations are not allowed here. Then, the only actions that can be performed with categorical data are exchange of categories by others that already exist, and generalization of some categories into a newer ones, so having only two different actions for protection makes it a difficult task.

J. Domingo-Ferrer and E. Magkos (Eds.): PSD 2010, LNCS 6344, pp. 97–106, 2010.

The Post Randomization Method (PRAM) was introduced in [6] as a method for categorical variables disclosure control in microdata files. In [4] and [5], the method and some of its implications were discussed in more detail. However, the PRAM method is still one of the least used statistical categorical data protection methods because of the difficulty to obtain a good transition matrix in order to obtain a good protection. This was demonstrated in the experiments done in [8] where PRAM protections were the ones with the worse scores.

In this paper, we present a new way to find a good transition matrix which provides us with a good categorical microdata protection, using an evolutionary algorithm applied to an initial mask. The method is bootstrapped with the PRAM matrices decribed in [4,5,6,8].

The reminder of the paper is organized as follows. A brief explanation of the Post Randomization Method and a description of the types of matrices we have used is provided in Section 2, followed by an outline of evolutionary algorithms and a description of our proposed algorithm in Section 3. Experimental results are given in Section 4.

2 The Post Randomization Method (PRAM)

PRAM is a probabilistic, perturbative method for disclosure protection of categorical variables.

This method is based on changing the scores on some categorical variables for certain records to a different score according to a prescribed Markov matrix. This matrix contains a row for each possible value of each variable to be protected, and each row contains the probabilities of changing the original data value to any other value. These probability matrices are very important in order to obtain a good protection.

2.1 PRAM Matrices

There are different ways to define the Markov matrices in the literature. We discuss here two of the approaches, which are the most commonly used. In the discussion we understand p_{kl} as the probability of changing a value k to a value l. Then, $\sum_{l=1}^{n} p_{kl} = 1$, where n is the number of categories. We choosed two types of matrices design to work. The first type is a fully-filled matrix with the off-diagonal elements depending on the corresponding frequencies in the original microdata file. This approach has been used in [2]. Formally, the probability p_{kl} for $k \neq l$ is defined by

$$p_{kl} = \frac{(1 - p_{kk})(\sum_{i=1}^{n} T_\xi(i) - T_\xi(k) - T_\xi(l)}{(n-2)(\sum_{i=1}^{n} T_\xi(i) - T_\xi(k))} \tag{1}$$

where $T_\xi(i)$ is the frequency of the category i inside the original dataset for the actual variable. In the approach p_{kk} is left as constant, that is, $p_{kk} = p$ for all k. The key point of this equation is that it assigns the higher exchange probabilities

to the categories with less frequency. In this way, the resultant dataset has more confusion.

The second type is a fully-filled matrix with the diagonal elements depending on the corresponding frequencies in the original microdata file. This approach has been used in [8]. In this case the row values are determined by the following expressions:

$$p_{kk} = 1 - (\theta T_\xi(K)/T_\xi(k)) \tag{2}$$

for $k = 1, \ldots, n$ and, then,

$$p_{kl} = \frac{1 - p_{kk}}{n - 1} \tag{3}$$

for $k \neq l$, where $T_\xi(K)$ is the lower value frequency higher than zero, and θ is a parameter in $[0, 1]$. In our experiments we have used $\theta = 0.7$.

2.2 Analytical Measures

There exist two measures to evaluate the performance of a protection method: the information loss and the disclosure risk.

Information loss is known as the quantity of harm that is inflicted to the data by a given masking method. This measure is small when the analytic structure of the masked dataset is very similar to the structure of the original dataset, so, the motivation for preserving the structure of the dataset is to ensure that the masked dataset will be analytically valid and interesting.

Assessment of the quality of a protection method cannot be limited to information loss because disclosure risk has also to be measured. Disclosure risk is known as the quantity of original data that can be obtained by an intruder from the masked dataset. This measure is small when the masked dataset values are very different to the original values.

The problem here is that both measures are inversely related so the higher information loss the lower disclosure risk, and the inverse. In order to perform a good protection there must be a minimised combination of both measures.

3 Outline of Evolutionary Algorithms

Evolutionary algorithms are stochastic optimization and search methods that mimic the metaphor of natural biological evolution. Those algorithms operate on a population of potential solutions P so, formally, the sequence $P(0), P(1), \ldots, P(t)$ is called an *evolution* of $P(0)$.

The population is mantained over all the t generations, where every individual $X_i' \in P(t)$ is related to a potential solution for the given problem. In order to guide the individuals through the generations any "fitness" measure for *evaluation* is needed. The search for new potential solutions is performed selecting some of the individuals and *altering* them using operators such as mutation and

crossover. These operators generate an offspring of new individuals from previous generations. Surviving individuals are going evaluated again, and the process is repeated until some stopping criterion is reached.

Using this basic scheme there are some settings that can be adapted to the problem like how to represent and evaluate the individuals, the stopping criteria, how are the individuals selected and altered from generation to generation.

Alg. 1 shows a pseudo-code summarizing our algorithm which is a generic evolutionary algorithm with some particularities that are described below.

Algorithm 1. Evolutionary Algorithm to Enhance PRAM Matrices

```
Input: P(0) = X initial population
Output: P(t) = X' final population
t ⇐ 0
evaluate(P(0))
while stopping(P(t)) ≠ true; do
  alter ⇐randomly choose between mutation and cross
  if alter by mutation then
    X' ⇐ mutation(X)
  else
    X' ⇐ cross(X)
  end if
  evaluate(X, X')
  t ⇐ t + 1
end while
return P(t)
```

Next subsections describe the key points of our evolutionary algorithm such as individual representation, genetic operators and evaluation function. We will also discuss how information loss and disclosure risk are integrated within an evolutionary algorithm.

3.1 Genotype Encoding

Usually the initial matrix contain values with a lot of decimals so, in order to simplify the values, all the values are multiplied by 100 and only the integer part of the value will be kept for the encoding.

Encoding of the individual X is done value by value transforming them into its Gray code representation. The decision of working with Gray-coded values was taken to avoid abrupt value changes when any bit is altered. Discussion of gray coding can be found in [1].

A complete file encoding example is shown in Fig. 1. The example includes all the steps required during the whole encoding process.

3.2 Genetic Operators

Our proposed algorithm uses two basic operators: crossover and mutation.

0.185	0.283	0.532
0.609	0.089	0.302
0.002	0.277	0.721

PRAM matrix.

185	283	532
609	89	302
2	277	721

PRAM matrix in integers mode.

0010111001	0100011011	1000010100
1001100001	0001011001	0100101110
0000000010	0100010101	1011010001

Binary matrix.

0011100101	0110010110	1100011110
1101010001	0001110101	0110111001
0000000011	0110011111	1110111001

Gray-coded genome matrix.

Fig. 1. Example of genotype encoding

Crossover of the individual X is performed by swapping two ranges of values inside the individual as follows. Take two value positions $\{s, r\}$ at random, and consider that the two values at this position are $x_s \in X$ and $x_r \in X$. Generate a random number m to indicate the length of the ranges. This number must be in the range $[0, min(length(X) - s, length(X) - r, |s - r|)]$, where length($X$) is the total number of values inside the individual X, and —— is the absolute value operator. Then the ranges $[x_s, x_{s+m}]$ and $[x_r, x_{r+m}]$ are swapped obtaining a new individual. For example, having $s < r$ and $X = \{x_1, \ldots, x_n\}$ the new individual will be $X' = \{x_1, \ldots, x_r, \ldots, x_{r+m}, \ldots, x_s, \ldots, x_{s+m}, \ldots, x_n\}$.

Mutation is performed by a simple value mutation as follows. Take a random value of the individual X and consider that the value at this position is x_i with $genome(x_i) = b_j b_{j-1} \ldots b_1$. Choose a bit position k at random, such that $1 \leq k \leq j$. Then a new individual is obtained just by replacing the bit b_k by its negation counterpart, $b'_k = not(b_k)$.

We decided to use the value 0.5 for both crossover rate and mutation rate in order to have mostly the same number of operations performed by each one. A random value (alter) between 0 and 1 decides the operation to perform, using 0.5 as a delimiter.

3.3 Fitness Function

During the evaluation of the PRAM matrix two steps are needed. First of all the matrix has to be used for an implementation of the PRAM method to protect the original file, and then the protected file needs to be used into the evaluator software in order to obtain the results of information loss and disclosure risk related to the matrix.

As we have two measures to minimize, this is a multi-objective optimization problem. To solve this we chose a multi-objective optimization method called Objective Weighting which allows us to combine both measures applying an individual weight to each one. We wanted to give the same importance to both Disclosure Risk (DR) and Information Loss (IL) measures so both have $\frac{1}{2}$ as a weigth value. Then, the individual score can be obtained as follows:

$$Y = PRAM(X') \tag{4}$$

$$Score(X') = \frac{DR(Y) + IL(Y)}{2} \tag{5}$$

As the PRAM protection method takes some random decisions, the method generates different protected files with the same Markov matrix, and these different files will have different scores. In order to have a more robust score we compute 5 protected files for each individual (i.e., each Markov matrix) and we take the average of their scores as the final score. Formally:

$$FinalScore(X') = \frac{\sum_{i=1}^{5} Score(X')}{5} \tag{6}$$

In both mutation and crossover cases, during the evaluation, an elitism replacement strategy is followed which means that the new individual and the old one are compared and only the one with best score will be selected as the individual of the population for the next generation.

4 Experimental Results

To test and empirically evaluate our proposed method we have done several experiments protecting some attributes of a dataset and analising the evolution of the score in each one. The dataset we used is a U.S. Housing Survey of 1993 with 1000 records and 11 categorical attributes.

Here we are going to present, the results for the protection of the DEGREE attribute, which has 8 ordinal categories available, using the two types of PRAM matrices that have been described in section 2.1.

In these experiments, we are going to denote the matrix computed by equation (1) as $nF(p)$ where n is the size of the square matrix and p is the value for the elements in its diagonal, and the matrix computed by equations (2) and (3) as $nD(p)$ where n is the size of the square matrix and p is the value of the parameter θ. In our case, we have used matrices 8F(0.5) and 8D(0.7).

Figure 2 shows the evolution of the information loss, disclosure risk and score of the matrix $8D(0.7)$ over more than 1600 generations. It is easy to see that not all the measures have been decreased (indeed the information loss has increased!) but the adjust of the two measures has performed a very high decrease of the score. In this experiment, the score has decreased from 24.18 to 8.34 what represents almost the third part of the initial score. It can be also seen the effect of the evolutionary algorithm looking for the best adjustment of the measures increasing and decreasing them irregularly like the results around generation 200.

For the second experiment we have used matrix 8F(0.5) to protect the same attribute and we obtained the results shown in Figure 3. In this case, unlike the first experiment, all the measures have decreased forcing to decrease the final score too. This fact demonstrate again that the evolutionary algorithm does not

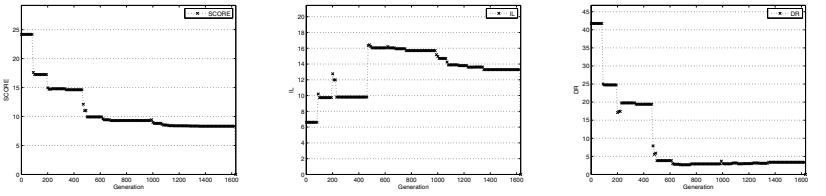


Fig. 2. Measures evolution for DEGREE attribute protection with $8D(0.7)$ matrix

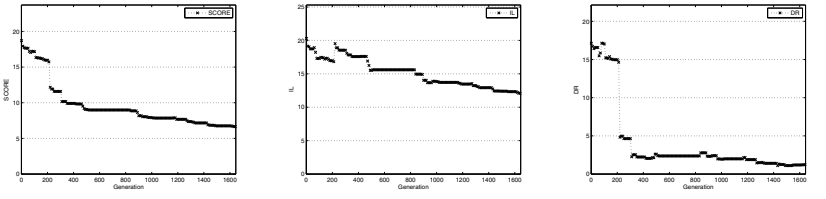


Fig. 3. Measures evolution for DEGREE attribute protection with $8F(0.5)$ matrix

follow any pattern in order to obtain the best result, it is just looking for the best option at the moment. In this experiment the fact of irregular variations of the measures also occurs, this can be observed in the figure within the first 200 generations. Using the $8F(0.5)$ matrix we obtained a decrement of the score from 18.71 to 6.65 what represents,like in the first experiment, alsmost the third part of the initial score.

A more detailed view of the results is given in Figure 4 where disaggregated measures are shown. The first one is the Interval Disclosure (ID) which decreases at the beginning and also from the 1600th generation, being mainly constant between those two decrements. The second measure is the Entropy-Based Information Loss (EBIL) which has a very similar behaviour. Finally the third measure is the Distance-Based Record Linkage (DBRL) which has a decrement at the beginning and after that is quite irregular but mantaining the range of values during all the generations. The rest of the measures that are not shown just mantain more or less their value.

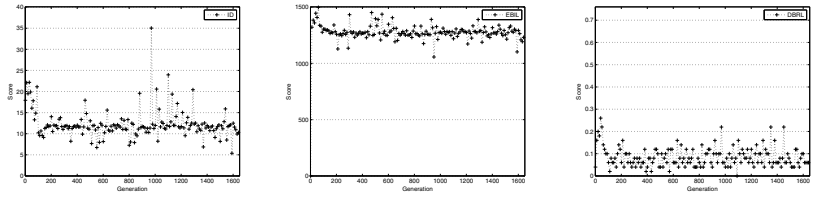


Fig. 4. Some disaggregated measures evolution

Table 1. Initial PRAM matrix 8D(0.7)

0.764	0.039	0.039	0.039	0.039	0.039	0.039	0.000
0.022	0.866	0.022	0.022	0.022	0.022	0.022	0.000
0.015	0.015	0.908	0.015	0.015	0.015	0.015	0.000
0.020	0.020	0.020	0.882	0.020	0.020	0.020	0.000
0.023	0.023	0.023	0.023	0.864	0.022	0.022	0.000
0.048	0.048	0.048	0.048	0.048	0.711	0.048	0.000
0.117	0.117	0.117	0.117	0.117	0.117	0.300	0.000
0.000	0.000	0.000	0.000	0.000	0.000	0.000	1.000

We also want to comment which changes have been performed between the original matrices and the final ones. Table 1 shows the initial 8D(0.7) matrix. It can be seen that the diagonal has the higher values of each row, but all values are different because they are computed depending on the frequency of each category. The rest of the elements in each row have the same value except the last one of each row which we decided to left it as 0.000 because it corresponds to the change to a reserved category.

The final matrix is shown in Table 2. Here we can see that the initial matrix structure has totally changed, so there is only one row with the highest value placed in the diagonal and not all the rows have the same off-diagonal values for all changes. An important point is that after the evolutionary algorithm is applied, there are three rows with their higher value at column 3 what means that almost all appearences of the categories corresponding to those rows will be changed to a single one (i.e., to the third category).

Table 2. Final PRAM matrix 8D(0.7)

0.352	0.026	0.005	0.005	0.173	0.429	0.010	0.000
0.026	0.026	0.026	0.026	0.846	0.026	0.026	0.000
0.057	0.057	0.057	0.739	0.011	0.023	0.057	0.000
0.094	0.774	0.019	0.019	0.019	0.057	0.019	0.000
0.039	0.016	0.921	0.003	0.003	0.014	0.005	0.000
0.006	0.006	0.814	0.006	0.047	0.019	0.102	0.000
0.469	0.000	0.055	0.055	0.055	0.227	0.139	0.000
0.062	0.062	0.430	0.000	0.000	0.000	0.033	0.413

The case of matrix $8F(0.5)$ is more or less the same. Table 3 shows the initial matrix. In this case it can be seen that all diagonal values have the same value, but the off-diagonal ones are all different in each row. This is because of the dependecy on the categories frequency when the matrix is computed. Note that, in this case, the higher values are still in the diagonal. For this matrix we wanted to use the reserved category in order to obtain a more different matrix, so the last column values are different than zero.

Finally, in Table 4 the final matrix is shown. Like in the 8D(0.7) matrix, this one also has the highest values outside of the diagonal. In addition, there are also two groups of two rows (i.e., categories) that are changed to different single categories for each group, obtaining a behaviour like in the case of 8D(0.7) matrix.

Table 3. Initial PRAM matrix 8F(0.5)

0.500	0.067	0.060	0.065	0.067	0.076	0.080	0.083
0.073	0.500	0.058	0.064	0.066	0.075	0.080	0.083
0.072	0.064	0.500	0.062	0.064	0.074	0.080	0.083
0.073	0.065	0.057	0.500	0.066	0.075	0.080	0.083
0.073	0.066	0.058	0.064	0.500	0.075	0.080	0.083
0.074	0.068	0.061	0.066	0.068	0.500	0.080	0.083
0.075	0.068	0.062	0.067	0.069	0.076	0.500	0.083
0.075	0.069	0.062	0.067	0.069	0.077	0.081	0.500

Table 4. Final PRAM matrix 8F(0.5)

0.011	0.009	0.011	0.009	0.926	0.009	0.011	0.014
0.000	0.004	0.971	0.005	0.005	0.005	0.004	0.004
0.027	0.625	0.009	0.027	0.205	0.036	0.036	0.036
0.049	0.037	0.432	0.074	0.062	0.259	0.049	0.037
0.032	0.005	0.006	0.928	0.006	0.008	0.006	0.008
0.892	0.008	0.033	0.008	0.011	0.017	0.017	0.014
0.003	0.030	0.004	0.004	0.466	0.487	0.003	0.003
0.430	0.416	0.005	0.006	0.019	0.006	0.006	0.111

More experiments have been done protecting other attributes and similar results have been obtained.

5 Conclusions

In this paper we have proposed an evolutionary algorithm to seek new and enhanced PRAM matrices in order to obtain better protections for categorical data. The experiments done in this paper have been presented using real survey data.

The $8D(0.7)$ PRAM matrix score got a 65.51% reduction -from 24.18 to 8.34-, and the $8F(0.5)$ PRAM matrix score got a 64.46% reduction -from 18.71 to 6.65- . These results demonstrate the effectiveness of our approach, and show that for some information loss measures the type of PRAM matrices found in this paper might be effective.

Our method has the advantage that can be extended to other measures of information loss and disclosure risk just by changing the fitness function. This property of decoupling of the algorithm from the measures is an important point because it may deserve future research.

On the contrary, the disadvantage is the cost in time for the evaluation of the information loss and disclosure risk. Aproximately, it takes 240 CPU seconds to compute both measures but, if we take in account that we need five computations per generation, a new individual complete evaluation takes 960 CPU seconds. As future work, this is a possible optimization to be explored.

Other lines of future work include the use of PRAM for protecting several variables at the same time and its comparison with other masking methods for categorical data.

In addition, experiments using other datasets with different sizes and structures will be also considered.

References

1. Caruana, R.A., Schaffer, J.D.: Representation and hidden bias: Gray vs binary coding for genetic algorithms. In: Proc. of the 5th Int. Conf. on Machine Learning, pp. 153–161. Morgan Kaufmann, Los Altos (1988)
2. De Wolf, P.P., Van Gelder, I.: An empirical evaluation of PRAM. Discussion paper 04012. Statistics Netherlands, Voorburg/Heerlen (2004)
3. Fienberg, S.E.: Conflict between the needs for access to statistical information and demands for confidentiality. Journal of Official Statistics 10(2), 115–132 (1994)
4. Gouweleeuw, J., Kooiman, P., Willenborg, L., de Wolf, P.P.: Post randomization for statistical disclosure control: Theory and implementation. Journal of Official Statistics 14(4), 463–478 (1998)
5. De Wolf, P.P., Gouweleeuw, J., Kooiman, P., Willenborg, L.: Reflections on pram. In: Statistical Data Protection, pp. 337–349. Office for Official Publications of the European Communities, Luxembourg (1998)
6. Kooiman, P., Willenborg, L., Gouweleeuw, J.: A method for disclosure limitation of microdata. Research paper 9705, Statistics Netherlands, Voorburg (1997)
7. Samarati, P.: Protecting respondents' identities in microdata release. IEEE Transactions on Knowledge and Data Engineering 13(6), 1010–1027 (2001)
8. Domingo-Ferrer, J., Torra, V.: A quantitative comparison of disclosure control methods for microdata. In: Doyle, P., Lane, J.I., Theeuwes, J.J.M., Zayatz, L.V. (eds.) Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies, ch. 6, pp. 111–133. Elsevier, Amsterdam (2001)

Multiplicative Noise Protocols

Anna Oganian

Georgia Southern University
aoganyan@georgiasouthern.edu

Abstract. Statistical agencies have conflicting obligations to protect confidential information provided by respondents to surveys or censuses and to make data available for research and planning activities. When the microdata themselves are to be released, in order to achieve these conflicting objectives, statistical agencies apply Statistical Disclosure Limitation (SDL) methods to the data, such as noise addition, swapping or microaggregation. In this paper, several multiplicative noise masking schemes are presented. These schemes are designed to preserve positivity and inequality constraints in the data together with means and covariance matrix.

Keywords and phrases: Statistical disclosure limitation (SDL), SDL method, multiplicative noise, positivity and inequality constraints.

1 Introduction

Official statistical agencies have long been aware of the tension between preserving confidentiality of data—identities of data subjects and values of sensitive attributes—and releasing useful information for policy, research or other purposes.

Multiple means of access to microdata records exist, including restricted data centers (*e.g.*, [ANES, MEPS, SSDS]), licensing [NCES1] and remote access servers [GKRP05]. These are effective, but they do not meet all needs, and many agencies also release deliberately altered microdata publicly.

For public microdata releases, the role of statistical disclosure limitation (SDL) is to alter the data in a way that maintains the utility but limits disclosure risk.

Many SDL methods can be used to prepare microdata releases. Of course, the initial step is to remove explicit identifiers for individuals — name, address and social security number.

Almost always, removal of identifiers alone is inadequate. Rare attribute combinations (for example, a 17-year old widow) can lead to re-identification. Moreover, in high-dimensional data, virtually every subject may have a unique set of attributes. Therefore, almost invariably, released data attributes must be modified. Some SDL techniques coarsen the resolution of the data; for example, date of birth can be replaced by age, and age may be reported in five-year intervals. Extreme attribute values can be top- or bottom-coded.

J. Domingo-Ferrer and E. Magkos (Eds.): PSD 2010, LNCS 6344, pp. 107–117, 2010.

Another approach is to generate synthetic records, which are draws from a distribution (typically, a posterior predictive distribution) representing original data.

Other methods actually change attribute values. Examples are addition of noise, data swapping and microaggregation [KKORS06, PSD06]. We term methods whose output is a perturbed version of the original data *perturbation methods*. This paper focuses on one of these — perturbation by means of externally generated "noise." Each specific perturbation method has consequences on both disclosure risk and data utility. Some limit risk effectively but are poor at preserving utility, while others yield high utility, but at the price of high risk. No method is superior with respect to both. It is possible, indeed, to combine two methods with the goal of capturing the good aspects of each [PSD06].

From a data utility perspective, it is important to preserve qualitative characteristics of data, for example, positivity constraints of the form $X \geq 0$ for some variables and inter-attribute relationships such as linear inequalities. Age, many economic variables (gross income, taxes) and many demographic variables (number of employees, number of students in the sixth grade) obey positivity constraints; examples of inequality constraints are "Federal taxes $\leq$ gross income", "number of salaried employees $\leq$ number of employees" and "year of birth $\leq$ year of death."

There is also a risk aspect. Because such characteristics are derived from domain knowledge available to both legitimate data users and intruders, failure to preserve them poses a disclosure risk: the extent to which constraints are violated can be informative about the nature and intensity of the SDL applied to the data.

Some SDL methods preserve structural characteristics more by coincidence than by design, and only partially. For instance, data swapping preserves positivity, but not multi-attribute constraints. Microaggregation preserves positivity, but whether it preserves linear inequalities depends on specifics of the implementation.

In this paper, we present several SDL protocols applicable to the numerical data that *preserves positivity constraints, inequality constraints and the first two moments*—the mean and covariance matrix.

For the purposes of this paper, the original and released (which we hereafter term masked) databases are flat files in which rows represent data subjects (individuals, households, business establishments, ...) and columns numerical attributes of those subjects. We denote the original data by $\boldsymbol{X_o}$ and the released (masked) data by $\boldsymbol{X_m}$. We assume that some variables in $\boldsymbol{X_o}$ are nonnegative, others can take positive and negative values. The goal is to obtain $\boldsymbol{X_m}(j) \geq 0$ for those variables j such that $\boldsymbol{X_o}(j) \geq 0$ and $\boldsymbol{X_m}$ should have the same mean and covariance matrix as $\boldsymbol{X_o}$.

As background, the analogous procedure for addition of noise to unconstrained numerical data is as follows. Let $\boldsymbol{\Sigma_o}$ be the covariance matrix of $\boldsymbol{X_o}$—in practice, one can use either the usual empirical estimator or a shrinkage-based estimator.

Let $k > 0$ be a parameter chosen by the agency; then

$$\boldsymbol{X_m} = E[\boldsymbol{X_o}] + \frac{(\boldsymbol{X_o} - E[\boldsymbol{X_o}]) + \boldsymbol{E}}{\sqrt{1+k}}, \tag{1}$$

where the noise $\boldsymbol{E}$ has distribution $N(\boldsymbol{0}, k\boldsymbol{\Sigma_o})$, has the requisite properties [PSD06]. Note that the value of k need not be released, even if it were made known that the method of SDL is addition of noise. As $k \to \infty$, $\boldsymbol{X_m}$ becomes a very simplistic form of synthetic data [Reit02], and any non-normal distributional characteristics of $\boldsymbol{X_o}$ are lost.

The structure of this paper is the following: several multivariate noise protocols are presented in §2, the extension of these protocols to satisfy inequality constraints is described in §3 and the results of the numerical experiments are reported in §4.

2 Multiplicative Noise Protocols

Suppose that $\boldsymbol{X_o}$ contain n records, each with d numerical attributes. Some of the attributes are nonnegative, denote them $\boldsymbol{X_o}^p$. We wish to construct and release a masked data set $\boldsymbol{X_m}$ with these characteristics:

$$\boldsymbol{X_m}^p \geq 0 \tag{2}$$
$$E[\boldsymbol{X_m}] = E[\boldsymbol{X_o}] \tag{3}$$
$$\Sigma(\boldsymbol{X_m}) = \Sigma(\boldsymbol{X_o}), \tag{4}$$

where $\boldsymbol{X_m}^p$ are the masked values of $\boldsymbol{X_o}^p$ and $\Sigma(\cdot)$ means "covariance matrix of $(\cdot)$."

In [OganKarr10], a masking scheme which preserves positivity, means and covariance matrix data was proposed. The basis of this scheme is to use multiplicative noise, implemented by taking logarithms, applying additive, normally distributed noise and exponentiating. This scheme works only if all the variables in the data set are nonnegative. Below are the details.

Let $\boldsymbol{E}$ be noise that is conditionally independent of $\boldsymbol{X_o}$ given $E[\boldsymbol{X_o}]$ and $\Sigma(\boldsymbol{X_o})$, and satisfies

$$E[\boldsymbol{X_o} \circ \exp(\boldsymbol{E})] = E[\boldsymbol{X_o}] \tag{5}$$
$$\Sigma(\boldsymbol{X_o} \circ \exp(\boldsymbol{E})) = (1+k)\Sigma(\boldsymbol{X_o}), \tag{6}$$

where $k > 0$ is an agency-chosen parameter and $\circ$ denotes elementwise matrix multiplication (Schur or Hadamard product). That is, the exponentiation in (5), (6) and elsewhere below also takes place componentwise. Then

$$\boldsymbol{X_m} = \frac{(\sqrt{1+k}-1)E[\boldsymbol{X_o}] + [\boldsymbol{X_o} \circ \exp(\boldsymbol{E})]}{\sqrt{1+k}} \tag{7}$$

satisfies (2)–(4).

For normally distributed noise $\boldsymbol{E}$, [OganKarr10] showed that the following vector of means $\boldsymbol{\mu_E}$ and covariance matrix $\boldsymbol{\Sigma_E}$ should be chosen for $\boldsymbol{E}$ to satisfy (5) and (6):

$$\boldsymbol{\Sigma_E}(i,j) = \log\left(1 + \frac{k\boldsymbol{\Sigma_o}(i,j)}{E[\boldsymbol{X_o}(i)\boldsymbol{X_o}(j)]}\right), \qquad i,j = 1,\ldots,d \tag{8}$$

$$\boldsymbol{\mu_E}(i) = -\boldsymbol{\sigma_E}(i)/2, \qquad i = 1,\ldots,d. \tag{9}$$

Here, d is the number of dimensions in the data.

If the data set contains not only nonnegative variables but variables with negative values as well, the scheme described above cannot be apply directly. The variables with negative and positive values may lead to

$$1 + \frac{k\boldsymbol{\Sigma_o}(i,j)}{E[\boldsymbol{X_o}(i)\boldsymbol{X_o}(j)]} < 0 \tag{10}$$

so, the covariance matrix (8) cannot be computed. After extensive experiments, it was also noticed that for some very rare distributions of values in $\boldsymbol{X_o}^p$, (10) may still hold. Example of such distribution is shown in Figure 1. The variables here are negatively correlated and aligned along the axes.

So, for the implementation of the multiplicative noise masking strategy, it would be helpful to have a scheme applicable to all the data sets. One possible solution is to convert all the variables to z-scores and make these z-scores nonnegative by adding some value (or vector – for multivariate data), denote it $\boldsymbol{lag}$, such that $\boldsymbol{lag} \geq |min(\boldsymbol{Z})|$. Denote these nonnegative z-scores by $\boldsymbol{Z_{pos}}$. Then we can apply the masking scheme described by (7), (8) and (9) to $\boldsymbol{Z_{pos}}$ and after that return the resulting data to the original scale:

$$\boldsymbol{Z_m} = \frac{(\sqrt{1+k}-1)\boldsymbol{lag} + [\boldsymbol{Z}^{pos} \circ \exp(\boldsymbol{E}^{\boldsymbol{z_{pos}}})]}{\sqrt{1+k}} \tag{11}$$

$$\boldsymbol{X_m} = (\boldsymbol{Z_m} - \boldsymbol{lag}) \circ \boldsymbol{\sigma_o} + E(\boldsymbol{X_o}) \tag{12}$$

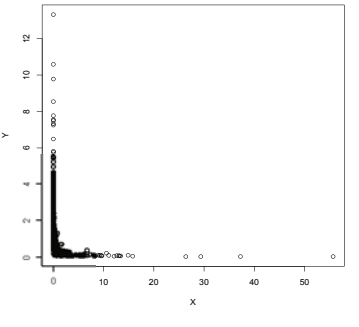


Fig. 1. Example of a data set when covariance matrix for noise cannot be computed

where $\boldsymbol{\sigma_o}$ is the main diagonal of $\boldsymbol{\Sigma_o}$ and $\boldsymbol{E^{z_{pos}}}$ has the following mean and covariance matrix:

$$\boldsymbol{\Sigma}_{\boldsymbol{E}_{z_{pos}}}(i,j) = \log\left(1 + \frac{k\boldsymbol{\Sigma}_{\boldsymbol{z}_{pos}}(i,j)}{E[\boldsymbol{Z_{pos}}(i)\boldsymbol{Z_{pos}}(j)]}\right), \qquad i,j = 1,\ldots,d \tag{13}$$

$$\boldsymbol{\mu}_{\boldsymbol{E}_{z_{pos}}}(i) = -\boldsymbol{\sigma}_{\boldsymbol{E}_{z_{pos}}}(i)/2, \qquad i = 1,\ldots,d. \tag{14}$$

where $\boldsymbol{\Sigma}_{\boldsymbol{z}_{pos}}(i,j)$ is the (i,j) element of the covariance matrix of positive z-scores.

Masked data $\boldsymbol{X_m}$ in this case can be represented as

$$\begin{aligned}\boldsymbol{X_m} &= \Big(\frac{(\boldsymbol{Z_{pos}} \circ \exp(\boldsymbol{E^{z_{pos}}}) + (\sqrt{1+k}-1)\boldsymbol{lag}}{\sqrt{1+k}} - \boldsymbol{lag}\Big) \circ \boldsymbol{\sigma_o} + E(\boldsymbol{X_o}) = \\ &= \frac{[\boldsymbol{X_o} \circ \exp(\boldsymbol{E^{z_{pos}}})] - E(\boldsymbol{X_o}) \circ exp(\boldsymbol{E^{z_{pos}}}) + \boldsymbol{\sigma_o} \circ \boldsymbol{lag} \circ exp(\boldsymbol{E^{z_{pos}}})}{\sqrt{1+k}} + \\ &+ \frac{E(\boldsymbol{X_o})\sqrt{1+k} - \boldsymbol{lag} \circ \boldsymbol{\sigma_o}}{\sqrt{1+k}}\end{aligned} \tag{15}$$

It is easy to see that such scheme preserves means and covariance matrix:

$$\begin{aligned}E(\boldsymbol{X_m}) &= \frac{1}{\sqrt{k+1}}[E(\boldsymbol{X_o}) - E(\boldsymbol{X_o}) + \boldsymbol{\sigma_o} \circ \boldsymbol{lag} - \boldsymbol{\sigma_o} \circ \boldsymbol{lag} + \\ &+ E(\boldsymbol{X_o})\sqrt{1+k}] = E(\boldsymbol{X_o})\end{aligned} \tag{16}$$

The equality in the formula above follows from the fact that noise is independent from $\boldsymbol{X_o}$ and $E(exp(\boldsymbol{E^{z_{pos}}})) = 1$.

$$\begin{aligned}&\boldsymbol{\Sigma_m}(i,j) = \boldsymbol{\Sigma}\Big(\frac{\boldsymbol{Z_{pos}}(i)\exp(\boldsymbol{E^{z_{pos}}}(i))\boldsymbol{\sigma_o}(i)}{\sqrt{1+k}}, \frac{\boldsymbol{Z_{pos}}(j)\exp(\boldsymbol{E^{z_{pos}}}(j))\boldsymbol{\sigma_o}(j)}{\sqrt{1+k}}\Big) = \\ &= \frac{\boldsymbol{\sigma_o}(i)\boldsymbol{\sigma_o}(j)}{1+k}(1+k)cov(\boldsymbol{Z_{pos}}(i), \boldsymbol{Z_{pos}}(j)) = \\ &= \boldsymbol{\sigma_o}(i)\boldsymbol{\sigma_o}(j)cor(\boldsymbol{X_o}(i), \boldsymbol{X_o}(j)) = \boldsymbol{\Sigma_o}(i,j)\end{aligned} \tag{17}$$

where $cov(\cdot)$ and $cor(\cdot)$ denote covariance and correlation of $(\cdot)$ respectively. Note that the second equality in the formula above follows from the property (6).

Now we will show that masking scheme (15) with the specific choice for $\boldsymbol{lag}$ will never lead to the case described by (10).

First, let us see what are the possible values for $\boldsymbol{lag}$ in this scheme. $\boldsymbol{lag}$ should be greater than $|min(\boldsymbol{Z})|$, however, a very big $\boldsymbol{lag}$ may lead to negative masked data (this follows from equation(12)), which violates positivity constrants for the variables ${\boldsymbol{X_o}}^p$.

From (11), $\boldsymbol{Z_m}$ is minimized when $E_n \to -\infty$:

$$min(\boldsymbol{Z_m}) > \frac{(\sqrt{1+k}-1)\boldsymbol{lag}}{\sqrt{1+k}}$$

From (12), $min(\boldsymbol{X_m})$ is larger than

$$\frac{-\boldsymbol{lag}}{\sqrt{1+k}}\boldsymbol{\sigma_o} + E(\boldsymbol{X_o}) \tag{18}$$

To preserve positivity in the masked data it would be enough to require positivity of (18). So, we have an upper bound for $\boldsymbol{lag}$:

$$\boldsymbol{lag} \leq \frac{E(\boldsymbol{X_o})}{\boldsymbol{\sigma_o}}\sqrt{1+k}$$

where division is done componentwise.

The lower bound for $\boldsymbol{lag}$ is $|min(\boldsymbol{Z})|$. For nonnegative variables with zeros $|min(\boldsymbol{Z})| = E(\boldsymbol{X_o})/\boldsymbol{\sigma_o}$. So, we can write the lower and upper bound for $\boldsymbol{lag}$ as:

$$\frac{E(\boldsymbol{X_o})}{\boldsymbol{\sigma_o}} \leq \boldsymbol{lag} \leq \frac{E(\boldsymbol{X_o})}{\boldsymbol{\sigma_o}}\sqrt{1+k} \tag{19}$$

Let us consider a few choices for $\boldsymbol{lag}$ in this range. If we choose $\boldsymbol{lag} = E(\boldsymbol{X_o})/\boldsymbol{\sigma_o}$, then the scheme with z-scores transformation (15) is equivalent to the scheme without transformation (7). In fact, it is straightforward to verify that masked data in this case can be written as:

$$\boldsymbol{X_m} = \frac{(\sqrt{1+k}-1)E[\boldsymbol{X_o}] + [\boldsymbol{X_o} \circ \exp(\boldsymbol{E^{z_{pos}}})]}{\sqrt{1+k}} \tag{20}$$

Expression (20) is almost identical to (7) except the second term in the nominator: $[\boldsymbol{X_o} \circ \exp(\boldsymbol{E^{z_{pos}}})]$.

Below we will show that even this term is identical in both schemes. In particular, after application of our masking scheme to the positive z-scores, noise $\boldsymbol{E^{z_{pos}}}$ has the mean and covariance matrix defined by (14) and (13) respectively.

Note, that

$$\begin{aligned}
&\frac{\boldsymbol{\Sigma_{z_{pos}}}(i,j)}{E[\boldsymbol{Z_{pos}}(i)\boldsymbol{Z_{pos}}(j)]} = \\
&= \frac{cor(\boldsymbol{X_o}(i), \boldsymbol{X_o}(j))}{E[(\frac{(\boldsymbol{X_o}(i)-E(\boldsymbol{X_o}(i)))}{\boldsymbol{\sigma_o}(i)} + \boldsymbol{lag}(i))(\frac{(\boldsymbol{X_o}(j)-E(\boldsymbol{X_o}(j)))}{\boldsymbol{\sigma_o}(j)} + \boldsymbol{lag}(j))]} = \\
&= \frac{cor(\boldsymbol{X_o}(i), \boldsymbol{X_o}(j))}{E[\boldsymbol{X_o}(i)/\boldsymbol{\sigma_o}(i) * \boldsymbol{X_o}(j)/\boldsymbol{\sigma_o}(j)]} = \frac{\boldsymbol{\Sigma_o}(i,j)}{E[\boldsymbol{X_o}(i)\boldsymbol{X_o}(j)]}
\end{aligned}$$

So, when $\boldsymbol{lag} = E(\boldsymbol{X_o})/\boldsymbol{\sigma_o}$, transformation to positive $z-$scores does not make any changes in the original scheme (7).

Now let us consider another extreme for $\boldsymbol{lag}$: $\boldsymbol{lag} = \sqrt{(}1+k)E(\boldsymbol{X_o})/\boldsymbol{\sigma_o}$.

It is easy to verify that masked data in this case can be written as:

$$\boldsymbol{X_m} = \frac{(\sqrt{1+k}-1)E[\boldsymbol{X_o}] \circ \exp(\boldsymbol{E^{z_{pos}}}) + [\boldsymbol{X_o} \circ \exp(\boldsymbol{E^{z_{pos}}})]}{\sqrt{1+k}} \tag{21}$$

Covariance matrix for noise for this scheme is:

$$\boldsymbol{\Sigma}_{\boldsymbol{E}_{\boldsymbol{z}_{\boldsymbol{pos}}}}(i,j) = \log\left(1 + \frac{k\boldsymbol{\Sigma}_{\boldsymbol{z}_{\boldsymbol{pos}}}(i,j)}{E[\boldsymbol{Z}_{\boldsymbol{pos}}(i)\boldsymbol{Z}_{\boldsymbol{pos}}(j)]}\right) \tag{22}$$

To prove that the expression under logarithm of (22) is always positive, let's express it in terms of original data.

$$\boldsymbol{Z}_{\boldsymbol{pos}}(i) = \frac{\boldsymbol{X}_{\boldsymbol{o}}(i) + E(\boldsymbol{X}_{\boldsymbol{o}}(i))(\sqrt{1+k}-1))}{\boldsymbol{\sigma}_{\boldsymbol{o}}(i)}$$

It is easy to see that

$$E[\boldsymbol{Z}_{\boldsymbol{pos}}(i)\boldsymbol{Z}_{\boldsymbol{pos}}(j)] = \frac{E[\boldsymbol{X}_{\boldsymbol{o}}(i)\boldsymbol{X}_{\boldsymbol{o}}(j)] + kE(\boldsymbol{X}_{\boldsymbol{o}}(i))E(\boldsymbol{X}_{\boldsymbol{o}}(j))}{\boldsymbol{\sigma}_{\boldsymbol{o}}(i)\boldsymbol{\sigma}_{\boldsymbol{o}}(j)}$$

$$\begin{aligned}\boldsymbol{\Sigma}_{\boldsymbol{E}_{\boldsymbol{z}_{\boldsymbol{pos}}}}(i,j) &= \log\left(1 + \frac{k\boldsymbol{\sigma}_{\boldsymbol{o}}(i)\boldsymbol{\sigma}_{\boldsymbol{o}}(j)cor(\boldsymbol{X}_{\boldsymbol{o}}(i),\boldsymbol{X}_{\boldsymbol{o}}(j))}{E[\boldsymbol{X}_{\boldsymbol{o}}(i)\boldsymbol{X}_{\boldsymbol{o}}(j)] + kE(\boldsymbol{X}_{\boldsymbol{o}}(i))E(\boldsymbol{X}_{\boldsymbol{o}}(j))}\right) = \\ &= \log\left(\frac{(1+k)E[\boldsymbol{X}_{\boldsymbol{o}}(i)\boldsymbol{X}_{\boldsymbol{o}}(j)]}{E[\boldsymbol{X}_{\boldsymbol{o}}(i)\boldsymbol{X}_{\boldsymbol{o}}(j)] + kE(\boldsymbol{X}_{\boldsymbol{o}}(i))E(\boldsymbol{X}_{\boldsymbol{o}}(j))}\right)\end{aligned} \tag{23}$$

The expression under logarithm in (23) is always positive for nonnegative $\boldsymbol{X}_{\boldsymbol{o}}$, so we can always compute $\boldsymbol{\Sigma}_{\boldsymbol{E}_{\boldsymbol{z}_{\boldsymbol{pos}}}}$. In the same way, it is possible to show that no other value for $\boldsymbol{lag}$ (in the range of its possible values) can guarantee positivity of (10) for all possible data sets.

When the data set contains variables which can take positive and negative values together with nonnegative variables, the scheme with z-scores transformations will work too. First the data should be made nonnegative by adding $|min(\boldsymbol{X}_{\boldsymbol{o}})|$ and then scheme (21) is applied to this data. Last, to return the data to the original location, we have to substract $|min(\boldsymbol{X}_{\boldsymbol{o}})|$ from the result of the previous step.

3 Preservation of Inequality Constraints

The scheme described above can be easily extended to satisfy inequality constraints of the form $X > Y$. For example, masking an income data with the variables "Gross income" and "Federal taxes" should produce a masked data such that "Gross income > Federal taxes".

Masking scheme. *To preserve inequality constraints* $\boldsymbol{X} > \boldsymbol{Y}$*:*

- *Apply multiplicative noise scheme to* $(\boldsymbol{Y}_{\boldsymbol{o}}, [\boldsymbol{X}_{\boldsymbol{o}} - \boldsymbol{Y}_{\boldsymbol{o}}])$*. Denote the result by* $(\boldsymbol{Y}^*, [\boldsymbol{X}_{\boldsymbol{o}} - \boldsymbol{Y}_{\boldsymbol{o}}]^*)$
- *Masked data corresponding to* $(\boldsymbol{X}_{\boldsymbol{o}}, \boldsymbol{Y}_{\boldsymbol{o}})$ *is* $(\boldsymbol{X}_{\boldsymbol{m}}, \boldsymbol{Y}_{\boldsymbol{m}}) = (\boldsymbol{Y}^* + [\boldsymbol{X}_{\boldsymbol{o}} - \boldsymbol{Y}_{\boldsymbol{o}}]^*, \boldsymbol{Y}^*)$

It is easy to see that this scheme preserves means and covariance matrix.

$$\begin{aligned}
E(\boldsymbol{X_m}, \boldsymbol{Y_m}) &= E(\boldsymbol{Y}^* + [\boldsymbol{X_o} - \boldsymbol{Y_o}]^*, \boldsymbol{Y}^*) = \\
&= (E(\boldsymbol{Y_o}) + E[\boldsymbol{X_o} - \boldsymbol{Y_o}]), (E(\boldsymbol{Y_o})) = (E(\boldsymbol{X_o}), E(\boldsymbol{Y_o})) \\
cov(\boldsymbol{X_m}, \boldsymbol{Y_m}) &= cov(\boldsymbol{Y}^* + [\boldsymbol{X_o} - \boldsymbol{Y_o}]^*, \boldsymbol{Y}^*) = var(\boldsymbol{Y_o}) + \\
&+ cov([\boldsymbol{X_o} - \boldsymbol{Y_o}]^*, \boldsymbol{Y}^*) = var(\boldsymbol{Y_o}) + cov([\boldsymbol{X_o} - \boldsymbol{Y_o}], \boldsymbol{Y_o}) = \\
&= var(\boldsymbol{Y_o}) + cov(\boldsymbol{X_o}, \boldsymbol{Y_o}) - var(\boldsymbol{Y_o}) = cov(\boldsymbol{X_o}, \boldsymbol{Y_o}) \\
var(\boldsymbol{X_m}) &= var((\boldsymbol{Y}^* + [\boldsymbol{X_o} - \boldsymbol{Y_o}]^*) = var(\boldsymbol{Y_o}) + var([\boldsymbol{X_o} - \boldsymbol{Y_o}]) + \\
&+ 2cov(\boldsymbol{Y_o}, [\boldsymbol{X_o} - \boldsymbol{Y_o}]) = var(\boldsymbol{X_o})
\end{aligned}$$

The scheme can be readily extended for the cases when multiple variables are related by inequality constraints.

Example. *Suppose we have d variables in a data set and these variables have following relationships:*

$$\boldsymbol{X}_{o1} > \boldsymbol{X}_{o2} > \boldsymbol{X}_{o3} \;\; \boldsymbol{X}_{o4} > \boldsymbol{X}_{o5} \;\; \boldsymbol{X}_{o6} \cdots \boldsymbol{X}_{od}$$

Masking scheme will be the following:

$$\begin{aligned}
\boldsymbol{X}_{m1} &= \boldsymbol{X}_3^* + [\boldsymbol{X}_{o2} - \boldsymbol{X}_{o3}]^* + [\boldsymbol{X}_{o1} - \boldsymbol{X}_{o2}]^* \\
\boldsymbol{X}_{m2} &= \boldsymbol{X}_3^* + [\boldsymbol{X}_{o2} - \boldsymbol{X}_{o3}]^* \\
\boldsymbol{X}_{m3} &= \boldsymbol{X}_3^* \\
\boldsymbol{X}_{m4} &= \boldsymbol{X}_5^* + [\boldsymbol{X}_{o4} - \boldsymbol{X}_{o5}]^* \\
\boldsymbol{X}_{m5} &= \boldsymbol{X}_5^* \\
\boldsymbol{X}_{m6} &= \boldsymbol{X}_6^* \\
&\vdots \\
\boldsymbol{X}_{md} &= \boldsymbol{X}_d^*
\end{aligned}$$

4 Numerical Experiments

Both multiplicative noise schemes (with and without z-scores transformation) were implemented and evaluated on different data sets. These data sets have different distributional characteristics: skewed distribution with many outliers and a symmetrical one without outliers. Symmetrical data sets had multivariate normal distribution and skewed sets were log-normally distributed. 500 replicates of three-dimensional normal and lognormal sets were generated. Each set had $10,000$ records. They were moderately correlated ($cor = 0.5$). Log-normal sets had means around 2 and variances ranging from 4 to 16. These sets had outliers—values close to 50 or larger.

Normal sets had means around 3.5 and variances ranging from 5 to 10. The variance inflation factor k was chosen to be 0.15 as recommended in [Ogan03].

The experiments showed that means were very well preserved for both schemes and both types of data: the ratio of masked and original means showed only a very small variation around 1. Results on variance/covariance matrix were different for skewed and symmetrical data sets. The experiments showed that covariance matrix was preserved for symmetrical data sets without outliers. There was slight variability in variance/covariance matrix inflation, defined as $\boldsymbol{\Sigma_m}/\boldsymbol{\Sigma_o}$, where / denotes elementwise division. Values of this ratio ranged from 0.98 to 1.02.

There was more variability in variance/covariance matrix inflation for skewed data sets with outliers. Values of this ratio ranged approximately from 0.7 to 1.3. Scheme with z-transformation resulted to be slightly more stable: variance/covariance inflation ranged approximately from 0.8 to 1.2. However, the average and most frequent value were 1 in both schemes and both types of data sets, as expected.

Such variability over replications is not very surprising in light of the nature of the noise and the variation in log-normal original data, which as noted above had a number of large outlying values. Records in the original data with big values—especially outliers—can undergo significant changes when multiplied by noise, distorting the covariance matrix.

One possible solution to reduce variability in the resulting masked data when the original is skewed and/or has many outliers is to apply different levels of noise to different zones of the data, as discussed in [OganKarr10]. It is illustrated in the Figure 2, where zone 1 is masked with the parameter k_1 and zone 2 with the parameter $k_2 < k_1$. Because all the protocols presented in §2 are designed to preserve the mean and covariance matrix of the original data, we can apply different independent noises to different zones of the data and the covariance matrix of the masked data should be the same as that of the original data.

Two-zone masking was implemented with different values of k for the same three-dimensional lognormal data as in the experiment with only one zone. The first zone consisted of all the points from 0 to 15; all the other records were included in the second zone. For the second zone we chose $k_2 = 0.01$. For the first zone we chose $k_1 = 0.15$.

This approach reduced variability in the covariance matrix of the skewed data significantly: in 95% of replicates of the masked data $\boldsymbol{X_m}/\boldsymbol{X_o}$ was in the interval of $[0.98, 1.02]$.

The best ways of variability reduction in masked data when the original have outliers and severe skewness is the subject of our current and future research.

Note, that multiple-zone masking may be used for other goals. For example, suppose a numerical variable in the data set has a lot of zeros, which happens often in the household data. Suppose the same numerical variable is paired with an indicator variable I, such that when $I = 0$, it is strictly positive and when $I = 1$, it is zero. Examples of I are "In the labor force" or "Income is greater than taxable min". If the agency wants to preserve such a relationship in the masked data, they can separately mask records paired with different values of an indicator variable leaving zeros in the numerical variable unchanged. Again,

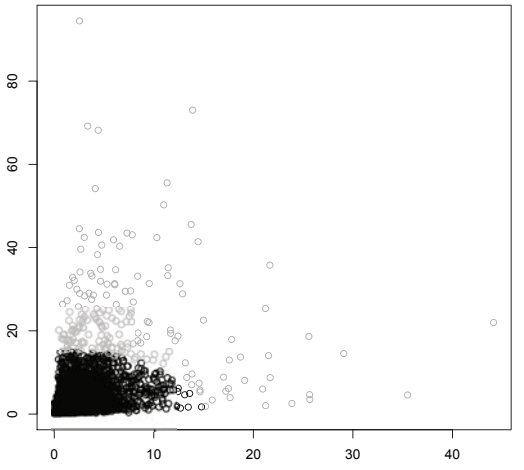


Fig. 2. Two zones of masking: black points correspond to the first zone of masking and grey points to the second zone

because our protocols preserve means and covariance matrix, the first two moments of the overall data should be preserved.

So, protocols described above are designed to preserve the first two moments and positivity constraints, but preservation of higher order moments is not guaranteed. However, because the masked data are scaled to have the same covariance matrix as the original data, higher-order moments seem unlikely to be grossly inflated, but it is possible. In most of our experiments with different data sets, third-order moments were only 3% larger than the original moments on average for skewed original data with outliers. For symmetrical data sets with the same covariance matrix as the lognormal ones, they were only .2% larger than the corresponding original ones. Fourth-order moments were about 15% larger on average for lognormal original data and only .9% larger for symmetrical data. In general, the discrepancy increases with the order of the moment, but only slowly. For the scheme with z-scores transformation the inflation of higher order moments was slightly smaller than for the scheme without transformation.

Last, we want to discuss the disclosure risk associated with the method. Our measures of disclosure risk focus on re-identification disclosure risk. Re-identification disclosure is defined as an average percentage of correctly identified records when record linkage techniques [J89] are used to match the original and masked data. Specifically, we assume that the intruder tries to link the masked file with an external database containing a subset of the attributes present in the original data [Ogan03]. The overall re-identification risk of the multiplicative noise is very small. Our experiments showed that only about 0.3% of records could be correctly identified in both schemes. So, multiplicative noise can be successfully compared with the most protective methods, like microaggregation and

rank swapping, at the same time performing significantly better than those in terms of utility.

Acknowledgments

This research was partly funded by NSF grant EIA–0131884 to the National Institute of Statistical Sciences (NISS). The sincerest thanks go to Alan Karr.

Any opinions, findings, and conclusions or recommendations expressed in this publication are those of the author and do not necessarily reflect the views of the National Science Foundation.

References

[AW01] Abowd, J.M., Woodcock, S.: Disclosure Limitation in Longitudinal Linked Data. In: Doyle, P., Lane, J., Zayatz, L., Theeuwes, J. (eds.) Confidentiality, Disclosure, and Data Access: Theory and Practial Applications for Statistical Agencies, pp. 215–277. North Holland, Amsterdam (2001)

[ANES] American National Election Studies Resticted Data Access, `http://www.electionstudies.org/rda/anes_rda.htm`

[SIPP] SIPP synthetic beta, Census Bureau, SSA, and IRS, `http://www.census.gov/sipp/synth_data.html`

[MEPS] Medical Expenditure Panel Survey, Restricted Data Files Available at Data Centers, `http://www.meps.ahrq.gov/mepsweb/data_stats/onsite_datacenter.jsp`

[SSDS] Social Science Data Services, `http://libraries.mit.edu/guides/subjects/data/access/restricted.html`

[NCESl] NCES Confidentiality procedures, `http://nces.ed.gov/StatProg/confproc.asp`

[GKRP05] Gomatam, S., Karr, A.F., Reiter, J.P., Sanil, A.P.: Data dissemination and disclosure limitation in a world without microdata: A risk-utility framework for remote access servers. Statistical Science 20, 163–177 (2005)

[J89] Jaro, A.M.: 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)

[KKORS06] Karr, A.F., Kohnen, C.N., Oganian, A., Reiter, J.P., Sanil, A.P.: Framework for Evaluating the Utility of Data Altered to Protect Confidentiality. The American Statistician 60(3), 224–232 (2006)

[Ogan03] Oganian, A.: Security and Information Loss in Statistical Database Protection. Ph. D. thesis, Universitat Politecnica de Catalunya (2003)

[PSD06] Oganian, A., Karr, A.F.: Combinations of SDC Methods for Microdata Protection. In: Domingo-Ferrer, J., Franconi, L. (eds.) PSD 2006. LNCS, vol. 4302, pp. 102–113. Springer, Heidelberg (2006)

[OganKarr10] Oganian, A., Karr, A.F.: Making Methods that Preserve Positivity Constraints in Microdata. Journal of Statistical Planning and Inference (to appear)

[Reit02] Reiter, J.P.: Satisfying disclosure restrictions with synthetic data sets. Journal of Official Statistics 18, 531–544 (2002)

Measurement Error and Statistical Disclosure Control

Natalie Shlomo

Southampton Statistical Sciences Research Institute, University of Southampton, Highfield, Southampton, SO17 1BJ, United Kingdom
N.Shlomo@Soton.ac.uk

Abstract. Statistical agencies release microdata to researchers after applying statistical disclosure control (SDC) methods. Noise addition is a perturbative SDC method which is carried out by adding independent random noise to a continuous variable or by misclassifying values of a categorical variable according to a probability mechanism. Because these errors are purposely introduced into the data by the statistical agency, the perturbation parameters are known and can be used by researchers to adjust statistical inference through measurement error models. However, statistical agencies rarely release perturbation parameters and therefore modifications to SDC methods are proposed that a priori ensure valid inferences on perturbed datasets.

Keywords: Additive noise, Post-randomisation method, Reliability ratio.

1 Introduction

Statistical disclosure control (SDC) methods are becoming increasingly important due to the growing demand for information provided by statistical agencies. More statistical agencies are releasing microdata from social surveys typically under licensing agreements or through data archives. SDC methods aim to prevent sensitive information about individual respondents from being disclosed.

In any released microdata, directly identifying key variables, such as name, address or id numbers, are removed. Disclosure risk arises from attribute disclosure where small counts on cross-classified indirect identifying key variables (such as: age, gender, place of residence, occupation, etc.) can be used to identify an individual and confidential information may be learnt. Identifying key variables are typically categorical since statistical agencies will often coarsen the data before its release. Sensitive variables are continuous (e.g., income) or categorical (e.g., health status). SDC methods can be non-perturbative by limiting the amount of information released or perturbative by altering the data in the microdata. Examples of non-perturbative SDC methods are global recoding, suppression and sub-sampling (see Willenborg and De Waal, 2001). Perturbative methods for continuous variables include adding random noise (Kim, 1986, Fuller, 1993, Brand, 2002), micro-aggregation (replacing values with their average within groups of records) (Anwar 1993, Domingo-Ferrer and Mateo-Sanz, 2002), rounding to a pre-selected rounding base, and rank swapping (swapping values between pairs of records within small groups) (Dalenius and Reiss, 1982, Fienberg and McIntyre, 2005). Perturbative methods for categorical variables

J. Domingo-Ferrer and E. Magkos (Eds.): PSD 2010, LNCS 6344, pp. 118–126, 2010

include record swapping (typically swapping geography variables) and a more general post-randomization method (PRAM) where categories of variables are changed or not changed according to a prescribed probability matrix and a stochastic selection process (Gouweleeuw, et al., 1998).

Perturbative methods can be applied to either the identifying key variables or the sensitive variables or both. In the first case identification of a unit is rendered more difficult, and the probability that a unit can be identified is reduced. In the second case, even if an intruder succeeds in identifying a unit by using the values of the indirect identifying key variables, the sensitive variable would hardly disclose any useful information on the particular unit as they have been perturbed.

In this paper, we focus on perturbative SDC methods which purposely introduce measurement errors into the microdata: additive random noise for a continuous variable and misclassification for a categorical variable. Assuming that the SDC parameters are released by the statistical agency, researchers can use these parameters to correct statistical inferences through measurement error models (Fuller, 1987). Following Fuller, 1993, we demonstrate a measurement error model for a simple linear regression on a perturbed dataset. Statistical agencies, however, rarely release SDC parameters due to confidentiality constraints. In this case, statistical agencies need to modify SDC methods so that researchers can make valid inferences on perturbed datasets.

Section 2 focuses on additive random noise to continuous variables and the impact on a simple regression analysis. An SDC method of correlated noise is proposed that preserves the sufficient statistics and allows valid inference from the regression model on the perturbed data. Section 3 focuses on misclassification of a categorical variable through PRAM and the impact on a simple regression model and a chi-square test for independence for a two dimensional table. By placing the property of invariance on the probability mechanism used in PRAM, some statistical inferences can be preserved exactly on the perturbed datasets. We conclude in Section 4 with a discussion on how these SDC methods can be implemented 'on the fly' so that they can be tailored specifically to the analysis.

2 Adding Noise to Continuous Variables

Additive random noise is an SDC method that is carried out on continuous variables. In its basic form, random noise is generated independently and identically distributed with a mean of zero and a positive variance which is determined by the statistical agency. A zero mean ensures that no bias is introduced into the original variable. The random noise is then added to the original variable. There are also more complex mixture models that can be used for adding noise which achieve higher protection levels since it has been found that additive random noise can yield high re-identification risk (Kargupta, et al., 2005).

Adding random noise to a continuous variable will not alter the mean value of the variable for large datasets but will introduce more variance depending on the variance parameter used to generate the noise. This will impact on the ability to make statistical inference, particularly for estimating parameters in a regression analysis. The ease of analysis in a regression model for a variable subject to additive noise

depends on whether the variable is used as the dependent variable or as the independent variable (or both). Standard regression model theory accounts for errors in the dependent variable and therefore adding more noise to the dependent variable should not affect the estimation of the slope parameters.

As an example, assume a simple regression model with a dependent variable y_i that has been subjected to independently generated Gaussian additive noise η_i with a mean of 0 and a positive variance σ_η^2. Assume also an independent variable x_i that is error free. The model is:

$$\begin{cases} y_i = \alpha + \beta x_i + \varepsilon_i, & i = 1,..., n \\ y_i = y_i^* + \eta_i \end{cases}$$

where y_i^* denotes the true but unobserved value of the dependent variable. If we regress y_i on x_i, then the least squares slope coefficient is:

$$\beta = \frac{Cov(y,x)}{Var(x)} = \frac{Cov(y^* + \eta, x)}{Var(x)} = \frac{Cov(y^*, x) + Cov(\eta, x)}{Var(x)} = \frac{Cov(y^*, x)}{Var(x)} \quad (1)$$

since $Cov(\eta, x) = 0$. The additive noise on the dependent variable does not bias the slope coefficient, however it will increase its standard error due to the increase in the variance: $Var(y) = Var(y^*) + Var(\eta)$.

Complications arise when the random noise η_i is added to the independent variable in the regression model. The model is now:

$$\begin{cases} y_i = \alpha + \beta x_i + \varepsilon_i, & i = 1,..., n \\ x_i = x_i^* + \eta_i \end{cases}$$

where x_i^* denotes the true but unobserved value of the independent variable. Now regressing y_i on x_i, we obtain for the least squares slope coefficient:

$$\beta = \frac{Cov(y,x)}{Var(x)} = \frac{Cov(y, x^* + \eta)}{Var(x^*) + Var(\eta)} = \frac{Cov(y, x^*) + Cov(y, \eta)}{Var(x^*) + Var(\eta)} = \frac{Cov(y, x^*)}{Var(x^*) + Var(\eta)} \quad (2)$$

since $Cov(y, \eta) = 0$. The additive noise on the independent variable biases the slope coefficient downwards. This is referred to as attenuation. In this case, the researcher needs suitable methodology to deal with the measurement error in the independent variable.

Noting that the estimate for the least squares slope coefficient follows:

$$\hat{\beta} \xrightarrow{p} \frac{Cov(y, x^*)}{Var(x^*) + Var(\eta)} = \frac{\beta \sigma_{x^*}^2}{\sigma_{x^*}^2 + \sigma_\eta^2} = \beta(1 + \sigma_\eta^2 / \sigma_{x^*}^2)^{-1} \quad (3)$$

Fuller, 1987 defines the term $(1 + \sigma_\eta^2 / \sigma_{x^*}^2)^{-1}$ as the reliability ratio denoted by λ. In a very simple measurement error model, a consistent estimate of the slope

coefficient can be obtained by dividing the least-squares estimate from the perturbed dataset by λ.

To calculate the reliability ratio and allow valid inferences, it is assumed that the variance parameter σ_η^2 used to generate the random noise is released by the statistical agency to researchers. This, however, is rarely the case since statistical agencies generally do not reveal parameters of SDC methods. In order to compensate for the measurement error, statistical agencies should employ a different method for adding random noise based on generating noise that is correlated with the original continuous variable. Kargupta, et al., 2005 noted that re-identification is more difficult when adding correlated noise. Correlated noise addition ensures that sufficient statistics (means, variances and correlations) of the original continuous variables are preserved (see also: Kim, 1986 and Tendick and Matloff, 1994). One algorithm for generating correlated random noise for a continuous variable x that is easy to implement is as follows:

Procedure for a univariate case: Define a parameter δ which takes a value greater than 0 and less than equal to 1. When $\delta = 1$ we obtain the case of fully modeled synthetic data. The parameter δ controls the amount of random noise added to the variable *x*. After selecting a δ, calculate: $d_1 = \sqrt{(1-\delta^2)}$ and $d_2 = \sqrt{\delta^2}$. Now, generate random noise ε independently for each record with a mean of $\mu' = \mu\{(1-d_1)/d_2\}$ and the original variance of the variable σ^2. Typically, a Normal Distribution is used to generate the random noise. Calculate the perturbed variable x_i' for each record i (i=1,..,n) as a linear combination: $x_i' = d_1 \times x_i + d_2 \times \varepsilon_i$. Note that

$E(x') = d_1 E(x) + d_2[(1-d_1)/d_2]E(x) = E(x)$ and

$Var\ (x') = (1-\delta^2)Var\ (x) + \delta^2 Var\ (x) = Var\ (x)$ since the random noise is generated independently to the original variable *x*. This algorithm can be extended to the multivariate case for simultaneously adding correlated random noise to several variables which preserves the sufficient statistics of each variable as well as the covariance matrix. (see Shlomo and De Waal, 2008).

Table 1 presents a simulation study which demonstrates the effects of adding random noise and correlated random noise to variables in a simple regression model. Each row in the table represents a different scenario consisting of the type of noise added (random or correlated) and whether the noise was added to the dependent variable, independent variable or both. We generate 1000 records where $x_i \sim N(20,9)$, $\varepsilon_i \sim N(0,3)$ and the model is: $y_i = 3 + 3x_i + \varepsilon_i$ (the true intercept is 3 and the true slope coefficient is 3). We generate Gaussian random noise: $u_i \sim N(0,1)$ as well as correlated noise according to the procedure described above with $\delta = 0.1$. Note that in this case, the reliability ratio is: $\lambda = 9/10$. We repeat for 1000 replications and present in Table 1 the average regression parameters and their standard errors.

The attenuation of the slope coefficient in Table 1 when adding random noise to the independent variable can be seen (from a value of 3.000 to a value of 2.701). We divide the slope coefficient that was estimated from the perturbed data by the

Table 1. Simulation study for estimating regression coefficients from data subjected to additive and correlated noise (average across 1000 replications)

Model	Intercept		Slope	
	Estimate	SE	Estimate	SE
Original model	2.997	0.363	3.000	0.018
Additive Random Noise on:				
Dependent variable	3.008	0.438	3.000	0.022
Independent variable	8.976	0.672	2.701	0.033
Both dependent and independent variables	6.985	0.512	2.801	0.025
Correlated Noise on:				
Dependent variable	3.285	0.413	2.986	0.020
Independent variable	3.299	0.409	2.985	0.020
Both dependent and independent variable (multivariate method)	3.010	0.444	2.999	0.022

reliability ratio, $\lambda = 9/10$ and obtain a consistent estimate for the slope, eg. $2.701 \times 10/9 = 3.000$. The intercept can then be consistently estimated.

As can be seen in Table 1, adding correlated noise to the independent variable, the dependent variable, or both variables provides estimates for the slope coefficient and intercept that are close to the true value. Standard errors are higher which reflect the added uncertainty due to the noise addition.

3 Misclassification of Categorical Variables

As described in Shlomo and De Waal (2008), we examine the use of the Post-randomization Method (PRAM) (Gouweleeuw, et al., 1998) to perturb a categorical variable. This method is a more general case of record swapping. Willenborg and De Waal (2001) describe the process as follows:

Let $\mathbf{P}$ be a $L \times L$ transition matrix containing conditional probabilities $p_{ij} = p(\text{perturbed category is } j \mid \text{original category is } i)$ for a categorical variable with L categories. Let $\mathbf{t}$ be the vector of frequencies and $\mathbf{v}$ the vector of relative frequencies: $\mathbf{v} = \mathbf{t}/n$, where n is the number of records in the microdata. For each record of the data, the category of the variable is changed or not changed according to the prescribed transition matrix $\mathbf{P}$ and the result of a random draw from a multinomial distribution with parameters p_{ij} (j=1,…,L). If the j-th category is selected, category i is moved to category j. When i = j, no change occurs.

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

re-identification under PRAM can generally be high and depends on the values of the diagonal of $\mathbf{P}$. The method introduces 'uncertainty' into the true values and this adds to the protection level.

Under PRAM, joint distributions between perturbed and unperturbed variables are distorted which impacts on statistical inference. Variables that typically undergo PRAM are the demographic and geographic identifiers in the microdata which are commonly used in statistical analysis as explanatory variables, for example in regression models. If the statistical agency releases the probability transition matrix **P** then measurement error models can be used. As an example, instead of generating the normally distributed x variable for 1000 records in Section 2, we generate a dichotomous z variable obtaining a value of 1 with a probability of 0.6 and 0 otherwise. Note that $Var\ (z) = 240$ in the dataset. The residuals are generated as before with $\varepsilon_i \sim N(0,3)$ and $y_i = 3 + 3z_i + \varepsilon_i$. We carry out a PRAM procedure on the z variable where the probability matrix P has the diagonal $p_{00} = 0.8$ for $z = 0$ and $p_{11} = 0.85$ for $z = 1$. The average least squares estimate for the slope coefficient after 1000 replications is reduced from $\hat{\beta}_1 = 3.000$ in the original dataset to $\hat{\beta}_1 = 1.931$ in the perturbed dataset. In order to calculate the reliability ratio defined in Section 2 to compensate for the measurement error, we need to calculate the additional variance to z due to PRAM, $Var\ (z^* \mid \mathbf{P})$, where z^* is the perturbed categorical variable. This is based on two independent binomially distributed random variables with parameters (z_0, p_{00}) and $(n - z_0, p_{11})$ respectively, where $z_0 = \sum I(z_i = 0)$ and n=1000:

$V(\mathbf{z}^* \mid \mathbf{P}) = z_0 p_{00}(1 - p_{00}) + (n - z_0) p_{11}(1 - p_{11}) = 137.2$ The reliability ratio is equal to: $\lambda = 240/(240 + 137.2) = 0.64$. Dividing the slope coefficient estimated from the perturbed dataset by the reliability ratio, $\lambda = 0.64$, we obtain a consistent estimate for the slope, eg. $1.931/0.64 = 3.035$. The calculation of the reliability ratio for the measurement error model depends on the release of the probability matrix **P.** As mentioned, statistical agencies do not generally release SDC parameters. For a regression model, the method of correlated noise addition described in Section 2 can also be applied to a categorical dummy variable to ensure consistent estimation of regression parameters and valid inferences in the perturbed dataset.

Categorical variables imply other types of statistical analysis, such as the chi-square test for independence. Statistical agencies can compensate for the measurement error induced by PRAM by ensuring that the marginal frequency counts of the perturbed variable are approximately equal to the marginal frequency counts of the original variable. This is done by placing the condition of invariance on the transition matrix $\mathbf{P}$, i.e. $\mathbf{tP} = \mathbf{t}$ where **t** is the vector of frequencies. The property of invariance means that the expected values of the marginal distribution of the variable under perturbation are preserved. In order to obtain the exact marginal distribution, we propose using a "without" replacement strategy for selecting the categories to change (or not change). This is carried out by calculating the expected number of categories to change according to the probability matrix and then drawing a random sample without replacement of those categories and changing their values. This

procedure ensures exact marginal distributions as well as reduces the additional variance that is induced by the perturbation.

For the purpose of carrying out a chi- square test for independence on a frequency table, the variables spanning the table should be perturbed as a single variable by cross-classifying the categories. For example, if we are interested in analyzing associations in health status with 2 categories and ethnicity with 7 categories, we combine the two variables to obtain a single variable with 14 categories. This single variable is perturbed using an invariant probability matrix of size 14×14 and drawing samples of categories to change without replacement. The resulting chi-square statistic from the perturbed dataset will be equal to the chi-square statistic of the original dataset. To demonstrate, we again generate a dichotomous z variable obtaining a value of 1 with a probability of 0.6 and zero otherwise and $\varepsilon_i \sim N(0,3)$ for 1000 records. We define

$u_i = \exp(3z_i + \varepsilon_i)/(1 + \exp(3z_i + \varepsilon_i))$ and classify into a dichotomous variable q obtaining the value of 1 if $u_i \geq 0.7$ and the value of zero otherwise. We are interested in a chi-square statistic for the two dimensional table spanned by z and q. Tables 2a to 2e contain the results of one realization out of a 1000 replications where Table 2a presents the counts and chi-square statistic from the original data. The other tables were calculated as follows:

- Table 2b: PRAM procedure on the z variable and an independent mechanism for changing categories, denoted by z^*,
- Table 2c: PRAM procedure on the z variable under the property of invariance and the without replacement strategy for selecting categories to change, denoted by z^{*I},
- Table 2d: Similar to Table 2b with PRAM applied on the combined single variable obtained by cross-classifying z and q , denoted z^* and q^*,
- Table 2e: Similar to Table 2c with PRAM applied on the combined single variable obtained by cross-classifying z and q, denoted z^{*I} and q^{*I} .

The diagonals of the probability matrices are dominant between 0.8 and 0.85.

All of the chi-square statistics in Tables 2b to 2e are significant and it is reassuring that none of the perturbed tables provided an erroneous conclusion of independence compared to the original table, but this may not always be the case. It is clear that only Table 2e can give the exact value for the chi-square statistic under the property of invariance and the without replacement strategy for selecting categories to change on the combined cross-classified variable.

4 Discussion

Statistical agencies prepare microdata for release by applying SDC methods according to their disclosure control standards and policies for data protection. The protected microdata are then typically delivered to a data archive where approved

Tables 2. Study of chi-square tests for independence under PRAM (one realization out of 1000 replications)

Table 2a: Original counts

$\chi^2 = 420.7$

q	z		
	0	1	Total
0	307	112	419
1	58	523	581
Total	365	635	1000

Table 2b: z perturbed randomly

$\chi^2 = 168.8$

q	z*		
	0	1	Total
0	258	180	438
1	107	455	562
Total	365	635	1000

Table 2c: z perturbed under invariance

$\chi^2 = 181.0$

q	z^{*I}		
	0	1	Total
0	254	165	419
1	111	470	581
Total	365	635	1000

Table 2d: z and q perturbed randomly

$\chi^2 = 287.6$

q*	z*		
	0	1	Total
0	288	133	421
1	91	488	579
Total	379	621	1000

Table 2e: z and q perturbed under invariance

$\chi^2 = 420.7$

q^{*I}	z^{*I}		
	0	1	Total
0	307	112	419
1	58	523	581
Total	365	635	1000

researchers can download the data to their personal computers. Since the microdata has many variables, the protection afforded by pre-defined SDC methods is limited. In addition, we have shown that when statistical agencies do not release the parameters of the SDC methods, it is almost impossible to develop measurement error models for analysis. We demonstrated how the analytical properties of the data can be preserved for a regression analysis and a chi-square test of independence by modifying standard SDC methods. However, developing SDC methods that a priori preserve the analytical properties of the data for all types of statistical analysis is a hard problem. Two possible ways of solving this problem are:

- Develop a remote analysis server where software code is submitted and run on the original data and the outputs checked for disclosure risk prior to their release.
- Develop specialized software that can tailor SDC methods applied to the microdata before its release according to the type of analysis specified. The SDC methods are applied 'on-the-fly' in the software package. The software would also include flexible table generation since aggregated data is a non-perturbative SDC method for microdata.

Implementing 'on the fly' SDC methods would not only increase the utility in the microdata for the specified analysis but would also reduce disclosure risk.

Statistical agencies need to carefully consider whether releasing SDC parameters, such as the variance used to generate additive noise, actually increases disclosure risk. While SDC methods can be modified to preserve some analytical properties of the perturbed microdata, it is only through the release of SDC parameters that researchers can compensate for measurement error and ensure correct inferences.

References

1. Anwar, N.: Micro-Aggregation-The Small Aggregates Method. Informe Intern. Luxembourg, Eurostat (1993)
2. Brand, R.: Micro-data Protection Through Noise Addition. In: Domingo-Ferrer, J. (ed.) Inference Control in Statistical Databases. LNCS, vol. 2316, pp. 97–116. Springer, Heidelberg (2002)
3. Dalenius, T., Reiss, S.P.: Data Swapping: A Technique for Disclosure Control. Journal of Statistical Planning and Inference 7, 73–85 (1982)
4. Domingo-Ferrer, J., Mateo-Sanz, J.: Practical Data-Oriented Microaggregation for Statistical Disclosure Control. IEEE Transactions on Knowledge and Data Engineering 14(1), 189–201 (2002)
5. Fienberg, S.E., McIntyre, J.: Data Swapping: Variations on a Theme by Dalenius and Reiss. In: Domingo-Ferrer, J., Torra, V. (eds.) PSD 2004. LNCS, vol. 3050, pp. 14–29. Springer, Heidelberg (2004)
6. Fuller, W.A.: Measurement Error Models. Wiley, New York (1987)
7. Fuller, W.A.: Masking Procedures for Micro-data Disclosure Limitation. Journal of Official Statistics 9, 383–406 (1993)
8. Gourweleeuw, J., Kooiman, P., Willenborg, L.C.R.J., De Wolf, P.P.: Post Randomisation for Statistical Disclosure Control: Theory and Implementation. Journal of Official Statistics 14, 463–478 (1998)
9. Kargupta, H., Datta, S., Wang, Q., Ravikumar, K.: Random Data Perturbation Techniques and Privacy Preserving Data Mining. Knowledge and Information Systems 7(4), 387–414 (2005)
10. Kim, J.J.: A Method for Limiting Disclosure in Micro-data Based on Random Noise and Transformation. In: ASA Proceedings of the Section on SRM, pp. 370–374 (1986)
11. Shlomo, N., De Waal, T.: Protection of Micro-data Subject to Edit Constraints Against Statistical Disclosure. Journal of Official Statistics 24(2), 1–26 (2008)
12. Tendick, P., Matloff, N.: A Modified Random Perturbation Method for Database Security. ACM Transactions on Database Systems 19(1), 47–63 (1994)
13. Willenborg, L.C.R.J., De Waal, T.: Elements of Statistical Disclosure Control in Practice. LNS, vol. 155. Springer, New York (2001)

Semantic Microaggregation for the Anonymization of Query Logs

Arnau Erola[1], Jordi Castellà-Roca[1],
Guillermo Navarro-Arribas[2], and Vicenç Torra[2]

[1] Departament d'Enginyeria Informàtica i Matemàtiques, UNESCO Chair in Data Privacy, Universitat Rovira i Virgili, Av. Països Catalans 26, E-43007 Tarragona, Spain
{arnau.erola,jordi.castella}@urv.cat

[2] IIIA, Institut d'Investigació en Intel·ligència Artificial - CSIC, Consejo Superior de Investigaciones Científicas, Campus UAB s/n, 08193 Bellaterra (Catalonia, Spain)
{vtorra,guille}@iiia.csic.es

Abstract. The publication of Web search logs is very useful for the scientific research community, but to preserve the users' privacy, logs have to be submitted to an anonymization process. Random query swapping is a common technique used to protect logs that provides k-anonymity to the users in exchange for loss of utility. With the assumption that by swapping queries semantically close this utility loss can be reduced, we introduce a novel protection method that semantically microaggregates the logs using the Open Directory Project. That is, we extend a common method used in statistical disclosure control to protect search logs from a semantic perspective. The method has been tested with a random subset of AOL search logs, and it has been observed that new logs improve the data usefulness.

1 Introduction

The Web search engines (WSE) in the Internet such as Google, Yahoo, or Bing, store information of the queries made by the users, normally referred as *search or query logs*. These data allow the WSE to provide personalized Web search to the users [8] and are a great source of information for researchers or marketing companies [10], but at the same time their publication may expose the privacy of the users from which the logs were generated. There is at least one well known case of released search logs with poor anonymization, which have been shown to reveal enough information to re-identify some users. The release was done by AOL in an attempt to help the information retrieval research community, and ended up with not only important damage to AOL users privacy, but also a major damage to AOL itself with several class action suits and complaints against the company [7,15]

Moreover, query logs are a great economic source for the WSE, for instance Google had a revenue of 21,128.5 million dollars in 2008 from advertisement [9], which is strongly based in the information gathered by their search engine.

J. Domingo-Ferrer and E. Magkos (Eds.): PSD 2010, LNCS 6344, pp. 127–137, 2010.

WSEs also charge law enforcement agencies for the access to user or group profiles [26,21].

In this paper we address the privacy problem exposed by the WSE query logs, which can be made publicly available without risking the privacy of their users. To that end we follow the same ideas found in statistical disclosure control, proposing a novel microaggregation method to anonymize query logs. This approach ensures a high degree of privacy, providing k-anonymity at user level, while preserving some of the data usefulness. Moreover, and unlike most of the previous work, our approach takes into account the semantics of the queries made by the user in the anonymization process making use of information obtained from the Open Directory Project [17].

The paper is organized as follows. Section 2 introduces microaggregation and our motivation and approach describe our approach for the semantic anonymization of query logs. In Section 3 we detail our proposal, and Section 4 presents our results in terms of protection and utility. Section 5 discusses the related work, and finally, Section 6 concludes the paper.

2 Towards a Semantic Microaggregation for Query Logs

Microaggregation is a popular statistical disclosure control technique, which provides privacy by means of clustering the data into small clusters and then replacing the original data by the centroids of the corresponding clusters.

In this paper we propose a novel microaggregation method for query logs taking into account the semantics of the queries made by the users. In this section, we overview microaggregation and discuss the motivations of our proposal.

2.1 Microaggregation

In microaggregation, privacy is ensured because all clusters have at least a predefined number of elements, and therefore, there are at least k records with the same value. Note that all the records in the cluster replace a value by the value in the centroid of the cluster. The constant k is a parameter of the method that controls the level of privacy. The larger the k, the more privacy we have in the protected data.

Microaggregation was originally [3] defined for numerical attributes, but later extended to other domains. E.g., to categorical data in [23] (see also [5]), and in constrained domains in [24].

From the operational point of view, microaggregation is defined in terms of partition and aggregation:

- **Partition.** Records are partitioned into several clusters, each of them consisting of at least k records.
- **Aggregation.** For each of the clusters a representative (the centroid) is computed, and then original records are replaced by the representative of the cluster to which they belong to.

From a formal point of view, microaggregation can be defined as an optimization problem with some constraints. We give a formalization below using u_{ij} to describe the partition of the records in the sensitive data set X. That is, $u_{ij} = 1$ if record j is assigned to the ith cluster. Let v_i be the representative of the ith cluster, then a general formulation of microaggregation with g clusters and a given k is as follows:

$$\begin{array}{ll} \text{Minimize} & SSE = \sum_{i=1}^{g} \sum_{j=1}^{n} u_{ij}(d(x_j, v_i))^2 \\ \text{Subject to} & \sum_{i=1}^{g} u_{ij} = 1 \text{ for all } j = 1, \ldots, n \\ & 2k \geq \sum_{j=1}^{n} u_{ij} \geq k \text{ for all } i = 1, \ldots, g \\ & u_{ij} \in \{0, 1\} \end{array}$$

For numerical data it is usual to require that $d(x, v)$ is the Euclidean distance. In the general case, when attributes $\mathbf{V} = (V_1, \ldots, V_s)$ are considered, x and v are vectors, and d becomes $d^2(x, v) = \sum_{V_i \in \mathbf{V}} (x_i - v_i)^2$. In addition, it is also common to require for numerical data that v_i is defined as the arithmetic mean of the records in the cluster. I.e., $v_i = \sum_{j=1}^{n} u_{ij} x_i / \sum_{j=1}^{n} u_{ij}$. As the solution of this problem is NP-Hard [18] when we consider more than one variable at a time (multivariate microaggregation), heuristic methods have been developed. One of such methods is MDAV (*Maximum Distance to Average Vector*) [4],

Note that when all variables are considered at once, microaggregation is a way to implement k-anonymity [20,22].

2.2 Motivations of Our Proposal

A key point for the microaggregation of search logs, is how the users are clustered. If the users in the same cluster do not share any interest, the protected search-logs can be useless, i.e. the resulting search logs are too much distorted and we can not obtain useful information from them.

For example, we can consider two soccer supporters, and two anti-sports users. If we create a cluster of size two with a soccer supporter and an anti-sports users we can obtain non-valid results. The entries of the protected search logs are confusing. On the other hand, if the two soccer supporters are in the same cluster, the protected search-logs provide more reliable results.

Thus, we should create the groups of users taking into consideration their interests. The users with common interests between them should be grouped in the same cluster. In order to do so, we should be capable to determine if their interests are closer, i.e. we need a tool to compute the semantic distance of two queries.

In this work, we use the Open Directory Project (ODP) [17] to compute the semantic distances between users. The ODP is the most widely distributed data base of Web content classified by humans. ODP data powers the core directory services for some the most popular portals and search engines on the Web, including AOL Search, Netscape Search, Google, Lycos, and HotBot, and hundreds of others. Thus, a query result using them is hardly influenced by the ODP classification. ODP uses a hierarchical ontology structure to classify sites according

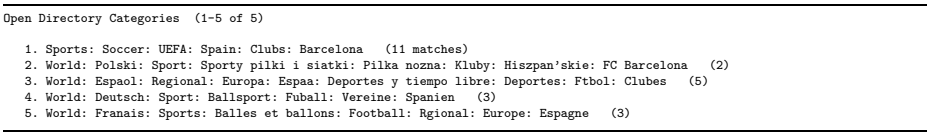


Fig. 1. Example of ODP query result

their themes. For example, when we search for *Barcelona FC*, ODP returns a list of categories which the query belongs (Figure 1). Each result starts with a root category and then are deeper categories in the ODP tree.

Our proposal groups users with common interests using the ODP classification. We consider that the users with common interest are those who have more terms in the same categories.

3 ODP-Based Microaggregation of Query Logs

The method proposed has the following steps:

- ODP classification (or data preparation).
- Partition.
- Aggregation.

These steps are described in the following sections.

3.1 ODP Classification

When querying the ODP, the returned categories can be divided in depth levels. Let l be a parameter of our system that identifies the maximum depth level in the ODP hierarchy, that our system works. For example, if we have the classification $Sports : Soccer : UEFA : Spain : Clubs : Barcelona$ and $l = 1$, we work with the root category $Sports$; when $l = 2$ we work with $Sports : Soccer$; ...

In the ODP classification step we know the set of users $\{u_1, \ldots, u_n\}$ and their sets ofqueries $Q = \{Q_1, \ldots, Q_n\}$, where $Q_i = \{q_1, \ldots, q_m\}$ are the queries of the user u_i. Every query q_j can have several terms t_s, i.e. $q_j = \{t_1, \ldots, t_r\}$.

For every term t_s, we obtain its classification at level $L \in \{1, \ldots, l\}$ using the ODP. Next, we create the matrix that contains the number of queries for each user and category at level L, M^L_{UxC} (classification matrix). Please, note that, we obtain one matrix for every level $L \in \{1, \ldots, l\}$. Finally, we use the M^L_{UxC} matrices in order to compute the incidence matrix that contains the semantic similiarity of the users M_{UxU}. The process works as follows:

1. Obtain the classification matrices M^L_{UxC} using Algorithm 1.
2. Obtain the incidence matrix M_{UxU} using Algorithm 2, i.e. the addition of all coincidences between two users in the classification matrix.

Algorithm 1. Algorithm for computing the matrices M^L_{UxC} where $L = \{1, \ldots, l\}$

Require: the maximum depth l for the ODP categories
Require: the set of users $U = \{u_i, \ldots, u_n\}$
Require: the set of queries $Q_i = \{q_1, \ldots, q_m\}$ of each user u_i
Require: the set of terms $\{t_1, \ldots, t_r\}$ of each query q_j
Ensure: $\{M^1_{UxC}, \ldots, M^l_{UxC}\}$, i.e. for every level L, the matrix M^L_{UxC} with the number of queries for each category and user in the depth L

```
for L = 1 to L = l do
  for u_i ∈ {u_1, ..., u_n} do
    for q_j ∈ Q_i = {q_1, ..., q_m} do
      for t_s ∈ q_j = {t_1, ..., t_r} do
        obtain the categories c at depth L for the term t_s using ODP;
        if c ∈ M^L_UxC then
          add one to the cell (u_i, c) of M^L_UxC;
        else
          add the column c to M^L_UxC;
          initialize the cell (u_i, c) of M^L_UxC to one;
        end if
      end for
    end for
  end for
end for
return {M^1_UxC, ..., M^l_UxC}.
```

Algorithm 2. Algorithm for computing the matrix M_{UxU}

Require: the calssification matrices $\{M^1_{UxC}, \ldots, M^l_{UxC}\}$
Ensure: M_{UxU}

```
for M^L_UxC ∈ {M^1_UxC, ..., M^l_UxC} do
  for each column c_j ∈ M^L_UxC do
    for each row u_i ∈ M^L_UxC do
      for each row u_ρ ∈ M^L_UxC do
        add x to the cell (u_i, u_ρ) of matrix M_UxU, where
        x = min((u_i, c_j), (u_ρ, c_j));
      end for
    end for
  end for
end for
return M_UxU.
```

3.2 Partition

The partition step creates groups of k users with similar interests using Algorithm 3.

Let assume that u_i and u_ρ are the most similar users in the set. We calculate the users' similarity using the user incidence matrix M_{UxU}, (see Section 3.1). The most similar users are those that have the highest value in the matrix. Next, we include u_i and u_ρ to the cluster. If the group size k is two, we delete u_i

Algorithm 3. Algorithm for computing the clusters $Z = \{z_1, \ldots, z_\gamma\}$ of users

Require: the set of users $U = \{u_1, \ldots, u_n\}$
Require: the incidence matrix M_{UxU}
Require: the clusters size k
Ensure: the clusters $Z = \{z_1, \ldots, z_\gamma\}$ of users
  $U' \leftarrow U$;
  while size$(U') \leq k$ **do**
    obtain the cluster z_j of k users using the Algorithm 4 and U';
    remove the users $u_i \in z_j$ form U';
    add z_j to the set Z;
  end while
  if $0 <$ size$(U') \leq k$ **then**
    create z_γ with the users $u_i \in U'$;
  end if
  return $Z = \{z_1, \ldots, z_\gamma\}$.

and u_ρ records from the incidence matrix and we repeat the process to obtain a new cluster. When, the group size is bigger than two, we merge the columns and rows of u_i and u_ρ creating a new user u'. u' is the addition of both users, u_i and u_ρ. Let assume, that u_ξ is the most similar user with u'. Next, we include u_ξ to the cluster with u_i and u_ρ. The method executes this process $k-2$ times.

Note that, if there are at least k users in the matrix we repeat the previous steps, and if there are less than k users we include them to a new group.

3.3 Aggregation

For every group z_j formed in the partition step, we compute its aggregation by selecting specific queries from each user in the group. That is, given the group of users $z_j = \{u_1, \ldots, u_k\}$ we obtain a new user u_{z_j} as the representative (or centroid) of the cluster, which summarizes the queries of all the users of the cluster. To select such queries we define the following priorities:

- The queries semantically close have more priority.
- Deeper levels in the ODP tree are more important.

The contribution of a user u_i ($Contrib_i$), in the centroid, depends on her number of queries $card(Qi)$, that can be calculated as follows:

$$Contrib_i = \frac{card(Qi)}{\sum_{i=1}^{k} card(Qi)} \tag{1}$$

The number of queries of the centroid is the average of the number of queries of each user u_i of the cluster z_j. Thus, the quota of each user u_i in the new centroid u_{z_j} can be computed as:

$$Quota_i = \frac{card(Qi)}{k} \tag{2}$$

The aggregation method runs the next protocol for each user:

Algorithm 4. Algorithm for computing a cluster z of k users

Require: the incidence matrix M_{UxU}
Require: the clusters size k
Ensure: a cluster z of k users
$M'_{UxU} \leftarrow M_{UxU}$
$z \leftarrow \emptyset$
obtain the two most similar users (u_i, u_ρ), i.e. the cell of M'_{UxU} with the highest value;
add (u_i, u_ρ) to the set z
$M''_{UxU} \leftarrow M'_{UxU}$;
while $(size(z_j) < k)$ **and** $(columns(M'_{UxU}) > 0)$ **do**
  for each column $c_s \in M''_{UxU}$ **do**
    add the cell (c_s, u_ρ) to the cell (c_s, u_i) of the matrix M''_{UxU}
  end for
  for each row $r_s \in M''_{UxU}$ **do**
    add (u_ρ, r_s) to the cell (u_i, r_s)
  end for
  delete the column u_ρ of matrix M''_{UxU};
  delete the row u_ρ of matrix M''_{UxU};
  obtain the new u_i's most similar user u_ρ, i.e. the cell of the user u_i with the highest value;
  add u_ρ to the set z;
end while
return z.

1. Sort logs from all users descending by query repetitions.
2. For each cluster z_j and for each user $u_i \in z_j$:
   (a) While not reaching $Quota_i$:
      i. Add the first query of her sorted list with a probability $Contrib_i \times \#q_j\_repetitions$. For example, if u_i has a query repeated 3 times, and $Contrib_i$ is 0.4, as $3 \cdot 0.4 = 1.2$, the method adds one query to the new log and randomly chooses if it adds the term another time according to the presence probability 0.2.
      ii. Delete the first query of the list.

4 Evaluation

We test our microaggregation method using real data from the AOL logs released in 2006, which corresponds to the queries performed by 650 000 users over three months. We randomly select 1 000 users, which correspond to 55 666 lines of query logs. The usefulness evaluation and the results are presented below.

4.1 Usefulness Evaluation Method

For each user we have her original set of queries and the corresponding protected ones by means of our microaggregation method. All queries can be classified in

categories, i.e., each query is classified in the l first depth levels of the ODP. In order to verify that our method preserves the usefulness of the data (i.e. does not introduce too much perturbation), for a given level l, we count the number of queries of each category that are in the original log as well as in the centroid, ρ. This number is divided by the number of original queries in l obtaining a *semantic remain percentage* (SRP) in the level, χ.

$$SRP = \frac{\rho}{\chi} \tag{3}$$

To summarize, our evaluation method does not only match two equal terms in both logs, but also a term in the protected log that replaces one with closest semantic in the original log.

Using a random partition algorithm, users of each cluster might not be semantically close.

Consider k users $\{u_1, \ldots, u_k\}$ forming a cluster, and her set of queries $Q = \{Q_1, \ldots, Q_k\}$. In the ODP classification:

- Intersection of Q_i and Q_j is $\emptyset$ if $j \neq i$;
- Intersection of Q_i and Q_j is Qi if $j = i$;

Thus, only the queries of a single user in a specific topic will appear in the centroid.

In that case, the number of queries of u_i that apper in the centroid can be calculated using formula 2 and it is known that the sum of all quotas is χ.

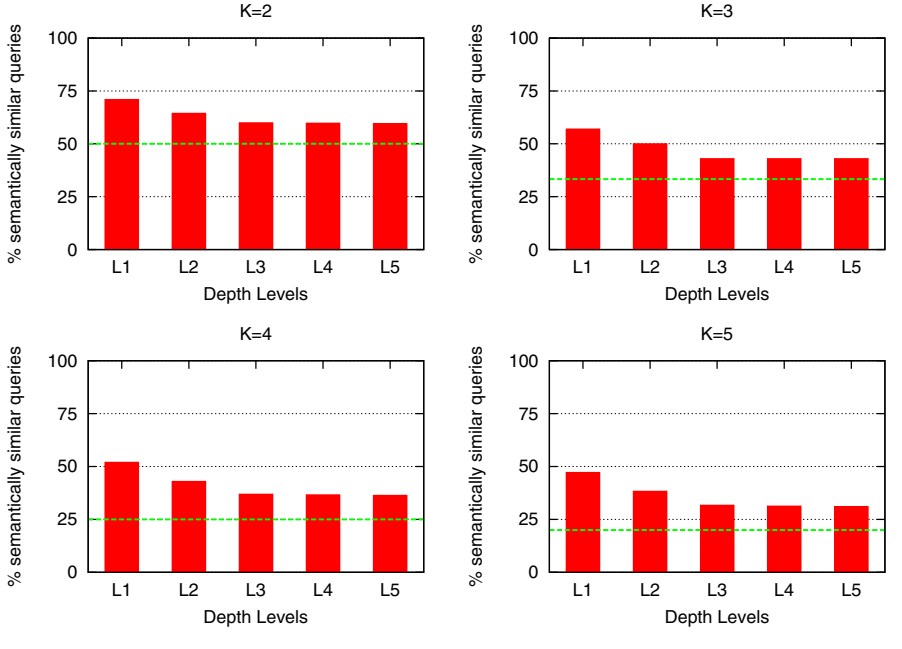


Fig. 2. Semantic similarity percentage of microaggregated logs

Therefore, in the worst case, i.e. when no common interests between users exists, we can calculate the average SRP as:

$$\frac{\sum_{i=1}^{k} \frac{card(Qi)}{\chi}}{k} = \frac{1}{k} \tag{4}$$

4.2 Results

Figure 2 shows, for cluster sizes 2, 3, 4, and 5, the average SRP that users obtain for various l. For a better comparison, we have included the theoretical SRP for a random partition algorithm. It can be observed that our system improves the SRP with all depth levels. When $l = 1$ and $l = 2$, we obtain a gain of nearly 20%. With $l = 3, l = 4$ and $l = 5$, this gain is approximately of 10%.

We have to bear in mind that we are working with a set of 1.000 users, randomly selected from the AOL files. We expect to achieve greater SRP values working with a larger set, because more similar users may be grouped.

5 Related Work

There are several approaches to anonymize query logs in the literature [2], but they are normally reduced to the deletion of specific queries or logs. For instance, in [1] the authors propose a technique to remove infrequent queries, while in [19] a more sophisticated technique is introduced to remove selected queries to preserve an acceptable degree of privacy, or in the case of [13] to choose the publishable queries. Common techniques used in statistical disclosure control (SDC) have not been applied to this specific problem until very recently [16,12]. Moreover, these systems use spelling similarities to link users, i.e. two users will be grouped if they had submit sintactic similar queries. Therefore, they cannot distinguish different senses of a term, if it has more than one.

The use of supporting semantic taxonomies to anonymize query logs was considered in [11] were the authors anonymize the set of queries made by a user by generalizing the queries using WordNet [14]. WordNet is a generic lexical database of the English language, where concepts are interlinked by means of conceptual-semantic and lexical relations. The problem of relying in WordNet when facing the anonymization of query logs is that the query introduced by the user, despite the fact that they might not be in English, can be meaningless in a generic dictionary. We think that better result can be obtained by gathering semantic information from the Open Directory Project (ODP), which its main purpose is precisely to serve as a catalog of the Web by providing a content-based categorization or classification of Web pages. Nevertheless, we need to introduce novel approaches to make the information obtained from the ODP useful. Unlike WordNet, which already has lots of published and tested distances functions, or aggregation operations, ODP lacks this extensive previous work.

It is important to remark that our proposal achieves k-anonymity [20,22] at user level, which guarantees that at least k users are indistinguishable in

the protected version. This guarantees a high degree of privacy, preventing the famous privacy leaks of the AOL logs.

Our proposal might resemble to some readers agglomerative hierarchical clustering methods such as the well known Ward method [25] . This method has been also adapted to perform microaggregation, although in another context, in [4].

6 Conclusions

The existing microaggregation techniques for query logs do not usually take into account the semantic proximity between users, which is negatively reflected in the usefulness of the resulting data. This paper presents a new microaggregation method for search logs based on a semantic clustering algorithm. We use ODP to classify the queries of all users and then aggregate the more semantically close logs. As we have seen, the resulting logs achieves higher usability while preserving k-anonymity.

The good obtained results encourage us to carry on researching in the same line. It will be our future work to find new semantic techniques that allow us to adress the comparison of full sentences. Moreover, new evaluation methods, such as as [6], will be tested to better assess the quality of the results obtained using our system.

Acknowledgment

Partial support by the Spanish MICINN (projects eAEGIS TSI2007-65406-C03-02, TSI2007-65406-C03-01, ARES-CONSOLIDER INGENIO 2010 CSD2007-00004), the Spanish Ministry of Industry, Commerce and Tourism (project TSI-020100-2009-720), and the Government of Catalonia (grant 2009 SGR 1135) is acknowledged. G. Navarro-Arribas enjoys a Juan de la Cierva grant (JCI-2008-3162) from the Spanish MICINN.

The authors are solely responsible for the views expressed in this paper, which do not necessarily reflect the position of UNESCO nor commit that organization.

References

1. Adar, E.: User 4xxxxx9: Anonymizing query logs. In: Query Logs workshop (2007)
2. Cooper, A.: A survey of query log privacy-enhancing techniques from a policy perspective. ACM Transactions on the Web 2(4) (2008)
3. Defays, D., Nanopoulos, P.: Panels of enterprises and confidentiality: the small aggregates method. In: Proc. of 1992 Symposium on Design and Analysis of Longitudinal Surveys, Statistics Canada, pp. 195–204 (1993)
4. Domingo-Ferrer, J., Mateo-Sanz, J.M.: Practical data-oriented microaggregation for statistical disclosure control. IEEE Transactions on Knowledge and Data Engineering 14(1), 189–201 (2002)
5. Domingo-Ferrer, J., Torra, V.: Ordinal, continuous and heterogeneous k-anonymity through microaggregation. Data Mining and Knowledge Discovery 11(2), 195–212 (2005)

6. Domingo-Ferrer, J., Solanas, A.: Erratum: Erratum to "a measure of variance for hierarchical nominal attributes". Inf. Sci. 179(20), 3732 (2009)
7. EFF. AOL's massive data leak. Electronic Frontier Foundation (2009), http://w2.eff.org/Privacy/AOL/
8. Gauch, S., Speretta, M.: Personalized search based on user search histories. In: Proc. of International Conference of Knowledge Management, CIKM 2004, pp. 622–628 (2004)
9. Google. 2008 annual report (December 2008), http://investor.google.com/order.html
10. Hansell, S.: Increasingly, internet's data trail leads to court. The New York Times (February 2006)
11. He, Y., Naughton, J.: Anonymization of set-valued data via top-down, local generalization. Proceedings of the VLDB Endowment 2(1), 934–945 (2009)
12. Hong, Y., He, X., Vaidya, J., Adam, N., Atluri, V.: Effective anonymization of query logs. In: CIKM 2009: Proceeding of the 18th ACM conference on Information and knowledge management, pp. 1465–1468 (2009)
13. Korolova, A., Kenthapadi, K., Mishra, N., Ntoulas, A.: Releasing search queries and clicks privately. In: WWW 2009: Proceedings of the 18th international conference on World wide web, pp. 171–180 (2009)
14. Miller, G.: WordNet - about us. WordNet. Princeton University (2009), http://wordnet.princeton.edu
15. Mills, E.: AOL sued over web search data release. CNET News (September 2006), http://news.cnet.com/8301-10784_3-6119218-7.html
16. Navarro-Arribas, G., Torra, V.: Tree-based microaggregation for the anonymization of search logs. In: WI-IAT 2009: Proceedings of the 2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology, pp. 155–158 (2009)
17. ODP. Open directory project (2010)
18. Oganian, A., Domingo-Ferrer, J.: On the complexity of optimal microaggregation for statistical disclosure control. Statistical Journal of the United Nations Economic Commision for Europe 18(4), 345–353 (2001)
19. Poblete, B., Spiliopoulou, M., Baeza-Yates, R.: Website privacy preservation for query log publishing. In: Bonchi, F., Ferrari, E., Malin, B., Saygın, Y. (eds.) PInKDD 2007. LNCS, vol. 4890, pp. 80–96. Springer, Heidelberg (2008)
20. Samarati, P.: Protecting respondents identities in microdata release. IEEE Transactions on Knowledge and Data Engineering 13(6), 1010–1027 (2001)
21. Summers, N.: Walking the cyberbeat. Newsweek (May 2009), http://www.newsweek.com/id/195621
22. Sweeney, L.: k-anonymity: a model for protecting privacy. International Journal on Uncertainty, Fuzziness and Knowledge-based Systems 10(5) (2002)
23. Torra, V.: Microaggregation for categorical variables: A median based approach. In: Domingo-Ferrer, J., Torra, V. (eds.) PSD 2004. LNCS, vol. 3050, pp. 162–174. Springer, Heidelberg (2004)
24. Torra, V.: Constrained microaggregation: Adding constraints for data editing. Transactions on Data Privacy 1(2), 86–104 (2008)
25. Ward, J.H.: Hierarchical grouping to optimize an objective function. Journal of the American Statistical Association 58(301), 236–244 (1963)
26. Zetter, K.: Yahoo issues takedown notice for spying price list. Wired (December 2009), http://www.wired.com/threatlevel/2009/12/yahoo-spy-prices/#more-11725

Data Environment Analysis and the Key Variable Mapping System

Mark Elliot, Susan Lomax, Elaine Mackey, and Kingsley Purdam

Centre for Centre and Survey Research, University of Manchester, Oxford Road, Manchester M13 9PL United Kingdom*
Mark.elliot@manchester.ac.uk

Abstract. It is now generally accepted that the measurement of statistical disclosure risk should be carried out with reference to the data environment into which a proposed dataset is to be released. This is normally considered through the development of intrusion or attack scenarios. Elliot and Dale's (1999) scheme set out a general set of principles for a scenario analysis, the output of which was a set of key variables. In this paper we outline an empirically based method, *Data Environment Analysis* which operationalises these ideas and a prototype tool the *Key Variable Mapping System* which has been designed to produce lists of key variables, with much more precise specification than was previously possible.

Keywords: Key Variables, Scenarios, Data Environment, Statistical Disclosure Risk.

1 Introduction

Following Paass (1990), Elliot and Dale (1999) and Purdam et al (2002, 2004) it is now generally accepted that the measurement of statistical disclosure risk should be carried out with reference to the data environment of a proposed data release. This is normally considered through the development of intrusion or attack scenarios.

This paper first considers the principles of disclosure risk assessment based upon scenarios and then describes how the Key Variable Mapping system enables the populating of scenario frames with well-specified key variables.

Elliot and Dale (1999) describe an 11-point system for analysing statistical disclosure scenarios. Taking a quasi-criminological view they express that first one must consider the means, motives and opportunity of a would-be *data intruder*. Only by considering why a would-be intruder would attack an anonymised dataset can we construct some measure of the prior likelihood of them make such an attempt. This reformulation came about through a consideration of Marsh et al's (1991) simple formulation of disclosure risk for samples of anonymised records:

$$p(\text{identification})=p(\text{identification}|\text{attempt}).p(\text{attempt}) \quad (1)$$

* The work reported in this paper was supported by the Office National Statistics as part of the Data Environment Analysis Service.

J. Domingo-Ferrer and E. Magkos (Eds.): PSD 2010, LNCS 6344, pp. 138–147, 2010

This formulation underlies the majority of attempts to model disclosure risk for individual level microdata. However, the focus has been on the conditional part. In other words p(attempt) has never been seriously modelled (not least because it is difficult to see how one would go about it) and so implicitly has been considered to equal 1.

Underlying Marsh et al's (1991) formulation is an implicit assumption that a data intruder would proceed by linking (or matching) known information to the anonymised target dataset. There are two consequent assumptions here: (i) that the known information includes unique formal identifiers and (ii) that the known information includes some information which is also found on the target database – this information is usually referred to as the *key variables.* It is assumed that the linkage of the known information to the target dataset would be done through these key variables.

Elliot and Dale's (1991) scenario scheme set out a general set of principles for a scenario analysis the output from which was a set of such key variables. These were based on a mixture of rational analysis and *ad hoc* data collection. In this paper we set out a case for empirically based methods for producing key variables lists, which capture in far more breadth and depth the information available to potential data intruders and through a technique known as *key variable matching* which allows more principled key variable specification than was previously possible.

2 The Data Environment Analysis

Data Environment Analysis (DEA) is a unique approach developed at Manchester University with funding from the Office of National Statistics. The goal of DEA to investigate, catalogue, categorise, and document available data in identification databases (those which could be used to link to target anonymised datasets in order to inform disclosure scenarios for data release).

Prior to this work there has been no (other) formal mechanism, within SDC, which allows the identification and classification what additional, external, information might be utilizable by a would-be data intruder. This has meant that a key element of the scenario structure the *means of an attempt* (which centrally revolves around what key variables is potentially available to a would-be intruder), is based largely on informed guesswork. In the absence of such a formal method, we have hit a barrier preventing further development our understanding of how a disclosure might occur. Without such understanding well grounded scenarios are impossible and we need well grounded scenarios to produce reasonably accurate disclosure risk measures to avoid disclosure management decisions that are either too conservative or too liberal, (which can lead to (i) potentially risky data are released; (ii) valuable low-risk data are not released; (iii) data releases are of limited utility because of the damage caused by data protection methods).

Another way to look at this is that there are two overarching themes captured in prior conditions of any scenario analysis: (i) *is it likely*- which is assessed (using Elliot and Dale classification scheme) by considering an plausible intruder 'motivations' and 'opportunities for attack' and (ii) *is it possible, and if so how* - which is centrally assessed considering what additional information an intruder would require to successful identify and/or disclose new information about respondents from a data

release. It is this second element, the how a disclosure might occur, that provides the rationale to DEA work.

We have identified that within a given DEA cycle there are two phases: (1) to investigate, catalogue, categorise and document what additional information may be available to an intruder and (2) to develop well grounded disclosure risk scenarios.[1] We are presently working within the first phase which has involved developing: (i) a series of methods for capturing what data is potential available to an intruder; (ii) a practical tool for cataloguing and categorizing the data captured i.e. the Key Variable Mapping System (KVMS). We will talk about each of these in turn.

2.1 Data Environment Analysis Methods

Data environment analysis uses several different but interrelated methods. For each method the principle is to capture metadata which can inform disclosure risk decision making.

1. Form Field Analysis – this is the primary method for capturing information held on *restricted access databases* and involves obtaining and cataloguing the data entry fields and data use specifications of paper and online forms from companies, organizations, political groups, clubs, charities, government departments etc. Information is also collected on whether the organization is registered with the UK's Information Commission and what details are given on data sharing.
2. Public domain data harvesting – this process involves an intensive search of all data in a given environment available in principle to any person within a given population (such information might include: directories, professional registers, vital life events data, self published data - such as blogs, land registry information, estate agents house sale information, electoral registers, commercial datasets and so on). The principle here is to create "grand keys" which combine multiple sources of information under various assumptions (regarding effort and financial resources).
3. Web Searching – the researcher searches the web for online resources and entry points for services, for example online shopping. Often this requires entering data for a fictitious person into a web form up to the point where payment details are requested.
4. Consultation with commercial data suppliers - commercial data companies increasingly hold detailed individual level information. Such information is often imputed from consumer surveys and combined with census and administrative data.
5. Attack Resourcing Simulation – an independent researcher is given a set of search parameters to identify as much information in the public domain about one or more variables that are in a particular target data set, within a particular time-frame. It is worth noting here that this method is non-general and the result will depend on the calibrated skills and resources of the researcher. The results, therefore, are equivalent in value to those obtained from re-identification/matching studies (i.e. non-abstract but *ad hoc*). Nevertheless the method does provide a workable

[1] These phases are necessary iterative, since available information might inform scenario structure (bottom up) or pre determined goal structures might drive data collection (both by the intruder and the analyst).

way of parameterising the concept of *reasonable effort*, which is sometimes used as a legal term by data stewardship organizations as one way to deal with the apparently paradoxical nature of the interaction between the categorical nature of data protection and other, related, legislation and the quasi-scalar nature of real word disclosure risk.

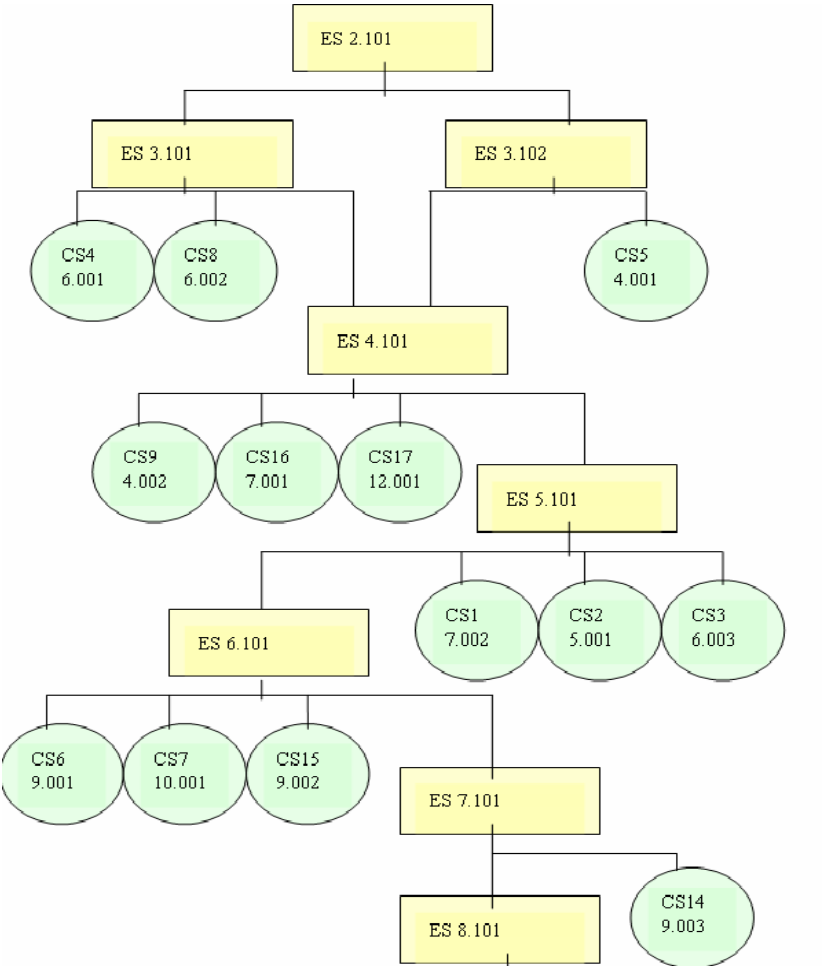
6. Security practices case studies - these involve establishing links with organizations holding individual level information and developing case studies of present and future plans for information gathering and data handling practices. In each organization formal interviews are conducted covering issues as: data provenance, data access and storage, anonymisation policies and practices, data quality and data linking.
7. Social network studies. Social networks can be defined by patterns of self disclosure; Elliot (2010) and therefore it is possible to use social network analysis to establish how much and what information individuals routinely know about other individuals within given social groups (work colleagues, neighbors etc.). This informs nosy neighbor type scenarios for example.

All of these methods of data collection are important, and different methods inform different scenarios and different aspects of those scenarios. However, the primary focus of our work to date has been form field analysis as it is this method that has enabled the structured development of the Key Variable Mapping System (KVMS) which is the subject of the second part of this paper.

Form field analysis works on the assumption that if a paper or electronic form asks for personal information then that information will be stored on a database of individual records (which could form an identification file in an attempt to attack another anonymised data file). The second assumption is that the data will be stored at the level of detail that it is collected. Given these two assumptions, it is plausible to infer that the forms provide direct metadata about the databases held by the organisations collecting information through them.

So, in form field analysis then the forms themselves are the raw data with each form providing information about the content of a database and that information itself becomes a record within a meta-database. The rows in such a database are the forms and the columns are possible variables and their codings. So, for example, each possible coding of the variable "ethnicity" will be represented as a column in the metadatabase. The various possible codings of a variable are not independent and can be arranged in graph structures such as that shown in figure 1 which shows part of an example graph for the variable employment status.

In the system used here, each coding has a simple classification code consisting of a set of letters followed by a number. The letters simply represent the construct captured by the variables in this case employment status is coded as ES. The numbers are in a pseudo decimal form. The digits before the decimal point refer to the number of categories in that coding. So a "2" here means a variable with two categories (such as "employed" and "not employed"). The first digit after the decimal point is a placeholder which indicates whether the node is actually captured data (0) or a harmonization coding (1). The final two digits are simply used to distinguish between coding for which the preceding information is identical. So ES3.101 and ES3.102 are two different harmonization codes with three categories.



KEY:
ES (boxes) represents the harmonized codes created for Employment Status necessary to link together two or more captured codes. CS (circles) represents the codes captured directly in the forms collected and catalogued in the database for Employment Status.
The number before the decimal point represents the number of categories in that coding. The number after the point completes the unique identifier for that coding.

Fig. 1. Example Taxonomic Graph for Employment Status Variable

The management of this meta-database is a complex process. To understand this consider that in response to each question on a new form there are three types of development that can happen:

1. *Assimilation to an existing structure* – here the question maps directly onto an existing coding and so the process is one of simply adding the form to that section of the database and indicate the code.

2. *Accommodation of an existing structure* – here the question does not directly map on to an existing code meaning that at least one new coding will have to be created. The new code will then have to be placed into the graph for that key variable. This can be extremely complex and can require the creation of harmonization codes which link the new coding to one or more existing codings.
3. *Creation of a new structure.* Occasionally a question is asked on a form which might result in the decision to create a new key variable.

With each new form the metadatabase develops and the overall structure, which in effect is a picture of the data environment, is enriched.

2.2 The Key Variable Mapping System (KVMS)

The Key Variable Mapping System (KVMS) has been produced to enable the production of precise key variable specifications. Its overall function is to generate a key variable set which is the metadata intersection of two data sets (or indeed classes of datasets).

KVMS compares codings of key variables across different datasets using a target dataset and either a single dataset or a summary of similar datasets. It is straightforward to add new target data sets of interest as they become available.

KVMS, written in Visual Basic, sits on top of the meta-database. The system works over a set of stored variable codings; as outlined in section 2, these codings are continually developed as new meta-data is added to the system. The codings are structured as a set of taxonomic graphs such as the one shown in Figure 1. The nodes in each graph are in two types:

1. observed data nodes – these correspond to coding systems actually used in collected meta-data.
2. harmonized data nodes which are codings produced by harmonizing two observed data nodes.

Sometimes, an observed data node can also be a harmonized data node. For example, because most external data sets collect Date of Birth information, the variable Age invariably harmonizes to whatever age coding is used on the target dataset.

When, asked to map two datasets, KVMS starts at the observed data nodes corresponding to the coding systems employed in both target and data environment datasets and then proceeds up the graph until it finds the node where the two paths to the top of the graph join. The join is the harmonized coding between the two datasets. It repeats that process for all variables. The full set of harmonized codes is then the key variable set for the pair of datasets.

An alternative and potentially more useful analysis is to use all of the meta-data for the set of forms, at present the system allows for two sets of classes: the sector in which the data is collected (e. g. banking or supermarket) and the purpose for which the data is collected (e.g. finance application). Expanding this to allow for other forms of classification would be relatively simple.

In order to use this set analysis, the user is required to enter a prevalence threshold value between 0 and 1, this indicates the proportion of datasets of a given classification that is required to have a given coding before it is considered. The system then

moves up the graph from the coding system of the target dataset until it arrives at a classification which meets the threshold. A value of zero means that if any form has a variable coding then it is considered. In effect, this will mean that the system will select as the key variable, the most detailed harmonization code between any dataset within the set and the target dataset. On the other hand a value of zero means that a coding will only be considered if all forms within the set either have that coding or harmonize to it.

The user interface to KVMS is shown in Figure 2. Once a user has run the system it returns a report which is in effect a detailed specification of the key variables. The Appendix shows an example of such a report produced with a mimic dataset.

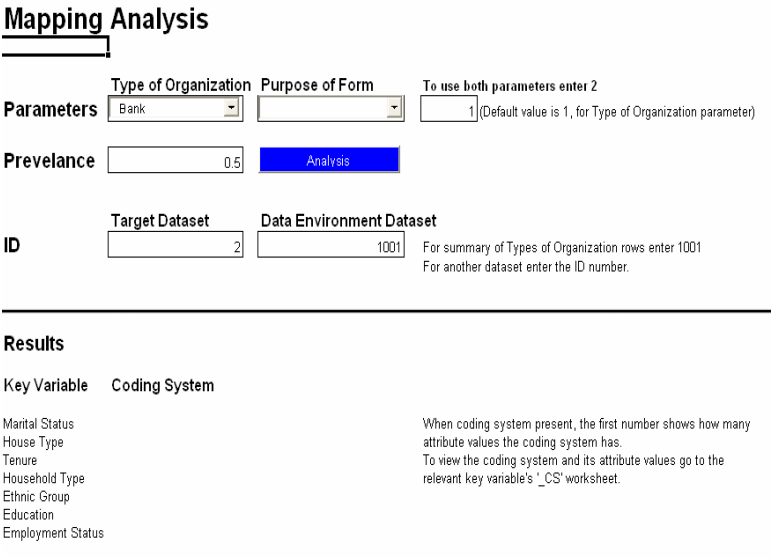


Fig. 2. Key Variable Mapping System Screenshot of the User Interface

3 Discussion

KVMS represents a step forward in our understanding of how key variables might be constructed by an intruder. The system has to date focused on form field analysis in order to drive forward the development of the system in a structured way. This has lead to a focus on database matching type scenarios.

In the next phase of the system's development we are gathering evidence on publicly available information which will allow a wider range of scenario types to be considered. Publicly available information can be differentiated in several ways - one three-way distinction is between involuntary (e.g. house sale information), participatory (e.g. telephone directories), and self-publication (blogs web sites) another key factor involves the resources required (mostly time and money) to obtain it. A further related class is the highly visible information such as type of housing. Different groupings of these classes of information and different assumptions about resource

investments will be required for and inform the development of a more subtle range of scenarios. However, in all cases the capacity of the Key Variable Mapping System to combine multiple data sources (through the prevalence threshold setting) ensures that it will always be able to drive the specification of the key variables that could conceivably be used.

It should be clear from the above that the development of and maintenance of the meta-database does itself require resources. However, given that inaccurate scenario specification could potentially introduce significant structural errors into risk assessments, it might be that such resource investment is worthwhile for National Statistical Institutes and other large organizations with data stewardship functions.

4 Summary

This paper sets out a case for, and outlines the principles of Data Environment Analyses (DEA) which amongst other things allows inferences to be made about the contents of databases for which no direct access is possible. The paper also describes a method - key variable mapping - which enables the production of more detailed and empirically grounded specifications of key variables for use in disclosure risk assessments.

In practical terms, the DEA methods outlined here could be employed by a National Statistical Institute who wishes to carry out their microdata release decisions and disclosure risk assessments based on contextual evidence.

References

Elliot, M., Dale, A.: Scenarios of Attack: A Data Intruder's Perspective on Statistical Disclosure Risk. Netherlands Official Statistics 14, 6–10 (1999)

Elliot, M.: Privacy, Confidentiality and Disclosure: conflicts and resolutions, Paper presented to Angela Dale's retirement Colloquium; Manchester (June 2010), http://www.ccsr.ac.uk/events/adc/index.html

Marsh, C., Skinner, C., Arber, S., Penhale, B., Openshaw, S., Hobcraft, J., Lievesley, D., Walford, N.: The case for samples of anonymized records from the 1991 census. Journal of the Royal Statistical Society series A 154, 305–340 (1991)

Paass, G.: Disclosure risk and disclosure avoidance for microdata. Journal of Business and Economic Statistics 6(4), 487–500 (1988)

Purdam, K., Elliot, M.J.: An evaluation of the availability of public data sources which could be used for identification purposes - A Europe wide perspective, CASC project report. University of Manchester, Manchester (2002)

Purdam, K., Mackey, E., Elliot, M.: The Regulation of the Personal. Policy Studies 25(4), 267–282 (2004)

Appendix: Example Report from KVMS

The text below is a mimic of the output of the key variable mapping process. The report described is a sector report which gives summary information across the whole of a sector, possibly sub-classified by the purpose for which the source data is collected (e.g. credit card application, customer loyalty scheme etc). The key variables identified will be those generated through the mapping of the sector results against the specified target database.

The prevalence threshold is in effect a risk parameter. If this number is set to zero this will provide the worst case and will return the key variable list which corresponds to the maximal list of variables of maximal detail collected anywhere in the sector (for the given purpose) – which is also available on the target dataset. If the number is set to 1 then this will provide the best case and will return the key variable list i.e. those collected by everyone in the sector (for the given purpose). Intermediate values will return intermediate lists.

The example given below shows the result for the hypothetical Widget industry against the hypothetical anonymised YYY survey dataset. In this case the prevalence threshold has been set to 0 which means that the resultant report lists the intersection of the variables available on the target dataset and the most detailed variables available anywhere in the sector.

Example Sector report for Widget Industry

The sector report queries take three parameters:

SECTOR: <a specification of the type of organization collecting the data>
PURPOSE: <a specification of the purpose of the source form (optional)>
PREVALENCE THRESHOLD: <a number between 0 and 1 representing the proportion of forms which have the information of a given type and level of detail.>

Exemplar Report

Report Run: 2nd July 2010
Target Database: 1990 YYY Survey
Sector: Widget
Purpose: Any
Inclusion Threshold: 0

Key variables (13)

Sex 2.101
1. Male
2. Female

Marital Status 4.101
1. Single
2. Married/living as couple/civil partnership
3. Divorced/separated
4. Widowed

Housing Type 4.101
1. Flat
2. Terraced
3. Detached
4. Other

Tenure 3.101
1. Own
2. Rent
3. Other

Using Support Vector Machines for Generating Synthetic Datasets

Jörg Drechsler

Institute for Employment Research, Regensburger Str. 104,
90478 Nuremberg, Germany

Abstract. Generating synthetic datasets is an innovative approach for data dissemination. Values at risk of disclosure or even the entire dataset are replaced with multiple draws from statistical models. The quality of the released data strongly depends on the ability of these models to capture important relationships found in the original data. Defining useful models for complex survey data can be difficult and cumbersome. One possible approach to reduce the modeling burden for data disseminating agencies is to rely on machine learning tools to reveal important relationships in the data.

This paper contains an initial investigation to evaluate whether support vector machines could be utilized to develop synthetic datasets. The application is limited to categorical data but extensions for continuous data should be straight forward. I briefly describe the concept of support vector machines and necessary adjustments for synthetic data generation. I evaluate the performance of the suggested algorithm using a real dataset, the IAB Establishment Panel. The results indicate that some data utility improvements might be achievable using support vector machines. However, these improvements come at the price of an increased disclosure risk compared to standard parametric modeling and more research is needed to find ways for reducing the risk. Some ideas for achieving this goal are provided in the discussion at the end of the paper.

Keywords: Support vector machines, Disclosure, Risk, Synthetic data, IAB Establishment Panel.

1 Introduction

Generating synthetic datasets is an innovative approach for statistical disclosure control. The basic idea that goes back to [28,21,12] is to develop a model for the joint distribution of the original data and then (repeatedly) sample from this distribution to generate synthetic data that could be released to the public without compromising the confidentiality of the survey respondents. While it is possible in theory to generate fully synthetic datasets [28] in which all values are completely synthetic, it is often sufficient to generate so called partially synthetic datasets [21] for which only the variables at risk – possibly only for a subset of respondents – are synthesized. Partially synthetic datasets enable the data disseminating agency to

J. Domingo-Ferrer and E. Magkos (Eds.): PSD 2010, LNCS 6344, pp. 148–161, 2010.

address the trade off between data utility and disclosure risk directly. On the one end, the release of the actual true data will obviously provide the highest possible data utility with a maximum of risk. On the other end, fully synthetic datasets will provide a maximum level of disclosure protection, but only the relationships captured in the model will be passed on to the synthetic data. If important relationships are missed or misspecified, results from the synthetic data can be misleading. With partial synthesis the agency can experiment with different amounts of synthetic data to find the optimal level of data utility and disclosure risk. For this reason, apart for some small experimental studies [9,26] only the partially synthetic approach has been implemented so far.

Recent research indicates that the partially synthetic data approach seems to be a promising alternative to standard statistical disclosure limitations especially for highly sensitive datasets for which the parameters that govern the level of perturbation for standard methods would have to be set to a high level. The approach has been successfully implemented for a product from the U.S. Census Bureau called OntheMap, illustrating commuting patterns, i.e. where people live and work, for the entire U.S. via maps available to the public on the web (`http://lehdmap.did.census.gov/`). The Census Bureau also protects the identities of people in group quarters (e.g., prisons, shelters) in the public use files of the American Communities Survey by replacing demographic data for people at high disclosure risk with imputations. Partially synthetic, public use datasets are in the development stage in the U.S. for the Longitudinal Business Database, the Longitudinal Employer-Household Dynamics survey, and the American Communities Survey veterans and full sample data. Outside the U.S. statistical agencies in Australia, Canada, Germany [8,11] and New Zealand [15,16] also are investigating the approach. The first partially synthetic dataset in Europe, a synthetic version of the IAB Establishment Panel, is scheduled to be released in 2010.

However, even for partially synthetic datasets the quality of the released data will strongly depend on the quality of the models that were used to generate the data. Defining useful models in a complex large scale survey is a labor intensive and very difficult task that can easily take several months and this is one of the reasons why many agencies are still reluctant to adopt the approach. In this context concepts from machine learning can be an attractive alternative for standard parametric modeling. The basic philosophy behind machine learning is to let the data decide, i.e. the relationships are derived from the data at hand and not from a superimposed model. [27] borrowed ideas from machine learning for the first time when he suggested using CART models to generate synthetic datasets. In the meantime this approach has been shown to be useful repeatedly [10,6]. Recently, [4] introduced random forests for synthesis and compared the results to standard CART models.

In this paper I investigate a machine learning concept that is fundamentally different from CART and random forests: support vector machines (SVM). SVMs use the predictors to find a hyperplane that separates the outcome of the response

variable. Some adjustments are necessary to obtain posterior probability models from SVMs to generate new data from the observed data and I will present them in the paper.

The remainder of the paper is organized as follows. In Section 2 I will present the main ideas behind support vector machines and the necessary adjustments for generating synthetic data. In Section 3 I will briefly illustrate the limits of parametric modeling using a toy dataset. In Section 4 I will compare the performance of SVMs to standard parametric modeling with a real data example based on the IAB Establishment Panel. Both data utility and disclosure risk will be evaluated. The paper concludes with some discussion on possible improvements for the method.

2 Support Vector Machines

The following introduction will necessarily be very brief. For more background information on the topic, I suggest [2,5,23]. Support vector machines introduced by [3] are a powerful tool for data classification. The basic concept was originally developed for predicting the outcome of a binary variable Y from a set of predictors X. However, the concept can easily be extended for categorical variables with more than two categories and can even be tuned for continuous variables (support vector regression). For the classification, the predictors are used to construct a hyperplane that "optimally" separates the different classes of the response variable. The aim is to select the hyperplane with the largest margin between the separated classes. To find this hyperplane the predictors might be mapped into a higher dimensional space using kernel functions.

It will not always be possible to find a mapping function for the predictor variables that will perfectly separate the outcome classes. To address this problem a particular loss function called the hinge loss function is introduced that is zero for every correctly classified point that is not inside the boundary of the margin and is linear otherwise. To keep the maximization problem tractable so called "slack" variables ξ_i, with $\xi_i \geq 0$ and $i = 1, ..., N$, are introduced that measure how inaccurate the classification is for each observation. Basically, they measure the distance of the reported value from the boundary of the separating hyperplane. By definition $\xi_i = 0$ for all observations that are correctly classified and outside the defined margin of the hyperplane.

Let Y be a binary variable, coded as $y_i \in \{-1, 1\}$. We can describe the maximization problem more formally as:

$$\max_{\beta, \beta_0, ||\beta||=1} \quad M \tag{1}$$

$$\begin{aligned} s.t. \qquad y_i(\beta_0 + \beta^T \Phi(x_i)) &\geq M(1 - \xi_i), \\ \xi_i &\geq 0, \quad i = 1, ..., N, \\ \sum_{i=1}^{N} \xi_i &\leq K, \end{aligned}$$

where M is the distance from the decision boundary to the margin boundary and Φ denotes some mapping of the original predictors into a possibly high dimensional feature space. The constant K can be seen as a tuning parameter. Note that as we allow K to increase, we allow the number of misclassifications to increase. It might seem to be advantage to set K as small as possible to keep the number of misclassifications low. However the necessary relationships for the link between the predictors and the outcome might be so complex that overfitting will become a serious issue, i.e. the found solution might describe the relation between the outcome and the predictors perfectly for the data at hand, but the results are not generalizable and thus are not useful for predicting unknown outcomes for a new set of observations. Selecting an appropriate K thus is a key element for support vector machines.

Without loss of generality we can rescale $|\beta_0 + \beta^T \Phi(x_i)| = 1$ for those points in each class nearest to the hyperplane. Note that the distance of the nearest point to the hyperplane in each class is now given by $1/||\beta||$. To maximize the margin we can restate (1) as a minimization problem concerning $||\beta||$ to obtain the notation found in most literature on support vector machines:

$$\min_{\beta,\beta_0,\xi} \quad \frac{1}{2}||\beta||^2 + C\sum_{i=1}^{N}\xi_i, \tag{2}$$

$$\begin{aligned} s.t. \qquad & y_i(\beta_0 + \beta^T\Phi(x_i)) \geq (1-\xi_i), \quad i = 1, ..., N \\ & \xi_i \geq 0, \quad i = 1, ..., N, \end{aligned}$$

with C defining the tuning parameter to balance the bias-variance trade-off.

Just like for most machine learning procedures, the tuning is performed by splitting the dataset for which both, the predictors X and the response Y are observed, into a training and a test dataset. The SVM is run on the training set with different tuning parameter combinations and each time the performance of the algorithm is evaluated by measuring how well it predicts the known classes of Y in the test dataset. The combination of tuning parameters that provides the best results on the test data is then used, if the SVM is applied to a new dataset for which Y is not observed.

Some adjustments are necessary before we can use support vector machines for generating synthetic datasets. Most importantly, it has been repeatedly pointed out that the output of the SVM can not be turned into posterior probability distributions for $P(Y|X)$ directly. As [1] state, the minimization of the hinge loss function "cannot correspond to fitting a probability model, since [the hinge loss] is indifferent to distinct values of the class probability." However, to generate synthetic datasets that can provide statistically valid inferences, we need to repeatedly draw from the posterior probability distributions to reflect the imputation uncertainty properly. Luckily, a number of authors [29,30,24,20,31] developed alternative loss functions for SVMs for which the necessary posterior probabilities can be derived. Describing them here would be beyond the scope

of this paper. In my application I rely on the methods developed by [31] and I refer to the paper for a detailed description of the derivations.

Secondly, overfitting is not as much an issue when generating synthetic datasets as it is for standard forecasting. In some sense, the whole data is already observed and we do not want to make any out of sample predictions. For that reason, I suggest to use a subset of the data to train the SVM just like in the standard setting, but then to evaluate the performance on the complete dataset and not only on the remaining test data. I found in extensive simulations not reported here for brevity that the tuning parameters found using this approach will generally lead to better results in terms of data utility as defined in Section 4.2 for the synthetic datasets.

Finally, the usual performance evaluation is based on how often the SVM predicts the correct class for Y in the test dataset. I found that this does not necessarily relate to the selection of the tuning parameters with "optimal" posterior probabilities. Sometimes a parameter combination is selected that poorly differentiates between the outcome classes of Y, i.e. for a binary variable, the probabilities are close to 0.5 almost for all records. To avoid these problems, I defined a different optimal performance criterion as:

$$\max \quad \frac{1}{N}\sum_{i=1}^{N}(Var(P(y_i|X))), \tag{3}$$

i.e. I calculate for each observation the variance between the posterior probabilities and average over all records. The parameter combination that maximizes this average variance is selected. In some sense this approach searches for the solution that places the highest confidence in the found classification. Again simulations not shown here indicate that using this evaluation criterion generally will lead to better results in terms of data utility for the synthetic datasets.

3 Illustrative Simulation

To illustrate that it might be a good idea in some situations to refrain from using parametric modeling for generating synthetic datasets I use a simple simulation. I generate a population of N=1,000,000 records consisting of three variables (Y, X_1, X_2), where $X = (X_1, X_2)$ are drawn from a bivariate standard normal distribution with zero mean and a correlation of 0.6. Y is a binary variable that is related to X through the following link:

$$Y \quad = \begin{cases} 1 \text{ if } P(logit^{-1}(-0.2x_1 - 0.2x_2 + 0.02x_1x_2)) > 0.5 \\ 0 \qquad\qquad\qquad\qquad\qquad\quad \text{else} \end{cases} \tag{4}$$

From this population I repeatedly draw simple random samples of size n=2,000. I treat Y as the sensitive variable and generate 10 synthetic datasets always replacing Y for all records using different imputation models. The first model (logit.correct) is the correct imputation model using a logistic regression including main effects and interactions. The second model (logit.wrong) is based on a logistic regression

Table 1. Simulation results

	Intercept	X_1	X_2	X_1X_2
true parameters	0.003	-0.200	-0.203	0.022
original sample	0.002(0.941)	-0.203(0.958)	-0.201(0.951)	0.024(0.952)
logit.correct	0.003(0.945)	-0.204(0.951)	-0.202(0.947)	0.024(0.953)
logit.wrong	0.004(0.949)	-0.321(0.528)	0.001(0.000)	0.022(0.961)
svm	0.014(0.974)	-0.196(0.985)	-0.196(0.982)	0.000(1.000)

including only X_1 and interactions. The last model (svm) is based on support vector machines. I assume the analyst is interested in the effect of X on Y so the analyst's model again is a logistic regression of Y on X using main effects and interactions. The analysis is performed using the original sample and the different generated synthetic samples respectively. I replicate the simulation 1,000 times. The results are presented in Table 1. The number in brackets reports the nominal coverage rate for a 95% confidence interval constructed from the generated datasets, i.e. how often the confidence interval contained the true parameter across the 1,000 replications. The reported point estimates for the different samples are the averages across the 1,000 simulation runs.

As expected, we get unbiased results for the original sample and the correct imputation model. The results for the logit imputation that excluded X_2 from the model are severely biased with only 52.8% coverage for X_1 and 0% coverage for X_2. This is a direct result of the uncongeniality [22] between the imputation model and the analysis model. Uncongeniality refers to the situation when the model used by the analyst of the data differs from the model used for the imputation. This can lead to biased results, if the analyst's model is more complex than the imputation model and the imputation model omitted important relationships present in the original data as is the case in my small toy example in which the imputation model basically assumes no direct relationship between Y and X_2.

The results for the SVM approach are somewhere in the middle. We can see that the results are a little biased especially for the intercept and for the interaction effect but generally the bias tends to be small. The coverage rates are all above 0.95. This is a result of the fact that the variance estimator for partially synthetic datasets that combines the variance within each dataset and the variance between the dataset (see for example [25]) tends to be conservative for the SVM approach. Explaining this phenomenon is an area of future research. Still, most researchers would agree that the results obtainable with the SVM approach are preferable to the results using the misspecified parametric model.

Of course, in this toy example it would have been easy to model the data correctly. Since it is the general advise for imputation modeling to always include as much information in the model as possible, any reasonable imputation model would have included main effects and interactions here. But in reality we have to deal with large scale complex surveys and including all variables with possible interactions is often not an option. So a decision has to be made, which effects

should be included and if the selected model omits important relationships between the variables found in the original data biased results are likely. In this case, it might be preferable to use a non parametric alternative that generally will introduce some bias, but might keep that bias at an acceptable level for any kind of analysis performed with the generated data.

4 Empirical Data Evaluation

In this section I illustrate that it is actually possible to obtain synthetic datasets with higher data utility from the SVM approach compared to results achievable using parametric modeling. The presented results are based on the IAB Establishment Panel, a large scale establishment survey conducted by the German Institute for Employment Research, so a short introduction to that dataset should prelude this section.

4.1 The IAB Establishment Panel

The IAB Establishment Panel is based on the German employment register aggregated via the establishment number as of 30 June of each year. The basis of the register, the German Social Security Data (GSSD) is the integrated notification procedure for the health, pension and unemployment insurances, which was introduced in January 1973. This procedure requires employers to notify the social security agencies about all employees covered by social security. As by definition the German Social Security Data only include employees covered by social security - civil servants and unpaid family workers for example are not included - approx. 80% of the German workforce are represented. However, the degree of coverage varies considerably across the occupations and the industries.

Since the register only contains information on employees covered by social security, the panel includes establishments with at least one employee covered by social security. The sample is drawn using a stratified sampling design. The stratification cells are defined by ten classes for the size of the establishment, 16 classes for the region, and 17 classes for the industry. These cells are also used for weighting and extrapolation of the sample. The survey is conducted by interviewers from TNS Infratest Sozialforschung. For the first wave, 4,265 establishments were interviewed in West Germany in the third quarter of 1993. Since then the Establishment Panel has been conducted annually - since 1996 with over 4,700 establishments in East Germany in addition. In the wave 2007 more than 15,000 establishments participated in the survey. Each year, the panel is accompanied by supplementary samples and follow-up samples to include new or reviving establishments and to compensate for panel mortality. The list of questions contains detailed information about the firms' personnel structure, development and personnel policy. For a detailed description of the dataset I refer to [13] or [19]. For the simulations I use one

dataset with all missing values imputed. I treat all imputed values like originally observed values for simplicity. See [7] for a description of the multiple imputation of the missing values in the survey.

4.2 Data Utility Evaluation

To compare the data utility achievable with the SVM approach to the standard parametric modeling, I use two regression analyses that have been previously used to evaluate the data quality of synthetic datasets generated from the IAB Establishment Panel ([6]). Both were suggested by colleagues at the IAB, who regularly use the survey for applied analyses. For the first regression (*regression*1), the binary dependent variable indicates if an establishment employs part-time employees. The 20 explanatory variables include among others dummies for the establishment size, whether the establishment expects changes in the number of employees, and information on the personnel structure. Industry dummies are included as covariates. The second regression (*regression*2) is an ordered probit regression with the expected employment trend in three categories (increase, no change, decrease) as the dependent variable. In that regression, I use 38 explanatory variables and the industry dummies as covariates. All results are computed for West Germany only.

For both regressions, I synthesized only the dependent variable generating 10 synthetic versions, once using the SVM approach and once using standard parametric modeling. For the SVM approach I followed the suggestions in [17] and used a radial basis function kernel $K(x_i, x_j) = \exp(-\gamma||x_i - x_j||^2)$ finding the optimal set of tuning parameter values γ and C by searching over the grid spanned by $\gamma = (2^{-15}, 2^{-13}, ..., 2^4)$ and $C = (2^{-5}, 2^{-3}, ..., 2^{15})$. To speed up the tuning, I split the dataset into 5 subsets defined by 5 quantiles based on the total gross wages paid by the establishment. For the parametric approach I used a logit model including all covariates not subject to any skip patterns to synthesize the dependent variable for *regression*1. To synthesize the dependent variable for *regression*2, I used a multinomial logit imputation model. In that model the number of explanatory variables had to be limited to 30 variables found by stepwise regression because of multicollinearity problems. Figures 1 and 2 present the results for the two regressions respectively. The upper two graphs present the plots of the point estimates from the regressions using the synthesized dependent variable plotted against the point estimates from the regressions using only the original data. The lower two graphs present box plots of the overlap of the 95% confidence intervals as suggested by [18]. This data utility measure can be computed as follows: For any estimand, first compute the 95% confidence intervals for the estimand from the synthetic data, (L_s, U_s), and from the collected data, (L_o, U_o). Then, compute the intersection of these two intervals, (L_i, U_i). The utility measure is

$$I = \frac{U_i - L_i}{2(U_o - L_o)} + \frac{U_i - L_i}{2(U_s - L_s)}. \tag{5}$$

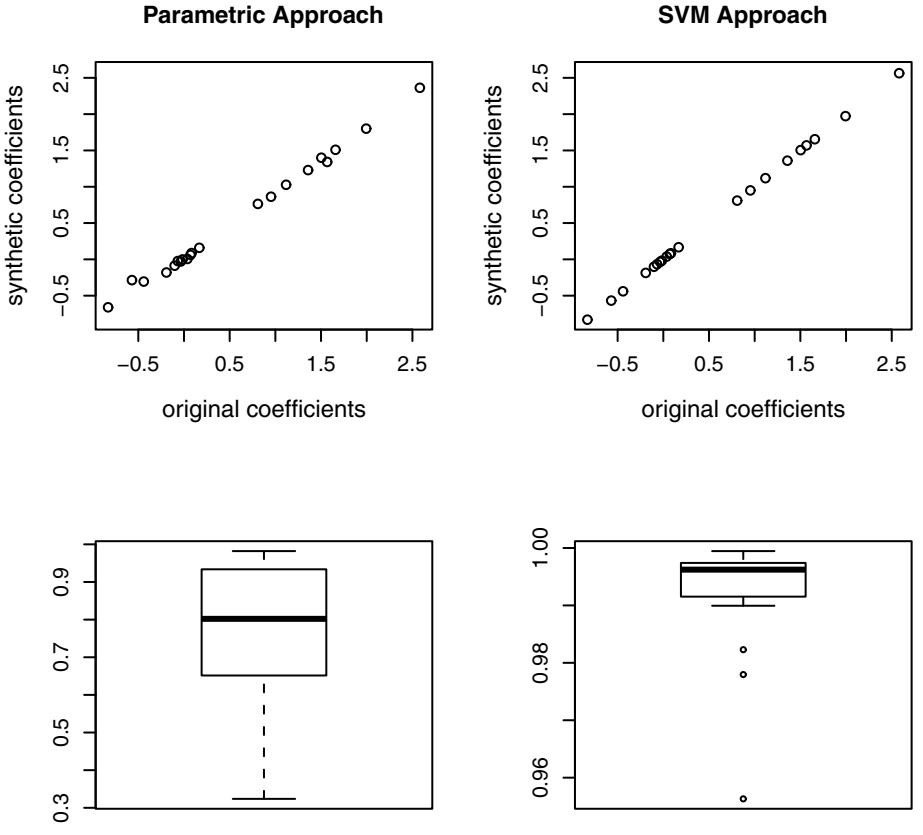


Fig. 1. Comparison of the point estimates and boxplots of the confidence interval overlap for the logit regression analysis

When the intervals are nearly identical, corresponding to high utility, $I \approx 1$. When the intervals do not overlap, corresponding to low utility, $I = 0$.

Obviously the SVM approach leads to higher data utility for both regressions. The synthetic point estimates are closer to their counterparts from the original data and the median confidence interval overlap is 0.996 (0.901) for the SVM approach compared to 0.802 (0.696) for the parametric approach for $regression1$ ($regression2$).

However, good data utility is only one side of the medal. The primary goal of any statistical disclosure limitation technique is to protect the data sufficiently. Therefore it is at least equally important to evaluate the level of protection offered by the two approaches.

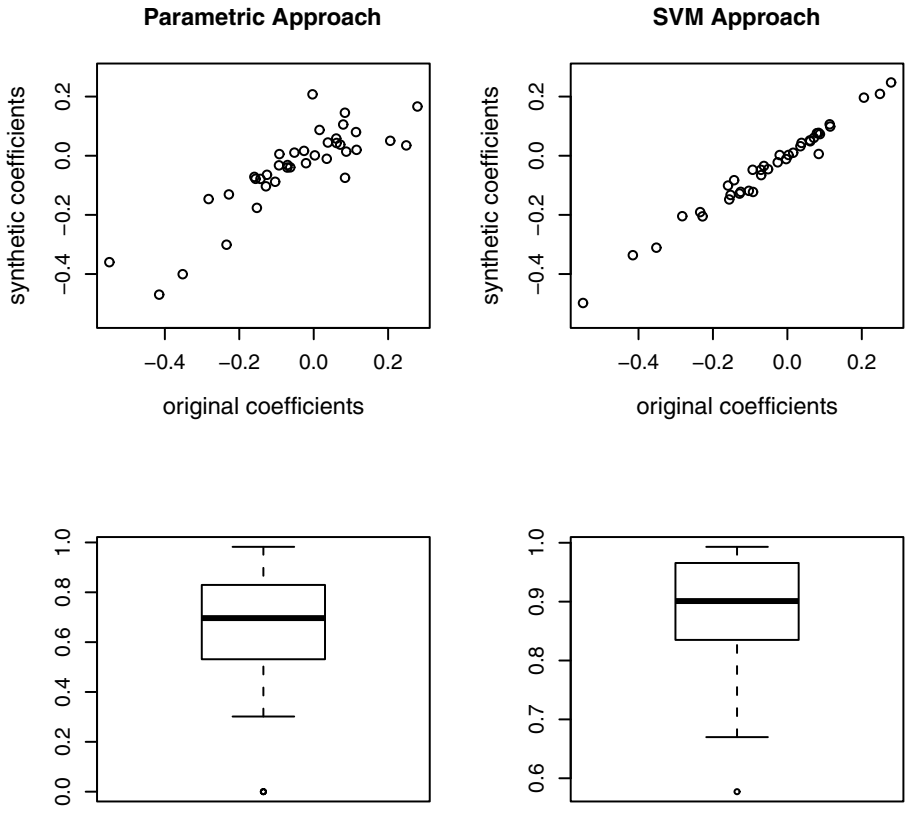


Fig. 2. Comparison of the point estimates and boxplots of the confidence interval overlap for the multinomial regression analysis

4.3 Disclosure Risk Evaluation

Since I only synthesized one variable in each of the simulations, it is difficult to come up with a realistic disclosure scenario. Releasing the dataset with just one variable synthesized would definitely not be an option. However, the main aim of this paper is to compare the two methods so it should be sufficient to compare the risks on a relative scale. I use two simple diagnostics to evaluate the risk from the two approaches. The first represents the percentage of records for which the mode across the synthetic responses equals the true response:

$$DR_1 = \frac{1}{N}\sum_{i=1}^{N} I(mode(y_{syn}^{(i,j)}) = y_{org}^{(i)}), \qquad i = 1, ..., N, \quad j = 1, ..., m, \qquad (6)$$

where $I(\cdot)$ is the indicator function, m is the number of imputations and $N = 15,644$ is number of records in the dataset. The second measure reports the

Table 2. Disclosure risk evaluations

		DR_1	DR_2
regression1	logit imputation	0.905	9.362
	SVM imputation	1.000	9.997
regression2	multinomial logit imp.	0.681	7.423
	SVM imputation	0.990	9.103

average number of times the true response is imputed across the 10 imputations for the records found by measure one.

$$DR_2 = \frac{1}{N_R}\sum_{i \in R} \#(y_{syn}^{(i,j)} = y_{org}^{(i)}), \qquad i = 1, ..., N, \quad j = 1, ..., m \tag{7}$$

where R is the set of records for which the mode of the synthetic records is equal to the true value and N_R is the total number of records in R. Note that DR_1 is bounded between 0 and 1, whereas DR_2 is bounded between 0 and the number of imputations (10 in this case).

The results for the two regressions are presented in Table 2. The relative "risks" for the SVM imputation are significantly larger than the ones for the parametric imputation especially for the second simulation. Obviously for the SVM approach this simple disclosure strategy would reveal the true reported value almost for all records in the dataset. On the other hand, knowing the reported value for a single binary variable (or a variable with three categories) will hardly identify a single respondent in the dataset. To evaluate the real risks, more variables would have to be synthesized to achieve a realistic data dissemination scenario. Then it would be possible to evaluate the risk of correctly identifying an individual record based on assumptions about external knowledge an intruder might use for re-identification purposes ([10]). These risks should be considerably lower. Nevertheless, the results clearly indicate the increased relative risk of the SVM approach compared to the parametric approach. Arguably the price in terms of increased risk is higher than the potential gains in data utility at least for this simulation.

The reason for this high risk for the SVM approach is probably a direct result of the way, support vector machines are searching for optimal solutions. High priority is given to sparse solutions, i.e. solutions for which a small number of support vectors are necessary to classify the data. Support vectors are the data points that drive the decisions for the classifiers. They are the points closest to the margin. All the other points do not have any influence on the classification. But this also means that the posterior probabilities derived for the data points that are not support vectors will always be close to one for the assigned category. As a result, drawing from these posterior probabilities likely will result in the same imputed value for most of the draws. The risk might be further increased by the tuning parameter selection approach defined in (3). This tuning approach

favors those tuning parameter combinations that lead to posterior probabilities $P(y_i|X)$ with high probabilities for one of the classes of y_i. This in turn will increase the probability that the same value is imputed in every imputation round and thus lead to an increase in DR_1 and DR_2.

5 Conclusions

Finding useful parametric imputation models for large scale surveys can be a challenging and time consuming task. One possible way of facilitating the search for useful models can be to rely on machine learning approaches that are based on the common idea of finding inherent structures in the data in a more or less automatic fashion by letting the data speak for themselves. Research on CART models ([27]) and random forests ([4]) already showed some promising results in this direction. In this paper I investigated, if the ideas behind support vector machines could be useful for generating synthetic datasets. The findings in the paper indicate that although some improvements in data utility might be possible with the approach, they might come at the price of an increased disclosure risk although the presented disclosure risk evaluations might be too simplified to allow a final statement. Clearly more research is needed in this area. The potentially increased risk is probably a direct result of the fact that support vector machines aim to use a limited number of observations for the actual classification to avoid overfitting. For this reason, many records will be assigned the same class across all imputations leading either to bad data quality if the classification does not provide consistent results or to a very high disclosure risk, if the classification is correct. Both are undesirable results. As I pointed out earlier, overfitting is not equally problematic in the context of synthetic data as it is in the context of forcasting. For that reason the results from the SVM approach might be improved if the search for sparse solutions could be relaxed. Along these lines it is reasonable to investigate different approaches for turning the SVM results into posterior probabilities than the ones I used in this paper. The approach I used explicitly tries to maintain the sparsity of the solution and I used it mainly for convenience reasons because it was readily available in R. Other approaches simply use some penalized regression to arrive at the posterior probabilities and this might actually be preferable in the synthetic data context. Given the promising results in terms of data utility, it would be interesting to see if alternative approaches to obtain the posterior probabilities from SVMs could lead to reduced risks of disclosure while maintaining the high data utility.

This paper can be seen as an initial investigation of the applicability of support vector machines for generating synthetic datasets. Besides the necessary extensions for continuous data, an important next step would be to compare this method to other machine learning approaches like CART or random forests that already have been demonstrated to work well as non parametric synthesizing tools.

Acknowledgments. This research was supported by grants from the German Research Foundation and the German Federal Ministry of Education and Research. I want to thank the two anonymous referees of this paper for their valuable comments which helped to improve the paper.

References

1. Bartlett, B., Jordan, M.I., McAuliffe, J.D.: Comment on: Moguerza, J.M. and Muñoz, A.: Support Vector Machines with Applications. Statistical Science (21), 341–345 (2006)
2. Berk, R.: Statistical Learning from a Regression Perspective. Springer, New York (2008)
3. Boser, B.E., Guyon, I., Vapnik, V.: A training algorithm for optimal marign classifiers. In: Proceedings of the Fifth ACM Workshop on Computation Learning Theory (COLT), pp. 144–152. ACM Press, New York (1992)
4. Caiola, G., Reiter, J.P.: Random Forests for Generating Partially Synthetic, Categorical Data. Transactions on Data Privacy 3, 27–42 (2010)
5. Cristianini, N., Shawe-Taylor, J.: Support Vector Machines and other kernel-based learning methods. Cambridge University Press, Cambridge (2000)
6. Drechsler, J.: Synthetic Datasets for the German IAB Establishment Panel. Working paper for the Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality (2009)
7. Drechsler, J.: Multiple imputation of missing values in the wave 2007 of the IAB Establishment Panel. IAB Discussion Paper (6) (2010)
8. Drechsler, J., Bender, S., Rässler, S.: Comparing fully and partially synthetic data sets for statistical disclosure control in the German IAB Establishment Panel. Transactions on Data Privacy 1, 105–130 (2008)
9. Drechsler, J., Dundler, A., Bender, S., Rässler, S., Zwick, T.: A new approach for disclosure control in the IAB Establishment Panel–Multiple imputation for a better data access. Advances in Statistical Analysis 92, 439–458 (2008)
10. Drechsler, J., Reiter, J.P.: Accounting for intruder uncertainty due to sampling when estimating identification disclosure risks in partially synthetic data. In: Domingo-Ferrer, J., Saygin, Y. (eds.) Privacy in Statistical Databases, pp. 227–238. Springer, Heidelberg (2008)
11. Drechsler, J., Reiter, J.P.: Disclosure risk and data utility for partially synthetic data: An empirical study using the German IAB Establishment Survey. Journal of Official Statistics 25, 589–603 (2009)
12. Fienberg, S.E.: A radical proposal for the provision of micro-data samples and the preservation of confidentiality. Tech. rep., Department of Statistics, Carnegie-Mellon University (1994)
13. Fischer, G., Janik, F., Müller, D., Schmucker, A.: The IAB Establishment Panel – from sample to survey to projection. Tech. rep., FDZ- Methodenreport No. 1 (2008)
14. Gomatam, S., Karr, A.F., Reiter, J.P., Sanil, A.P.: Data dissemination and disclosure limitation in a world without microdata: A risk-utility framework for remote access servers. Statistical Science 20, 163–177 (2005)
15. Graham, P., Penny, R.: Multiply imputed synthetic data files. Tech. rep., University of Otago (2005), http://www.uoc.otago.ac.nz/departments/pubhealth/pgrahpub.htm

16. Graham, P., Young, J., Penny, R.: Multiply imputed synthetic data: Evaluation of hierarchical bayesian imputation models. Journal of Official Statistics 25, 407–426 (2009)
17. Hsu, C.-W., Chang, C.-C., Lin, C.-J.: A Practical Guide to Support Vector Classification. Technical report, Department of Computer Science, National Taiwan University (2010)
18. Karr, A.F., Kohnen, C.N., Oganian, A., Reiter, J.P., Sanil, A.P.: A framework for evaluating the utility of data altered to protect confidentiality. The American Statistician 60, 224–232 (2006)
19. Kölling, A.: The IAB-Establishment Panel. Journal of Applied Social Science Studies 120, 291–300 (2000)
20. Lin, H.-T., Lin, C.-J., Weng, R.C.: A note on Platt's probabilistic outputs for support vector machines. Technical report, Department of Computer Science, National Taiwan University (2003)
21. Little, R.J.A.: Statistical analysis of masked data. Journal of Official Statistics 9, 407–426 (1993)
22. Meng, X.-L.: Multiple-imputation inferences with uncongenial sources of input (disc: P558-573). Statistical Science 9, 538–558 (1994)
23. Moguerza, J.M., Muñoz, A.: Support Vector Machines with Applications (with discussion). Statistical Science (21), 322–362 (2006)
24. Platt, J.: Probabilities for SV machines. In: Smola, A., Bartlett, P., Schölkopf, B., Schuurmans, D. (eds.) Advances in Large Margin Classifiers, pp. 61–74. MIT Press, Cambridge (2000)
25. Reiter, J.P.: Inference for partially synthetic, public use microdata sets. Survey Methodology 29, 181–189 (2003)
26. Reiter, J.P.: Releasing multiply-imputed, synthetic public use microdata: An illustration and empirical study. Journal of the Royal Statistical Society, Series A 168, 185–205 (2005)
27. Reiter, J.P.: Using CART to generate partially synthetic, public use microdata. Journal of Official Statistics 21, 441–462 (2005)
28. Rubin, D.B.: Discussion: Statistical disclosure limitation. Journal of Official Statistics 9, 462–468 (1993)
29. Wahba, G.: Multivariate function and operator estimation, based on smoothing splines and reproducing kernels. In: Casdagli, M., Eubank, S. (eds.) Proc. of Nonlinear Modeling and Forcasting, SFI Studies in the Science of Complexity, vol. XII, pp. 95–112. Addison-Wesley, Reading (1992)
30. Wahba, G.: Support vector machines, reproducing kernel hilpert spaces and the erndomized GACV. In: Schölkopf, B., Burges, C.J.C., Smola, A. (eds.) Advances in Kernel Methods – Support Vector Learning, pp. 69–88. MIT Press, Cambridge (1999)
31. Wu, T.-F., Lin, C.-J., Weng, R.C.: Probability estimates for multi-class classification by pairwise coupling. Journal of Machine Learning Research 5, 975–1005 (2004)

Synthetic Data for Small Area Estimation

Joseph W. Sakshaug and Trivellore E. Raghunathan

University of Michigan, Ann Arbor MI 48104, USA

Abstract. Increasingly, researchers are demanding greater access to microdata for small geographic areas to compute estimates that may affect policy decisions at local levels. Statistical agencies are prevented from releasing detailed geographical identifiers in public-use data sets due to privacy and confidentiality concerns. Existing procedures allow researchers access to restricted geographical information through a limited number of Research Data Centers (RDCs), but this method of data access is not convenient for all. An alternative approach is to release fully-synthetic, public-use microdata files that contain enough geographical details to permit small area estimation. We illustrate this method by using a Bayesian Hierarchical model to create synthetic data sets from the posterior predictive distribution. We evaluate the analytic validity of the synthetic data by comparing small area estimates obtained from the synthetic data with estimates obtained from the U.S. American Community Survey.

Keywords: Synthetic Data, Small Area Estimation, Disclosure, Microdata.

1 Introduction

The demand for greater access to microdata for counties, municipalities, neighborhoods, and other small geographic areas is ever increasing [1]. Analysts require such data to answer important research questions that affect policy decisions at local levels. Statistical agencies regularly collect data from small areas, but are prevented from releasing detailed geographic identifiers due to the risk of disclosing respondent identities and their sensitive attributes.

Existing data dissemination practices for small geographic areas include: 1) releasing summary tables containing aggregate-level data only; 2) suppressing geographical details in public-use microdata files for areas that do not meet a predefined population threshold (e.g., 100,000 persons) and; 3) permitting access to restricted geographical identifiers through a limited number of Research Data Centers (RDCs). Although useful in some situations, none of these methods is likely to satisfy the various needs of researchers, students, policy-makers, and community planners, who are fueling the demand for small area estimates.

This article investigates a fourth approach that statistical agencies may implement to release more detailed geographical information in public-use data sets. The approach builds on the statistical disclosure control method, originally proposed by Rubin [2], of creating multiple synthetic populations conditional on the observed data and releasing samples from each synthetic population which comprise the public-use

J. Domingo-Ferrer and E. Magkos (Eds.): PSD 2010, LNCS 6344, pp. 162–173, 2010

data files. Valid inferences on a variety of estimands are obtained by analyzing each data file separately and combining the results using methods described in [3].

The synthetic data literature focuses on preserving statistics about the entire sample, but preserving small area statistics is not addressed. Statistics about small areas can be extremely valuable to data users, but detailed geographical identifiers are almost always excluded from public-use microdata sets. Significant theoretical and practical research on model-based small area estimation has led to a greater understanding of how small area data can be summarized (and potentially simulated) by statistical models [4,5]. The majority of this research involves the use of Bayesian Hierarchical models, which are used to "borrow strength" across related areas and to increase the efficiency of the resulting small-area estimates. The use of Bayesian Hierarchical models for multi-level imputation, and, particularly, for synthetic data applications, is rare [20,21,22].

Under a fully-synthetic design all variables are synthesized and few (if any) observed data values are released. This design offers greater privacy and confidentiality protection compared to synthesizing only a subset of variables [10], but the analytic validity of inferences drawn from the synthetic data may be poor if important relationships are omitted or misspecified in the imputation model. A less extreme approach involves synthesizing a partial set of variables or records that are most vulnerable to disclosure [6,7,8,9]. If implemented properly, this approach yields high analytic validity as inferences are less sensitive to the specification of the imputation model, but it may not provide the same level of protection as fully-synthetic data because the observed sample units and the majority of their data values are released to the public [10].

At the present time, statistical agencies have only released partially synthetic data files [11,12,13]. There are worthy reasons why fully-synthetic data may be more appropriate for small area applications. The most important reason is that full synthesis can offer stronger levels of disclosure protection than partial synthesis. Data disseminators are obligated by law to prevent data disclosures and may face serious penalties if they fail to do so. Hence, maintaining high levels of privacy protection takes precedence over maintaining high levels of analytic validity. This point is particularly important for small geographic areas, which may contain sparse subpopulations and higher proportions of unique individuals who are especially susceptible to re-identification. A secondary benefit of creating fully-synthetic data sets is that an arbitrarily large sample size may be drawn from the synthetic population, facilitating analysis for data users who would otherwise have to exclude or apply complicated indirect estimation procedures to areas with sparse (or nonexistent) sample sizes. Synthetic sample sizes may be deliberately chosen to facilitate the use of direct estimation methods and standard statistical software and ease the burden of analysis for data users.

In this article, we propose an extension to existing synthetic data procedures for the purpose of creating synthetic, public-use microdata sets for small geographic areas from which valid small area inferences may be obtained. A Bayesian hierarchical model is developed that accounts for the hierarchical structure of the geographical areas and "borrows strength" across related geographic areas. A sequential multivariate regression procedure [14] is used to approximate the joint distribution of the observed data and to simulate synthetic values from the resulting posterior predictive

distribution. We demonstrate how statistical agencies may generate fully-synthetic data for small geographic areas on a subset of data from the U.S. American Community Survey. Synthetic data is generated for several commonly used household- and person-level variables and their analytic validity is evaluated by comparing small area inferences obtained from the synthetic data with those obtained from the observed data. We do not evaluate the disclosure risk properties of the proposed synthetic data approach and leave this to future work.

2 Review of Fully Synthetic Data

The general framework for creating and analyzing fully synthetic data sets is described in [3] and [15]. Suppose a sample of size *n* is drawn from a finite population $\Omega = (X, Y)$ of size *N*, with $X = (X_i;\ i = 1,2, \ldots, N)$ representing the design or geographical variables available on all *N* units in the population, and $Y = (Y_i;\ i = 1,2, \ldots, N)$ representing the survey variables of interest, observed only for the sampled units. Let $Y_{obs} = (Y_i; i = 1,2, \ldots, n)$ be the observed portion of *Y* corresponding to sampled units and $Y_{nobs} = (Y_i, i = n+1, n+2, \ldots, N)$ be the unobserved portion of *Y* corresponding to the nonsampled units. The observed data set is $D = \{X, Y_{obs}\}$. For simplicity, we assume there are no item missing data in the observed survey data set, but methods exist for handling this situation [16].

Fully synthetic data sets are constructed based on the observed data *D* in two steps. First, multiple synthetic populations are generated by simulating $\left(Y_{nobs}^{(l)}; l = 1,2, \ldots M\right)$ for the nonsampled units using independent draws from the Bayesian posterior predictive distribution, $(Y_{nobs}|X, Y_{obs})$, i.e., conditional on the observed data *D*. Alternatively, one can generate synthetic values of *Y* for all *N* units based on the posterior predictive distribution of "future" or "super" populations $\left(Y_f|D\right)$, conditional on the observed data. This procedure ensures that the synthetic populations contain no real values of *Y*, thereby avoiding the release of any observed value of *Y*. Second, a random sample (e.g., simple random sample) of size n_{syn} is drawn from each of *M* synthetic populations. These sampled units comprise the public-use data sets that are released to, and analyzed by, data users.

From these publicly-released synthetic data sets, data users can make inferences about a scalar population quantity $Q = Q(X, Y)$, such as the population mean of *Y* or the population regression coefficients of *Y* on *X*. In each synthetic data set, the user estimates *Q* with some point estimator *q* and an associated measure of uncertainty *v*. Let $\left(q^{(l)}, v^{(l)}; l = 1,2, \ldots, M\right)$ be the values of *q* and *v* computed on the *M* synthetic data sets. We assume that these quantities are estimated based on a simple random sampling design. Under assumptions described in [3], the data user can obtain valid inferences for scalar *Q* by combining the $q^{(l)}$ and $v^{(l)}$ using the following quantities:

$$\bar{q}_M = \sum_{l=1}^{M} q^{(l)}/M \tag{1}$$

$$b_M = \sum_{l=1}^{M} \left(q^{(l)} - \bar{q}_M\right)^2 / (M-1) \tag{2}$$

$$\bar{v}_M = \sum_{l=1}^{M} v^{(l)} / M \tag{3}$$

where $\bar{q}_M$ is used to estimate Q, and

$$T_M = (1 + M^{-1}) b_M - \bar{v}_M \tag{4}$$

is used to approximate the variance of $\bar{q}_M$. A disadvantage of T_M is that it can be negative. Negative values generally can be avoided by making M and n_{syn} large. A more precise variance estimator that is always positive is outlined in [3]. Inferences for scalar Q are based on a normal distribution when $T_M > 0$, n, M, and n_{syn} are large. For moderate M, inferences can be based on t-distributions [17].

3 Creation of Synthetic Data Sets for Small Area Estimation

We adopt a Bayesian approach, using a hierarchical imputation model, to generate synthetic data for small area estimation. Hierarchical models have been used in several applications of small area estimation [18,19]; see [5] for a comprehensive review of design-based, empirical Bayes, and fully Bayesian approaches for small area estimation. Hierarchical models have also been used for imputation of missing data in multilevel data structures [20,21].

Our approach involves three stages. In the first stage, incremental regression models are fit using the observed data within small areas to approximate the joint conditional density of the set of variables to be synthesized. In the second stage, the joint sampling distribution of population parameters is approximated and the between-area variation is modeled by incorporating covariates from larger geographical areas (e.g., states). In the final stage, the population parameters are simulated by taking independent draws from the posterior predictive distribution and are used to generate the synthetic microdata.

In illustrating the modeling steps, we take a pragmatic approach by keeping the models relatively simple from a computational perspective. Despite the simplified presentation, the basic framework can potentially handle more sophisticated modeling approaches on a routine basis. Limitations of our approach and alternatives are discussed in Section 5.

3.1 Stage 1: Direct Estimates

For descriptive purposes, we introduce the following notation. We define small areas as *counties*, nested within *states*, which could be nested within an even larger area (e.g., region). Specifically, suppose that a sample of size n is drawn from a finite population of size N. Let n_{cs} and N_{cs} denote the respective sample and population

sizes for county $c = 1,2,\ldots,C_s$ within state $s = 1,2,\ldots,S$. Let $Y_{cs} = \left(Y_{ics,p}; i = 1,2,\ldots,n_{cs};\ p = 1,2,\ldots,P\right)$ represent the $n_{cs} \times P$ matrix of survey variables collected from each survey respondent located in county *c* and state *s*. Let $X_{cs} = \left(X_{ics,j}; i = 1,2,\ldots,n_{cs}, n_{cs}+1,\ldots,N_{cs};\ j = 1,2,..,J\right)$ represent the $N_{cs} \times J$ matrix of auxiliary or administrative variables known for every population member in a particular county and state. Although we consider synthesis of the survey variables Y_{cs} only, it is straightforward to synthesize the auxiliary variables X_{cs} as well.

A desirable property of synthetic data is that the multivariate relationships between the observed variables are maintained in the synthetic data, i.e., the joint distribution of variables is preserved. The first task is to specify the joint conditional distribution of the observed county-level variables to be synthesized $\left(Y_{cs,1}, Y_{cs,2}, \ldots, Y_{cs,P} | X_{cs,j}\right)$, where synthetic values are drawn from a corresponding posterior predictive distribution. Specifying and simulating from the joint conditional distribution can be difficult for complex data structures involving large numbers of variables representing a variety of distributional forms. Alternatively, one can approximate the joint density as a product of conditional densities [14]. Drawing synthetic variables from the joint posterior density $\left(Y_{cs,1}, Y_{cs,2}, \ldots, Y_{cs,P} | X_{cs,j}\right)$ can be achieved by sampling from$\left(Y_{cs,1} | X_{cs,j}\right)$, $\left(Y_{cs,2} | Y_{cs,1}, X_{cs,j}\right)$,…,$\left(Y_{cs,P} | Y_{cs,1}, \ldots, Y_{cs,P-1}, X_{cs,j}\right)$. In practice, a sequence of generalized linear models is fit on the observed county-level data where the variable to be synthesized comprises the outcome variable and any auxiliary variables or previously fitted variables are used as predictors, e.g., $Y_{ics,p+1} = \left(X_{ics}, Y_{ics,1,\ldots,p}\right)\beta_{cs} + \varepsilon_{ics}$. The choice of model (e.g., Gaussian, binomial) is dependent on the type of variable to be synthesized (e.g., continuous, binary). It is assumed that any complex survey design features are incorporated into the general linear models and that each variable has been appropriately transformed, if needed, to satisfy linear regression assumptions. After fitting each conditional density, estimates of the population regression coefficients $\hat{\beta}_{cs}$, the corresponding covariance matrix $\hat{V}_{cs}$, and the residual variance $\hat{\sigma}^2_{cs}$ are obtained and incorporated into the hierarchical structure described below in Section 3.2.

3.2 Stage 2: Sampling Distribution and Between-Area Model

In the second stage of synthetic data creation, the joint sampling distribution of the design-based county-level regression estimates $\hat{\beta}_{cs}$ (obtained from each conditional density in Stage 1) is approximated by a multivariate normal distribution,

$$\hat{\beta}_{cs} \sim MVN\left(\beta_{cs}, \hat{V}_{cs}\right),$$

where β_{cs} is a $(J+p) \times 1$ matrix of population regression coefficients and $\hat{V}_{cs}$ is the $(J+p) \times (J+p)$ corresponding covariance matrix estimated from the first stage. Further, the county-level population parameters β_{cs} are assumed to follow a multivariate normal distribution,

$$\beta_{cs} \sim MVN(\beta Z_s, \Sigma),$$

where $Z_s = \left(Z_{s,k}; k = 1,2,\ldots,K\right)$ is a $K \times 1$ matrix of state-level covariates, β is a $(J+p) \times K$ matrix of population regression coefficients, and Σ is a $(J+p) \times (J+p)$ covariance matrix. State-level covariates are incorporated into the hierarchical model in order to "borrow strength" from related areas. Prior distributions may be assigned to the unknown parameters β and Σ, but for ease of presentation, we assume that β and Σ are fixed at their respective maximum likelihood estimates (MLE), a common assumption in hierarchical models for small area estimation [18,23,24].

3.3 Stage 3: Generating Synthetic Populations within Small Areas

The ultimate objective is to generate synthetic populations within a small area using an appropriate posterior distribution. To this end, one can simulate the unknown population regression parameters β_{cs} specified in the hierarchical model described in Section 3.2. Based on standard theory of the normal hierarchical model [25], the posterior predictive distribution of the population regression coefficients is,

$$\tilde{\beta}_{cs} \sim MVN\left[\left(\hat{V}_{cs}^{-1} + \hat{\Sigma}_{\text{MLE}}^{-1}\right)^{-1}\left(\hat{V}_{cs}^{-1}\hat{\beta}_{cs} + \hat{\Sigma}_{\text{MLE}}^{-1}\hat{\beta}_{MLE}Z_s\right), \left(\hat{V}_{cs}^{-1} + \hat{\Sigma}_{\text{MLE}}^{-1}\right)^{-1}\right],$$

where $\tilde{\beta}_{cs}$ is a simulated vector of values for the vector of population regression coefficients β_{cs} . Simulating a synthetic variable $\tilde{Y}_{cs}$ for observed variable Y_{cs} can then be achieved by drawing $\tilde{Y}_{cs}$ from a parametric distribution with location and scale parameters $X_{cs}\tilde{\beta}_{cs}$ and σ_{cs}^2 , respectively, where σ_{cs}^2 may be drawn from an appropriate posterior predictive distribution $(\sigma_{cs}^2|Y_{cs}, X_{cs})$, or the maximum likelihood estimate $\hat{\sigma}^2$ obtained from Section 3.1 may be used. For example, to simulate a normally distributed variable $Y_{cs,1}$ one can draw $\tilde{Y}_{cs,1}$ from the distribution $N\left(X_{cs}\tilde{\beta}_{cs}, \hat{\sigma}^2\right)$. Generating a second (normally distributed) synthetic variable $\tilde{Y}_{cs,2}$ from the posterior predictive distribution $\left(Y_{cs,2}|Y_{cs,1}, X_{cs}\right)$ is achieved by drawing $\tilde{Y}_{cs,2}$ from $N\left(X_{cs}^*\tilde{\beta}_{cs}, \hat{\sigma}^2\right)$, where $X_{cs}^* = \left(X_{cs}, \tilde{Y}_{cs,1}\right)$. If the second synthetic variable is binary, then $\tilde{Y}_{cs,2}$ is drawn from $Bin\left(1, \hat{p}\left(X_{cs}^*\tilde{\beta}_{cs}\right)\right)$, where $\hat{p}\left(X_{cs}^*\tilde{\beta}_{cs}\right)$ is the predicted probability computed from the inverse-logit of $X_{cs}^*\tilde{\beta}_{cs}$. For polytomous variables, the same procedure is adapted to obtain posterior probabilities for each categorical response and the synthetic values are sampled from a multinomial distribution. This iterative process continues until all synthetic variables $\left(\tilde{Y}_{cs,1}, \tilde{Y}_{cs,2}, \ldots, \tilde{Y}_{cs,P}\right)$ are generated. Multiple conditioning cycles can be implemented to minimize ordering effects [14]. The entire procedure is repeated M times to create multiple replicates of synthetic variables $\left(\tilde{Y}_{cs,1}^{(l)}, \tilde{Y}_{cs,2}^{(l)}, \ldots, \tilde{Y}_{cs,P}^{(l)}; l = 1,2,\ldots,M\right)$.

The entire synthetic populations may be disseminated to data users, or a simple random sample of arbitrary size may be drawn from each population and released. Stratified random sampling may be used if different sampling fractions are to be applied within the small areas. Inferences for a variety of small-area estimands Q_{cs} and large-area estimands Q_s or Q can be obtained using the combining rules in Section 2.

4 Evaluation of Synthetic Data for Small Area Inferences

In this section, we illustrate the above procedure on a subset of public-use microdata from the U.S. American Community Survey (ACS). We generate fully-synthetic data sets for relatively small geographic areas and evaluate the analytic validity of the resulting estimates. The data consist of seven household-level variables and eight person-level variables measured on 846,832 households and 2,093,525 persons during years 2005-2007. The variables, shown in Table 1, were chosen by researchers at the U.S. Census Bureau for this project. The smallest geographic unit that is identified in the ACS microdata is a Public-Use Microdata Area (PUMA). A PUMA is a census area containing around 100,000 people. All such areas are non-overlapping and are nested within a state. We restrict our sample to the Northeast region, which contains 9 states and 405 PUMAs.

We generate $M = 10$ fully-synthetic data sets for each "small area" (i.e., PUMA). To ensure that each synthetic data set contains ample numbers of households and/or persons within PUMAs, we create synthetic samples that are larger than the observed samples in each PUMA. Specifically, we generate synthetic sample sizes that are equivalent to 20% of the total number of U.S. households located within each PUMA based on the 2000 decennial census counts. This yielded a total synthetic sample size of 4,436,085 households for the Northeast region. Conceptually, this is equivalent to drawing a stratified random sample of households from each of $M = 10$ synthetic data populations.

The first survey variable to be synthesized is household size. Creating a household size variable will facilitate the creation of synthetic person-level variables in a later step. Because no administrative or other conditioning variables X_{cs} are available for

Table 1. List of ACS variables used in the synthetic data evaluation

Variable	Range/Categories
Household variables	
Household Size	1 - 20
Sampling weight	1 - 516
Total bedrooms	0 - 5
Electricity bill/mo.	1 - 600
Total rooms (excl. bedrooms)	1 - 7
Tenure	mortgage/loan, own free and clear, rent
Income	-33,998 – 2,158,100
Person variables	
Sampling weight	1 - 814
Gender	male, female
Education	16 categories, recoded less than high school, high school, some college, and college graduate
Ethnicity	Hispanic, non-Hispanic
Age	0 - 95
Race	9 categories, recoded white, black, other
Moved in last year	yes, no
Living in poverty	yes, no

this application, household size is simulated using a Bayesian Poisson-gamma model conditional on the observed household size variable with unknown hyperparameters estimated using maximum likelihood estimation. The remaining household-level variables are synthesized using the hierarchical modeling procedure described in Section 3. The sampling weights (both household and person) are included among the set of variables to be synthesized. State-level covariates Z_s, including population size (log-transformed), number of metropolitan, and number of micropolitan areas, are incorporated into the hierarchical model.

Linear regression models are used within PUMAs to obtain design-based estimates of population parameters for all numerical variables (with the previously noted exception of household size). Synthetic values of numerical variables are sampled from a Gaussian posterior predictive distribution. For binary variables, logistic regression models are used to obtain design-based population coefficients and corresponding synthetic values are sampled from a binomial posterior predictive distribution; the same procedure is applied to polytomous variables, which are broken up into a series of binary variables. To increase the stability of the design-based population regression coefficients, we apply a minimum sample size rule of $15 \cdot p$ within each PUMA. If a PUMA did not meet this minimum threshold, then nearby PUMAs were pooled together until the criterion was met.

Once the household variables were synthesized, the synthetic household data sets were transformed to person-level data sets and the person-level variables were synthesized conditional on the household variables. Taylor series linearization [26] was used to obtain design-based regression coefficients, accounting for the clustering of persons within households. To reduce the ordering effect of synthesizing the household variables first, we performed an additional conditioning cycle where each synthetic variable is conditioned on the full set of household- and person-level variables from the previous implementation.

4.1 Univariate Inferences for Small Areas

We evaluate the analytic validity of the synthetic data by comparing PUMA estimates obtained from the synthetic data with those obtained from the observed data for all 405 PUMAs. First, we compute basic univariate estimates, namely, means (or proportions) and standard errors for each PUMA; multivariate estimates are evaluated in Section 4.2. The sampling weights (synthetic and observed) are used to obtain adjusted point estimates and standard errors.

Table 2 presents the overall mean of the (weighted) PUMA means and standard errors obtained from the synthetic and observed data. The last column contains the slope (β_1) of the observed point estimates regressed against the synthetic point estimates for all 405 PUMAs. A slope equal to (or close to) 1 indicates a strong linear correspondence between the synthetic and observed estimates. On average, the synthetic PUMA means are generally within two standard errors of the observed PUMA means and the estimated slopes are reasonably close to the desired value, i.e., $(\beta_1 = 1)$. One exception is the Age variable, which is overestimated by the synthetic data. Upon inspection, the observed age variable has a bimodal distribution, which is not ideally simulated with a Gaussian distribution; this is a limitation of the parametric Bayesian simulation framework. Nonparametric strategies are likely to produce

Table 2. Mean of synthetic and observed PUMA means/proportions and standard errors and regression slope of actual means on the synthetic means for all 405 PUMAs

	Synthetic	Observed	
Variable	Mean (SE)	Mean (SE)	Slope (β_1)
Household variables			
Household Size	2.32 (0.03)	2.32 (0.03)	0.99
Sampling weight	33.70 (0.38)	33.96 (0.50)	0.99
Total bedrooms	2.70 (0.03)	2.66 (0.03)	1.02
Electricity bill/mo	115.53 (2.14)	114.19 (2.32)	1.04
Total rooms	3.05 (0.03)	2.99 (0.03)	1.01
Tenure			
Own free & clear	0.22 (0.01)	0.21 (0.01)	0.97
Rent	0.34 (0.01)	0.35 (0.01)	1.01
Mortgage/loan	0.44 (0.01)	0.44 (0.01)	0.98
Income	76144.90 (1576.65)	73658.80 (1780.44)	0.93
Person variables			
Sampling weight	35.80 (0.51)	35.49 (0.34)	0.98
Gender: male	0.49 (0.01)	0.49 (0.01)	0.57
Education			
Less than HS	0.33 (0.01)	0.33 (0.01)	0.99
HS graduate	0.24 (0.01)	0.24 (0.01)	0.98
Some college	0.20 (0.01)	0.20 (0.01)	0.95
College graduate	0.23 (0.01)	0.23 (0.01)	1.05
Hispanic ethnicity	0.12 (0.06)	0.11 (0.01)	1.16
Age	41.33 (0.34)	38.00 (0.38)	1.07
Race			
White	0.75 (0.01)	77.0 (0.01)	1.02
Black	0.13 (0.01)	0.11 (0.01)	1.03
Other	0.12 (0.01)	0.12 (0.01)	1.05
Moved in last year	0.11 (0.01)	0.12 (0.01)	1.13
Living in poverty	0.12 (0.01)	0.12 (0.01)	1.04

more desirable results. The Gender variable yields a relatively low slope value due to small variations in observed proportions of males across PUMAs. Overall, we believe the quality of the synthetic estimates is good relative to the observed data for obtaining univariate estimates. Aggregating the synthetic data to the state- and region-levels yielded estimates with similar correspondence to the observed data (not shown), indicating that synthetic data may be useful for producing valid estimates across multiple levels of geography.

4.2 Multivariate Inferences for Small Areas

Next we evaluate the analytic validity of the synthetic data for multivariate estimates. Table 3 presents summary results of two multiple regression models fitted within each PUMA. The first model regresses household income on the remaining household-level variables, and the second model regresses a recoded binary variable indicating college attendance (some college/college degree vs. less than high school/high school graduate) against all other person-level variables. Pseudo-maximum likelihood

Table 3. Mean of synthetic and observed PUMA regression coefficients and standard errors and regression slope of actual coefficients on the synthetic coefficients for all 405 PUMAs

	Synthetic	Observed	
	Coef (SE)	Coef (SE)	Slope (β_1)
Regression coefficients of household income (cube root):	25.37 (1.24)	26.05 (1.53)	0.93
Intercept	1.64 (0.19)	1.52 (0.24)	0.90
Household Size	1.27 (0.28)	1.20 (0.35)	0.98
Total bedrooms	0.94 (0.20)	0.89 (0.26)	0.98
Electricity bill/mo. (cube root)	1.34 (0.24)	1.35 (0.29)	0.99
Total rooms (excl. bedrooms)			
Tenure			
Own free & clear	-3.99 (0.61)	-4.13 (0.79)	1.08
Rent	-5.97 (0.70)	-6.15 (0.85)	0.97
Regression coefficients of college attendance on:			
Intercept	-0.70 (0.08)	-1.07 (0.09)	1.17
Gender: male	-0.09 (0.06)	-0.07 (0.08)	0.97
Hispanic ethnicity	-0.62 (0.15)	-0.53 (0.24)	1.07
Age	0.01 (0.01)	0.02 (0.01)	1.36
Race			
Black	-0.30 (0.14)	-0.18 (0.27)	1.01
Other	0.01 (0.14)	0.03 (0.17)	1.01
Poverty	-0.73 (0.11)	-0.80 (0.17)	0.87
Moved in last year	0.57 (0.13)	0.45 (0.14)	0.82

estimation is used to incorporate the relevant sampling weights [27]. The summary measures shown in Table 3 consist of overall means of the estimated regression coefficients and corresponding standard errors obtained from each PUMA. The last column represents the slope of the observed point estimates regressed against the synthetic point estimates for all 405 PUMAs. Because these models resemble those used earlier to approximate the joint posterior distribution of county-level parameters, we should expect close correspondence between the synthetic and observed point estimates, and more efficient synthetic data estimates. Indeed, we find that, on average, the synthetic point estimates correspond well with the observed point estimates in both direction and magnitude. The synthetic point estimates lie within about two standard errors of the observed point estimates, on average, and are generally more efficient than the observed point estimates. We find similar correspondence between the synthetic and observed data estimates when the data are aggregated to higher levels of geography (e.g., states, region).

5 Conclusions

This paper addresses an important data dissemination issue facing statistical agencies, which is how to meet the growing demand for high quality, public-use microdata for small geographic areas while protecting data confidentiality and privacy of respondents. These competing aims are likely to garner even more attention in the future as

research into small area effects and societal sensitivity to privacy and confidentiality continues to grow.

We propose a fully-synthetic data approach that utilizes a hierarchical model for creation of microdata for small geographic areas. The resulting data sets could conceivably be released to the public, along with additional data products that contain finer levels of data than are currently being released. The methodology is flexible, easy to implement, and can be straightforwardly adapted to a variety of data sources representing various geographical structures and variable types.

Results of the empirical evaluation suggest that valid small area inferences can be obtained from fully-synthetic data for basic descriptive and multivariate estimands. Although PUMAs are generally not considered to be "small areas," small-scale evaluations of the proposed synthetic data method using county-level data in the Census Research Data Center has yielded similarly valid results, with a slight loss in efficiency for the smallest areas.

One issue that was not empirically addressed in this paper is the level of disclosure protection offered by the synthetic data for small areas. Although there is evidence that fully-synthetic data offers better protection against disclosure than partially-synthetic data [10], this may not be true for small geographic areas or sparse subpopulations. Further research is needed to determine whether fully-synthetic data offers adequate levels of disclosure protection to be suitable for public release in a small area context.

In the evaluation, we did not assess the validity of the synthetic data for obtaining subgroup estimates or modeling interactions. Such estimates are particularly important to researchers studying subpopulations segregated within small geographic areas. Current work is underway to build complex relationships and interactions into the synthetic data generation process. An area for future work is the development of easy-to-implement, nonparametric approaches that weaken the dependence of the synthetic data inferences on the specification of the imputation models. In addition, further evaluations of the repeated sampling properties of the resulting synthetic data are needed to assess confidence interval coverage and data utility.

Acknowledgments. This research was supported by grants from the U.S. Census Bureau (YA-132309SE0354) and the U.S. National Science Foundation (SES-0918942).

References

1. Tranmer, M., Pickles, A., Fieldhouse, E., Elliot, M., Dale, A., Brown, M., Martin, D., Steel, D., Gardiner, C.: The case for small area microdata. J. Roy. Stat. Soc. A 168, 29–49 (2005)
2. Rubin, D.B.: Satisfying confidentiality constraints through the use of synthetic multiply-imputed microdata. J. Off. Stat. 9, 461–468 (1993)
3. Raghunathan, T.E., Reiter, J.P., Rubin, D.B.: Multiple imputation for statistical disclosure limitation. J. Off. Stat. 19, 1–16 (2003)
4. Platek, R., Rao, J.N.K., Sarndal, C.E., Singh, M.P.: Small area statistics. Wiley, New York (1987)
5. Rao, J.N.K.: Small Area Estimation. Wiley, New York (2003)

6. Little, R.J.A.: Statistical analysis of masked data. J. Off. Stat. 9, 407–426 (1993)
7. 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, pp. 248–267. National Academy Press, Washington DC (1997)
8. Liu, F., Little, R.J.A.: Selective multiple imputation of keys for statistical disclosure control in microdata. In: Proceedings of the Joint Statistical Meetings, pp. 2133–2138. American Statistical Association, Blacksburg (2002)
9. Reiter, J.P.: Inference for partially synthetic public use microdata sets. Surv. Methodol. 29, 181–188 (2003)
10. Drechsler, J., Bender, S., Raessler, S.: Comparing fully and partially synthetic datasets for statistical disclosure control in the German IAB establishment panel. Trans. Data Priv. 1(3), 105–130 (2008)
11. Rodriguez, R.: Synthetic data disclosure control for American Community Survey group quarters. In: Proceedings of the Joint Statistical Meetings, pp. 1439–1450. American Statistical Association, Salt Lake City (2007)
12. Abowd, J.M., Stinson, M., Benedetto, G.: Final report to the Social Security Administration on the SIPP/SSA/IRS public use file project. Technical report, U.S. Census Bureau Longitudinal Employer-Household Dynamics Program (2006)
13. Kinney, S.K., Reiter, J.P.: Making public use, synthetic files of the Longitudinal Business Database. In: Privacy in Statistical Databases: UNESCO Chair in Data Privacy International Conference Proceedings, Istanbul, Turkey (2008)
14. Raghunathan, T.E., Lepkowski, J.M., Van Hoewyk, J., Solenberger, P.: A multivariate technique for multiply imputing missing values using a sequence of regression models. Surv. Methodol. 27, 85–95 (2001)
15. Reiter, J.P.: Releasing multiply-imputed, synthetic public use microdata: an illustration and empirical study. J. Roy. Stat. Soc. A 168, 185–205 (2005)
16. Reiter, J.P.: Simultaneous use of multiple imputation for missing data and disclosure limitation. Surv. Methodol. 30, 235–242 (2004)
17. Reiter, J.P.: Satisfying disclosure restrictions with synthetic data sets. J. Off. Stat. 18, 531–544 (2002)
18. Fay III., R.E., Herriot, R.A.: Estimates of income for small places: an application of James-Stein procedures to Census data. J. Amer. Stat. Assoc. 74(366), 269–277 (1979)
19. Malec, D., Sedransk, J., Moriarity, C.L., LeClere, F.B.: Small area inference for binary variables in the National Health Interview Survey. J. Amer. Stat. Assoc. 92(439), 815–826 (1997)
20. Yucel, R.M.: Multiple imputation inference for multivariate multilevel continuous data with ignorable non-response. Phil. Trans. R. Soc. A 366(2008), 2389–2403 (1874)
21. Reiter, J.P., Raghunathan, T.E., Kinney, S.: The importance of modeling the sampling design in multiple imputation for missing data. Surv. Methodol. 32, 143–150 (2006)
22. Yu, M.: Disclosure Risk Assessments and Control. Doctoral Dissertation, University of Michigan (2008)
23. Datta, G.S., Fay, R.E., Ghosh, M.: Hierarchical and empirical Bayes multivariate analysis in small area estimation. In: Proceedings of the Bureau of the Census 1991 Annual Research Conference, pp. 63–79. U.S. Bureau of the Census, Washington (1991)
24. Rao, J.N.K.: Some recent advances in model-based small area estimation. Surv. Methodol. 25, 175–186 (1999)
25. Lindley, D.V., Smith, A.F.M.: Bayes estimates for the linear model. J. Roy. Stat. Soc. B 34(1), 1–41 (1972)
26. Binder, D.A.: On the variances of asymptotically normal estimators from complex surveys. Int. Stat. Rev. 51, 279–292 (1983)
27. Skinner, C.J., Holt, D., Smith, T.M.F.: Analysis of complex surveys. Wiley, Chichester (1989)

Disclosure Risk of Synthetic Population Data with Application in the Case of EU-SILC*

Matthias Templ[1,2] and Andreas Alfons[1]

[1] Department of Statistics and Probability Theory, Vienna University of Technology, Wiedner Hauptstraße 7, 1040 Vienna, Austria
templ@tuwien.ac.at, alfons@statistik.tuwien.ac.at
[2] Methods Unit, Statistics Austria, Guglgasse 13, 1110 Vienna, Austria

Abstract. In survey statistics, simulation studies are usually performed by repeatedly drawing samples from population data. Furthermore, population data may be used in courses on survey statistics to support the theory by practical examples. However, real population data containing the information of interest are in general not available, therefore synthetic data need to be generated. Ensuring data confidentiality is thereby absolutely essential, while the simulated data should be as realistic as possible. This paper briefly outlines a recently proposed method for generating close-to-reality population data for complex (household) surveys, which is applied to generate a population for Austrian EU-SILC (European Union Statistics on Income and Living Conditions) data. Based on this synthetic population, confidentiality issues are discussed using five different worst case scenarios. In all scenarios, the intruder has the complete information on key variables from the real survey data. It is shown that even in these worst case scenarios the synthetic population data are confidential. In addition, the synthetic data are of high quality.

Keywords: Survey Statistics, Synthetic Population Data, Data Confidentiality.

1 Introduction

In the analysis of survey data, variability due to sampling, imputation of missing values, measurement errors and editing must be considered. Statistical methods thus need to be evaluated with respect to the effect of these variabilites on point and variance estimates. A frequently used strategy to adequately measure such effects under different settings is to perform simulation studies by repeatedly drawing samples from population data (possibly using different sampling designs) and to compare the estimates with the true values of the sample frame.

* This work was partly funded by the European Union (represented by the European Commission) within the 7th framework programme for research (Theme 8, Socio-Economic Sciences and Humanities, Project AMELI (Advanced Methodology for European Laeken Indicators), Grant Agreement No. 217322). For more information on the project, visit http://ameli.surveystatistics.net

J. Domingo-Ferrer and E. Magkos (Eds.): PSD 2010, LNCS 6344, pp. 174–186, 2010.

Evaluating and comparing various statistical methods within such a *design-based* simulation approach under different *close-to-reality* settings is daily work for survey statisticians and has been done, e.g., in the research projects DACSEIS [1], EurEdit [2] and AMELI [3].

Furthermore, population data may be used for teaching courses on survey statistics. Realistic examples may help students to better understand issues in survey sampling, e.g., regarding different sampling designs.

Since suitable population data are typically not available, it is necessary to generate synthetic data. The generation of population microdata for selected surveys as a basis for Monte Carlo simulation studies is described in [1,4]. These procedures were extended in [3,5] to simulate close-to-reality population data for more complex surveys such as EU-SILC (*European Union Statistics on Income and Living Conditions*). However, confidentiality issues of such synthetic population data are only briefly addressed in these contributions.

Generation of population microdata for simulation studies is closely related to the field of *microsimulation* [6], yet the aims are quite different. Microsimulation models attempt to reproduce the behavior of individual units within a population for policy analysis purposes and are well-established within the social sciences. Nevertheless, they are highly complex and time-consuming. On the other hand, synthetic population microdata for simulation studies in survey statistics are used to evaluate the behavior of statistical methods. Thus fast computation is preferred to over-complex models.

Another approach towards the generation of microdata is to use multiple imputation to create *fully* or *partially* synthetic data sets, as proposed in [7,8]. This approach is further discussed in [9,10,11]. However, these contributions do not consider some important issues such as the generation of categories that do not occur in the original sample or the generation of structural zeros.

The rest of the paper is organized as follows. Section 2 outlines the framework for generating synthetic populations proposed in [5]. Then the data investigated in this paper are introduced in Section 3. Sections 4 and 5 discuss statistical disclosure control issues related to survey and population data. In Section 6, several scenarios for evaluating the confidentiality of synthetic population data are described, while Section 7 lists the obtained results for these scenarios. The final Section 8 concludes.

2 Generation of Synthetic Population Data

The generation of synthetic population data for surveys is described in great detail in [5]. Therefore, only the basic ideas of this framework are presented here. Several conditions for simulating population data are listed in [1,4,5]. The most important requirements are:

- Actual sizes of regions and strata need to be reflected.
- Marginal distributions and interactions between variables should be represented accurately.

- Heterogeneities between subgroups, in particular regional aspects, should be allowed.
- Data confidentiality must be ensured.

In general, the framework for generating synthetic population data consists of four steps:

1. In case of household data, set up the household structure.
2. Simulate categorical variables.
3. Simulate continuous variables.
4. Split continuous variables into components.

Not all of these steps need to be performed, depending on the survey of interest.

Step 1. When generating household data, the household structure is simulated separately for the different household sizes within each strata. Using the corresponding sample weights, the number of households is simply estimated by the Horvitz-Thompson estimator [12]. The structure of the population households is then simulated by resampling some basic variables from the sample households with probabilities proportional to the sample weights. For disclosure reasons, information from as few variables as possible should be used to construct the household structure (e.g., only age and gender information).

Step 2. For each stratum, the conditional distribution of any additional categorical variable is estimated with a multinomial logistic regression model. The previously simulated variables are thereby used as predictors. Furthermore, the sample weights are considered and it is possible to account for structural zeros. The main advantage of this approach is that it allows to generate combinations that do not occur in the sample, which is not the case for the procedure introduced in [1,4].

Step 3. Two approaches for simulating continuous variables are proposed in [5], but only the approach that performs better in the case of EU-SILC data is considered in this paper. First, the variable to be simulated is discretized using suitable breakpoints. The discretized variable is then then simulated as described in the previous step. Finally, the values of the continuous variable are randomly drawn from uniform distributions within the respective intervals. Note that the idea behind this approach is to divide the data into relatively small subsets so that the uniform distribution is not too much of an oversimplification.

Step 4. Splitting continuous variables into components is based on conditional resampling of fractions from the sample households with probabilities proportional to the sample weights. Only very few highly influential categorical variables should thereby be considered for conditioning. The resampled fractions are then multiplied with the previously simulated total.

The data simulation framework proposed in [5] is implemented in the R [13] package `simPopulation` [14]. In addition to the four steps of the procedure and a wrapper function to generate synthetic EU-SILC populations, various diagnostic plots are available.

3 Synthetic EU-SILC Population Data

The *European Union Statistics on Income and Living Conditions* (EU-SILC) is a complex panel survey conducted in EU member states and other European countries. It is mainly used for measuring risk-of-poverty and social cohesion in Europe [15]. The generation of synthetic population data based on Austrian EU-SILC survey data from 2006 is discussed and evaluated in [5]. The resulting synthetic population is investigated in this paper with respect to confidentiality issues. Table 1 lists the variables that are used in the analysis. A detailed description of all variables included in EU-SILC data and their possible outcomes is given in [16].

In order to demonstrate that the synthetic population data are of high quality, they are compared to the underlying sample data. Figure 1 contains mosaic plots of gender, region and household size for the sample and synthetic population data, respectively. Clearly, the plots show almost identical structures. In addition, the distribution of personal net income is visualized in Figure 2. On the left hand side, the cumulative distribution functions for the sample and population data, respectively, are displayed. For better visibility, only the main parts of the data are shown, which are nearly in perfect superposition. On the right hand side, the conditional distributions with respect to gender are represented by box plots. The fit of the distribution within the subgroups is excellent and heterogeneities between the subgroups are very well reflected. Note that points outside the extremes of the whiskers are not plotted. For extensive collections of results showing that the multivariate structure of the data is well preserved, the reader is referred to [5,17].

Table 1. Variables of the synthetic EU-SILC population data used in this paper

Variable	**Type**	
Region	Categorical	9 levels
Household size	Categorical	9 levels
Age category	Categorical	15 levels
Gender	Categorical	2 levels
Economic status	Categorical	7 levels
Citizenship	Categorical	3 levels
Personal net income	Semi-continuous	

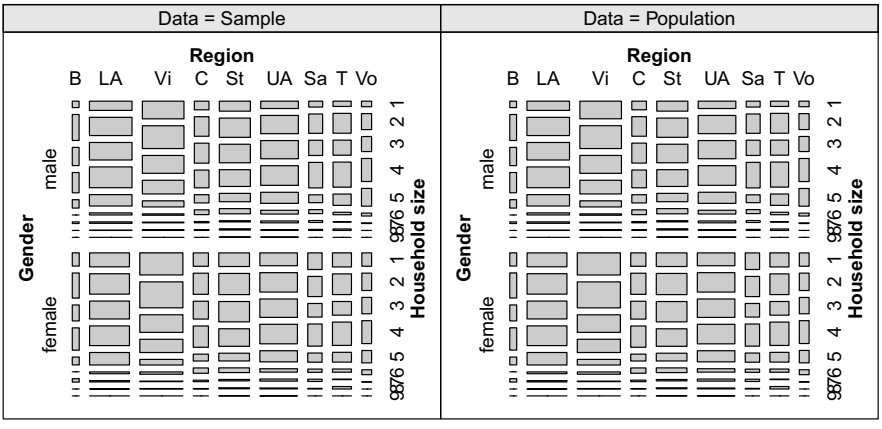


Fig. 1. Mosaic plots of gender, region and household size of the Austrian EU-SILC sample from 2006 and the resulting synthetic population

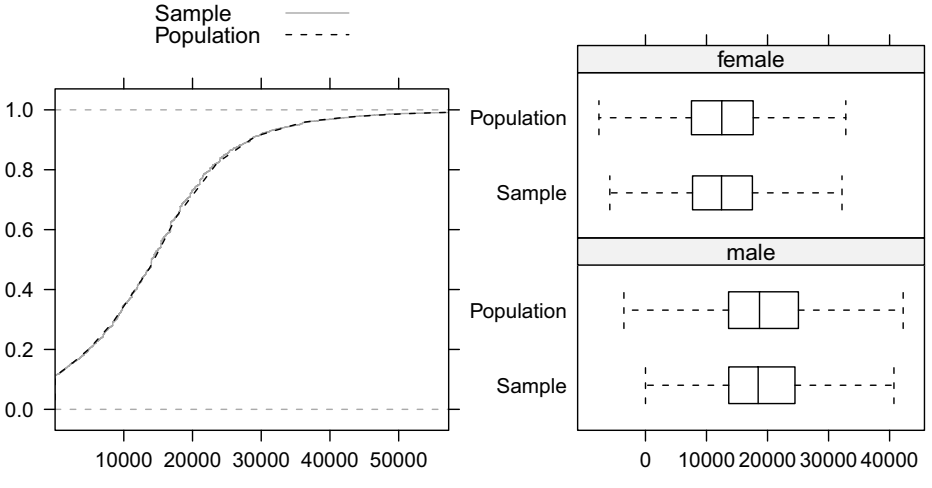


Fig. 2. Personal net income in the Austrian EU-SILC sample from 2006 and the resulting synthetic population. *Left*: Cumulative distribution functions of personal net income. Only the main parts of the data are shown for better visibility. *Right:* Box plots of the conditional distributions with respect to gender. Points outside the extremes of the whiskers are not plotted.

4 A Global Disclosure Risk Measure for Survey Data

A popular global measure of the reidentification risk for survey data is given by the number of uniquenesses in the sample that are unique in the population as well. Let m categorical key variables in the sample and population

data be denoted by $\boldsymbol{x}_j^S = (x_{1j}^S, \ldots, x_{nj}^S)'$ and $\boldsymbol{x}_j^P = (x_{1j}^P, \ldots, x_{Nj}^P)'$, respectively, $j = 1, \ldots, m$, where n and N give the corresponding number of observations. For an observation in the sample given by the index $c = 1, \ldots, n$, let J_c^S and J_c^P denote the index sets of observations in the sample and population data, respectively, with equal values in the m key variables:

$$\begin{aligned} J_c^S &:= \{j = 1, \ldots, n : x_{jk}^S = x_{ck}^S,\ k = 1, \ldots, m\}, \\ J_c^P &:= \{j = 1, \ldots, N : x_{jk}^P = x_{ck}^S,\ k = 1, \ldots, m\}. \end{aligned} \tag{1}$$

Furthermore, a function $\mathcal{I}$ is defined as

$$\mathcal{I}(J) := \begin{cases} 1 & \text{if } |J| = 1, \\ 0 & \text{else.} \end{cases} \tag{2}$$

The global disclosure risk measure can then be expressed by

$$\tau_0 := \sum_{c=1}^{n} \mathcal{I}(J_c^S) \cdot \mathcal{I}(J_c^P). \tag{3}$$

Note that the notation in (3) differs from the common definition. For comparison, see, e.g., the risk measures in [18,19]. The notation used in (3) describes the same phenomenon, but provides more flexibility in terms of software implementation [20] and allows to formulate the adapted risk measures given in the following section.

Clearly, the risk of reidentification is lower the higher the corresponding population frequency count. If the population frequency count is sufficiently high, it is not possible for an intruder to assign the observation for which they hold information with absolute certainty. Hence the intruder does not know whether the reidentification was successful. However, the true frequency counts of the population are usually unknown and need to be estimated by modeling the distribution of the population frequencies.

In Section 6, the global disclosure risk measure τ_0 is modified to estimate the disclosure risk for synthetic population data in certain scenarios instead of survey data.

5 Confidentiality of Synthetic Population Data

The motivation for generating close-to-reality population data is to make the resulting data sets publicly available to researchers for use in simulation studies or courses on survey statistics. Therefore, the disclosure risk of such data needs to be low, while at the same time the multivariate structure should be as realistic as possible.

If population data are generated from perturbed survey data, confidentiality is guaranteed whenever the underlying survey data are confidential. Perturbing survey data is typically done by performing recodings and suppression such that k-anonymity [21,22] is provided for categorical key variables, as well as low

risk of reidentification on the individual level is ensured [23,24,25, and references therein]. In any case, perturbation implies information loss. Usually not all combinations of categorical key variables are still included in the perturbed sample and outliers in continuous variables are often modified to a great extent. It is thus favorable to use information of the unperturbed sample to generate synthetic populations, as this increases the quality of the resulting data.

Based on the ideas proposed in [7,8], the generation of *fully* or *partially* synthetic population data using multiple imputation is discussed in [9,10,11]. More precisely, let p be the number of variables in the sample and let the first k, $1 \leq k < p$, categorical variables be available for the population data from administrative sources. These first k variables are released unchanged, while the remaining $p - k$ variables are estimated using regression based multiple imputation. It is important to note that the first k variables of real population data may still contain unique combinations after cross tabulation, therefore further investigation may be necessary to ensure confidentiality. Probabilities of reidentification for such synthetic data have been studied in [26], based on the work of [27,28], by matching the synthetic data with the intruder's data on some predefined key variables.

The situation for synthetic population data generated by the approach of [5] is somewhat different. A very low number of basic categorical variables are generated in the first step by resampling from the actual survey data. Since the sample weights are thereby used as probability weights, on average k-anonymity is provided with respect to these basic variables, where k denotes the smallest sample weight. In surveys, k is typically very high (> 500), hence the disclosure risk is very low. However, additional categorical and continuous variables are generated based on models obtained from the actual survey data. In particular, the generation of continuous variables involves random draws from certain intervals.

With the additional categorical variables, some unique combinations may be introduced in the synthetic population data. If such a combination is not unique in the real population, it is not useful to an intruder. On the other hand, if such a combination is unique in the real population as well, it must be ensured that the values of the other variables in the synthetic population data are not too close to the real values. Most notably, it is of interest to measure the difference in continuous variables of the successfully matched statistical units.

In addition, unique combinations in the real population may even be critical if they are not unique in the synthetic population data. An intruder could in this case look for all occurrences of such a combination in the synthetic population. If the corresponding units have too similar values in a (continuous) variable of interest, the intruder may be able to infer some information on the original value, since the synthetic values have been predicted with models obtained from the real sample data.

In order to investigate these issues in more detail, various disclosure scenarios are introduced in the following section. Section 7 then presents the results for the synthetic EU-SILC population data described in Section 3.

6 Disclosure Scenarios for Synthetic Population Data

Five different scenarios are considered to evaluate the confidentiality of synthetic data generated with the framework proposed in [5]. These scenarios are motivated by the synthetic EU-SILC population data, hence only a continuous variable is considered to contain confidential information, while there are m categorical key variables. In the case of EU-SILC, the confidential variable is *personal net income* and the key variables are *region*, *household size*, *age category*, *gender*, *economic status* and *citizenship* (see Table 1). Let the confidential continuous variable for the original sample and synthetic population, respectively, be denoted by $\boldsymbol{y}^S = (y_1^S, \ldots, y_n^S)'$ and $\boldsymbol{y}^U = (y_1^U, \ldots, y_N^U)'$, while the categorical key variables are denoted by $\boldsymbol{x}_j^S = (x_{1j}^S, \ldots, x_{nj}^S)'$ and $\boldsymbol{x}_j^U = (x_{1j}^U, \ldots, x_{Nj}^U)'$, $j = 1, \ldots, m$, analogous to the definitions in Section 4. Furthermore, let J_c^S be defined as in (1) and let J_c^U be defined accordingly as

$$J_c^U := \{j = 1, \ldots, N : x_{jk}^U = x_{ck}^S,\ k = 1, \ldots, m\}. \tag{4}$$

In the following scenarios, the intruder has knowledge of the m key variables for all observations from the original sample and tries to acquire information on the confidential variable.

It should be noted that the link to the global risk measure from (3) is loosened in the following. Disclosure is considered to occur if the value of the confidential variable for a unique combination of key variables in the sample can be closely estimated from the synthetic population data, given a prespecified value of accuracy p. However, such a sample uniqueness does not need to be unique in the true population, in which case close estimation of the confidential variable would not necessarily result in disclosure. In this sense, the following scenarios can be considered worst case scenarios and the reidentification risk is thus overestimated. Proper analysis with estimation of the true population uniquenesses is future work.

6.1 Scenario 1: Attack Using One-to-One Matches in Key Variables with Information on the Data Generation Process

The intruder in this scenario tries to find one-to-one matches between their data and the synthetic population data. Moreover, they know the intervals from which the synthetic values were drawn. For details on the data generation procedure, the reader is referred to [5]. Let these intervals be denoted by $[l_j, u_j]$, $j = 1, \ldots, N$, and let l be a function giving the length of an interval defined as $l([a, b]) := b - a$ and $l(\emptyset) := 0$. With a prespecified value of accuracy p defining a symmetric interval around a confidential value, (3) is reformulated as

$$\tau_1 := \sum_{c=1}^{n} \mathcal{I}(J_c^S) \cdot \mathcal{I}(J_c^U) \cdot \frac{l([y_c^S(1-p), y_c^S(1+p)] \cap [l_{j_c}, u_{j_c}])}{l([l_{j_c}, u_{j_c}])}, \tag{5}$$

where $j_c \in J_c^U$ if $|J_c^U| = 1$, i.e., j_c is the index of the unit in the synthetic population with the same values in the key variables as the cth unit in the intruder's

data if such a one-to-one match exists, otherwise it is a dummy index. The last term in (5) thereby gives the probability that for the successfully matched unit, the synthetic value drawn from the interval $[l_{j_c}, u_{j_c}]$ is sufficiently close to the original value y_c^S.

6.2 Scenario 2: Attack Using One-to-One Matches in Key Variables without Information on the Data Generation Process

In general, an intruder does not have any knowledge on the intervals from which the synthetic values were drawn. In this case, reidentification is successful if the synthetic value itself of a successfully matched unit is sufficiently close to the original value. The risk of reidentification thus needs to be reformulated as

$$\tau_2 := \sum_{c=1}^{n} \mathcal{I}(J_c^S) \cdot \mathcal{I}(J_c^U) \cdot \mathbb{I}_{[y_c^S(1-p), y_c^S(1+p)]}(y_{j_c}^U), \tag{6}$$

where j_c is defined as above and $\mathbb{I}_A$ denotes the indicator function for a set A.

6.3 Scenario 3: Attack Using All Occurrences in Key Variables with Information on the Data Generation Process

This scenario is an extension of Scenario 1, in which the intruder does not only try to find one-to-one matches, but looks for all occurrences of a unique combination from their data in the synthetic population data. Keep in mind that the intruder in this scenario knows the intervals from which the synthetic values were drawn. For a unique combination in the intruder's data, reidentification is possible if the probability that the synthetic values of all matched units are sufficiently close to the original value. Hence the disclosure risk from (5) changes to

$$\tau_3 := \sum_{c=1}^{n} \mathcal{I}(J_c^S) \cdot \prod_{j \in J_c^U} \frac{l([y_c^S(1-p), y_c^S(1+p)] \cap [l_j, u_j])}{l([l_j, u_j])}. \tag{7}$$

6.4 Scenario 4: Attack Using All Occurrences in Key Variables without Information on the Data Generation Process

In an analogous extension of Scenario 2, reidentification of a unique combination from the intruder's data is successful if the synthetic values themselves of all matched units are sufficiently close to the original value. Equation (6) is in this case rewritten as

$$\tau_4 := \sum_{c=1}^{n} \mathcal{I}(J_c^S) \cdot \prod_{j \in J_c^U} \mathbb{I}_{[y_c^S(1-p), y_c^S(1+p)]}(y_j^U). \tag{8}$$

6.5 Scenario 5: Attack Using Key Variables for Model Predictions

In this scenario, the intruder uses the information from the synthetic population data to obtain a linear model for $\boldsymbol{y}^U$ with predictors $\boldsymbol{x}_j^U$, $j = 1, \ldots, m$:

$$\boldsymbol{y}^U = \beta_0 + \beta_1 \boldsymbol{x}_1^U + \ldots + \beta_m \boldsymbol{x}_m^U + \boldsymbol{\varepsilon}. \tag{9}$$

For a unique combination of the key variables, reidentification is possible if the corresponding predicted value is sufficiently close to the original value. Let the predicted values of the intruder's data be denoted by $\hat{\boldsymbol{y}}^S = (\hat{y}_1^S, \ldots, \hat{y}_n^S)'$. Then the disclosure risk can be formulated as

$$\tau_5 := \sum_{c=1}^{n} \mathcal{I}(J_c^S) \cdot \mathbb{I}_{[y_c^S(1-p), y_c^S(1+p)]}(\hat{y}_c^S). \tag{10}$$

Note that for large population data, the computational costs for fitting such a regression model are very high, so an intruder needs to have a powerful computer with very large memory. Furthermore, the intruder could also perform a stepwise model search using an optimality criterion such as the AIC [29].

7 Results

The disclosure risk of the synthetic Austrian EU-SILC population data described in Section 3 is analyzed in the following with respect to the scenarios defined in the previous section. In these scenarios, the intruder has knowledge of the categorical key variables *region*, *household size*, *age category*, *gender*, *economic status* and *citizenship* for all observations in the original sample used to generate the data. In addition, the intruder tries to obtain information on the confidential variable *personal net income* (see Table 1 for a description of these variables). The original sample thereby consists of $n = 14\,883$ and the synthetic population of $N = 8\,182\,218$ observations.

Note that this paper only evaluates the risk of reidentification for this specific synthetic data set. In order to get more general results regarding confidentiality of the data generation process, many data sets need to be generated in a simulation study and the average values need to be reported. This task, however, is beyond the scope of this paper.

Table 2 lists the results for the risk measures for the investigated scenarios using different values of the accuracy parameter p. Besides the absolute values, the relative values with respect to the size of the intruder's data set are presented, which give the probabilities of successful reidentification.

The results show that even if an intruder is able to reidentify an observation, they do not gain any useful information, as the probability that the obtained value is sufficiently close to the original value is extremely low.

In particular if the intruder tries to find one-to-one matches (Scenarios 1 and 2), the probability of a successful reidentification is only positive for $p = 0.05$ and if they have information on the data generation process, i.e., the intervals from which the synthetic values were drawn.

Table 2. Results for Scenarios 1-5 using different values for the accuracy parameter p

Scenario	Risk measure	p = 0.01	p = 0.02	p = 0.05
1	τ_1	0	0	0.052
2	τ_2	0	0	0
3	τ_3	$1.1 \cdot 10^{-8}$	$1.2 \cdot 10^{-6}$	0.053
4	τ_4	15	15	15
5	τ_5	20	43	110
1	τ_1/n	0	0	$3.5 \cdot 10^{-6}$
2	τ_2/n	0	0	0
3	τ_3/n	$6.7 \cdot 10^{-13}$	$8.6 \cdot 10^{-11}$	$3.5 \cdot 10^{-6}$
4	τ_4/n	0.001	0.001	0.001
5	τ_5/n	0.001	0.003	0.007

If the intruder looks for all occurrences of a unique combination from their data in the synthetic population, using information on the data generation process hardly changes the probabilities of reidentification (Scenario 3). This is not a surprise given the formula in (7), since for such a unique combination, the probabilities that the corresponding synthetic values are sufficiently close to the original value need to be multiplied. On the other hand, if the intruder uses only the synthetic values (Scenario 4), some observations are successfully reidentified. Nevertheless, the probabilities of reidentification are extremely low.

Among the considered scenarios, Scenario 5 leads to the highest disclosure risk. However, the regression model in this scenario comes with high computational costs and the probabilities of reidentification are still far too low to obtain any useful information.

8 Conclusions

Synthetic population data play an important part in the evaluation of statistical methods in the survey context. Without such data, it is not possible to perform design-based simulation studies.

This paper gives a brief outline of the flexible framework proposed in [5] for simulating population data for (household) surveys based on available sample data. The framework is applicable to a broad class of surveys and is implemented along with diagnostic plots in the R package `simPopulation`. In the case of EU-SILC, the data generation procedure led to excellent results with respect to information loss.

In this paper, confidentiality issues of the generated synthetic EU-SILC population are discussed based on five different worst case scenarios. The results show that while reidentification is possible, an intruder would not gain any useful information from the purely synthetic data. Even if they successfully reidentify a

unique combination of key variables from their data, the probability that the obtained value is close to the original value is extremely low for all considered worst case scenarios.

Due to our experiences and the results from the investigated scenarios, we can argue that synthetic population data generated with the methodology introduced in [5] and implemented in `simPopulation` are confidential and can be distributed to the public. Researchers could then use this data to evaluate the effects of different sampling designs, missing data mechanisms and outlier models on the estimator of interest in design-based simulation studies.

References

1. Münnich, R., Schürle, J., Bihler, W., Boonstra, H.-J., Knotterus, P., Nieuwenbroek, N., Haslinger, A., Laaksonen, S., Eckmair, D., Quatember, A., Wagner, H., Renfer, J.-P., Oetliker, U., Wiegert, R.: Monte Carlo simulation study of European surveys. In: DACSEIS Deliverables D3.1 and D3.2, University of Tübingen (2003)
2. Chambers, R.: Evaluation criteria for statistical editing and imputation. In: EurEdit Deliverable D3.3, Department of Social Statistics, University of Southhampton (2001)
3. Alfons, A., Templ, M., Filzmoser, P., Kraft, S., Hulliger, B.: Intermediate report on the data generation mechanism and on the design of the simulation study. In: AMELI Deliverable 6.1, Vienna University of Technology (2009)
4. Münnich, R., Schürle, J.: On the simulation of complex universes in the case of applying the German Microcensus. DACSEIS research paper series, vol. 4. University of Tübingen (2003)
5. Alfons, A., Kraft, S., Templ, M., Filzmoser, P.: Simulation of synthetic population data for household surveys with application to EU-SILC. Research Report CS-2010-1, Department of Statistics and Probability Theory, Vienna University of Technology (2010), `http://www.statistik.tuwien.ac.at/forschung/CS/CS-2010-1complete.pdf`
6. Clarke, G.P.: Microsimulation: an introduction. In: Clarke, G.P. (ed.) Microsimulation for Urban and Regional Policy Analysis, Pion, London (1996)
7. Rubin, D.B.: Discussion: Statistical disclosure limitation. J. Off. Stat. 9(2), 461–468 (1993)
8. Little, R.J.A.: Statistical analysis of masked data. J. Off. Stat. 9(2), 407–426 (1993)
9. Raghunathan, T.E., Reiter, J.P., Rubin, D.B.: Multiple imputation for statistical disclosure limitation. J. Off. Stat. 19(1), 1–16 (2003)
10. Drechsler, J., Bender, S., Rässler, S.: Comparing fully and partially synthetic datasets for statistical disclosure control in the German IAB Establishment Panel. Trans. Data Priv. 1(3), 105–130 (2008)
11. Reiter, J.P.: Using multiple imputation to integrate and disseminate confidential microdata. Int. Stat. Rev. 77(2), 179–195 (2009)
12. Horvitz, D.G., Thompson, D.J.: A generalization of sampling without replacement from a finite universe. J. Am. Stat. Assoc. 47(260), 663–685 (1952)
13. R Development Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna (2010) ISBN 3-900051-07-0, `http://www.r-project.org`
14. Kraft, S., Alfons, A.: simPopulation: Simulation of synthetic populations for surveys based on sample data, R package version 0.1.2 (2010)

15. Atkinson, T., Cantillon, B., Marlier, E., Nolan, B.: Social Indicators: The EU and Social Inclusion. Oxford University Press, New York (2002), ISBN: 0-19-925349-8
16. Eurostat. Description of target variables: Cross-sectional and longitudinal. EU-SILC 065/04, Eurostat, Luxembourg (2004)
17. Kraft, S.: Simulation of a Population for the European Living and Income Conditions Survey. Master's thesis, Vienna University of Technology (2009)
18. Rinott, Y., Shlomo, N.: A generalized negative binomial smoothing model for sample disclosure risk estimation. In: Domingo-Ferrer, J., Franconi, L. (eds.) PSD 2006. LNCS, vol. 4302, pp. 82–93. Springer, Heidelberg (2006) ISBN 978-3-540-49330-3
19. Elamir, E.A.H., Skinner, C.J.: Record level measures of disclosure risk for survey microdata. J. Off. Stat. 22(3), 525–539 (2006)
20. Templ, M.: New Developments in Statistical Disclosure Control and Imputation: Robust Statistics Applied to Official Statistics. Südwestdeutscher Verlag für Hochschulschriften, Germany (2009) ISBN 3838108280
21. Samarati, P., Sweeney, L.: Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression. Technical Report SRI-CSL-98-04, SRI International (1998)
22. Sweeney, L.: k-anonymity: a model for protecting privacy. Int. J. Uncertain. Fuzziness Knowl. Syst. 10(5), 557–570 (2002)
23. Franconi, L., Polettini, S.: Individual risk estimation in μ-ARGUS: A review. In: Domingo-Ferrer, J., Torra, V. (eds.) Privacy in Statistical Databases. LNCS, vol. 3050, pp. 262–272. Springer, Heidelberg (2004) ISBN 978-3-540-22118-2
24. Domingo-Ferrer, J., Mateo-Sanz, J.M.: Practical data-oriented microaggregation for statistical disclosure control. IEEE Trans. Knowl. Data Eng. 14(1), 189–201 (2002)
25. Templ, M., Meindl, B.: Robust statistics meets SDC: New disclosure risk measures for continuous microdata masking. In: Domingo-Ferrer, J., Saygın, Y. (eds.) PSD 2008. LNCS, vol. 5262, pp. 113–126. Springer, Heidelberg (2008) ISBN 978-3-540-87470-6
26. Reiter, J.P., Mitra, R.: Estimating risks of identification disclosure in partially synthetic data. J. Priv. Confid. 1(1), 99–110 (2009)
27. Duncan, G.T., Lambert, D.: Disclosure-limited data dissemination. J. Am. Stat. Assoc. 81(393), 10–28 (1986)
28. Fienberg, S.E., Makov, U.E., Sanil, A.P.: A bayesian approach to data disclosure: Optimal intruder behavior for continuous data. J. Off. Stat. 13(1), 75–89 (1997)
29. Akaike, H.: Statistical predictor identification. Ann. Inst. Stat. Math. 22(2), 203–217 (1970)

Differential Privacy and the Risk-Utility Tradeoff for Multi-dimensional Contingency Tables

Stephen E. Fienberg[1,2,3], Alessandro Rinaldo[1,2], and Xiaolin Yang[1,⋆]

[1] Department of Statistics, Carnegie Mellon University, Pittsburgh, PA 15213, USA
fienberg@stat.cmu.edu, arinaldo@stat.cmu.edu, xyang@stat.cmu.edu
[2] Machine Learning Department, Carnegie Mellon University, Pittsburgh, PA 15213, USA
[3] Cylab, and i-Lab, Carnegie Mellon University, Pittsburgh, PA 15213, USA

Abstract. The methodology of differential privacy has provided a strong definition of privacy which in some settings, using a mechanism of doubly-exponential noise addition, also allows for extraction of informative statistics from databases. A recent paper extends this approach to the release of a specified set of margins from a multi-way contingency table. Privacy protection in such settings implicitly focuses on small cell counts that might allow for the identification of units that are unique in the database. We explore how well the mechanism works in the context of a series of examples, and the extent to which the proposed differential-privacy mechanism allows for sensible inferences from the released data.

1 Introduction

Contingency tables, databases arising from the cross-classification of a sample or a population according to a collection of categorical variables, are among the most prevalent forms of statistical data, especially in the context of official statistics and sample surveys. When the data displayed are a random sample from a population, the most widely used statistical methods for analyzing the data are log-linear model methods. A key feature of log-linear models applied to multi-dimensional contingency tables is the fact that the minimal sufficient statistics are sets of possibly overlapping marginals, from which one can compute maximum likelihood estimates, e.g., see [2,9,12]. Fienberg and Slavkovic [11] review the statistical literature on privacy protection of results from contingency tables focusing on the exact release of minimal sufficient marginals under a well-fitting log-linear model and they discuss this method in the context of the Risk-Utility (RU) trade-off initially proposed in Duncan et al. [5], where risk was defined in terms of protection of small counts in the table. Dobra et al. [4] further insight into the RU-trade-off problem for large sparse tables using recent results from algebraic statistics. Winkler [15] proposed a method to reduce re-identification risk while preserving analytic properties by placing upper and lower bounds on key aggregates needed for loglinear modeling and also on large sets of small cells and sampling zeros.

The methodology of differential privacy [6,7] has provided a strong definition of privacy which in some settings, using a mechanism of doubly-exponential noise addition, also allows for extraction of informative statistics from databases. A recent paper by

⋆ Corresponding author.

J. Domingo-Ferrer and E. Magkos (Eds.): PSD 2010, LNCS 6344, pp. 187–199, 2010.

Barak et al. [1] extends this approach to the release of a specified set of margins from a multi-way contingency table. Adding non-integer noise in such contexts poses a variety of additional problems: violation of non-negativity, incompatible margins, and infeasible tables. The proposed methodology purports to handle all of these problems. In this paper, we explore how well the mechanism works in the context of a series of examples, and the extent to which the proposed differential-privacy mechanism allows for sensible inferences from the released data.

2 Differential Privacy

Let $\mathcal{D}$ denote the set of databases. A privacy protecting mechanism is a randomized function $K: \mathcal{D} \rightarrow \mathcal{D}$. The output of K is a random database called the sanitized database.

Definition 1. *The privacy protecting mechanism K satisfies ϵ-differential privacy if, for all databases D_1 and D_2 in $\mathcal{D}$ differing on at most one record, and all measurable subsets S of the range of K,*

$$Pr[K(D_1) \in S] \leq \exp(\epsilon) Pr[K(D_2) \in S].$$

The smaller ϵ, the greater the privacy provided by the mechanism, in the sense that the probability distribution of the sanitized database is rather insensitive to a one-record change in the input database. Wasserman and Zhou [13, Theorem 2.4] provide a related statistical interpretation of differential privacy based on hypothesis testing theory.

3 Notation for Binary Contingency Tables

A 2^k contingency table arises from the cross classification of n individuals according to k binary categorical variables, where each cell of the table corresponds to the number of times a given combination of the k variables occurred in the sample. It is convenient for us to think of a table x as a vector in R^{2^k}. We represent each cell i of the table x as a vertex of the k-dimensional unit hypercube: $x = \left\{x_i, i \in \{0,1\}^k\right\}$. For a given subset $\alpha \subset \{1, \ldots, k\}$, we write $i_\alpha = \{i_j, j \in \alpha\} \in \{0,1\}^{|\alpha|}$ for the α-coordinate projection of i. The α-marginal table of x is the $|\alpha|$-dimensional binary array $x^\alpha = \left\{x_{i_\alpha}, i_\alpha \in \{0,1\}^{|\alpha|}\right\}$, whose i_α entry is obtained by summing over the cells j of x having identical α-coordinate projection:

$$x_{i_\alpha} = \sum_{j: j_\alpha = i_\alpha} x_j. \tag{1}$$

We will write compactly $x^\alpha = C^\alpha x$, were C^α is the $2^{|\alpha|} \times 2^k$ matrix realizing the sums in equaition (1). Also, with a slight abuse of notation, we refer to both α and x^α as margins.

4 The Risk-Utility Trade-Off

Let $A \subset 2^{\{0,1\}^k}$ be a collection of margins, such that $\cup_{\alpha \in A} = \{1, \ldots, k\}$ and $\alpha_1 \not\subset \alpha_2$ for any $\alpha_1, \alpha_2 \in A$.

From the theory of log-linear models [2,12], we know that each such collection $A \subset 2^{\{0,1\}^k}$ encode a statistical model for the the probabilistic dependence among the k attributes, each of which as a categorical random variable. Specifically, each A specify a collection of positive probability distributions over $\{0,1\}^k$ obeying a set of rules known as Markov properties. Each probability distribution is a point in the simplex in R^{2^k} such that p_i denotes the probability of observing cell i. The corresponding marginal tables $\{x^\alpha, \alpha \in A\}$ are minimal sufficient statistics for the model determined by A. This means that, from an inferential standpoint, the A-margins of x contains as much statistical information as x itself. Furthermore, they determine the maximum likelihood estimator (MLE) $\hat{p}$, which is the unique probability distribution in the model encoded by A that makes x the "most likely" sample that we could have observed. The MLE possess many optimal properties and, in particular, and we can use it to assess the fit of the model A using the likelihood ratio test statistic

$$\sum_i x_i \log\left(\frac{x_i}{n\hat{p}_i}\right). \tag{2}$$

From a privacy protection perspective the table x contains potentially sensitive information whose public release would entail a violation of privacy. Because the release of some information from such databases is a public utility, a database curator overseeing the table seeks to implement a mechanism of partial data release that are safe from the privacy standpoint. While the A-margins contain only aggregate (partial) information about x and thus appear to be a natural candidates for a data release [11,4], marginal releases may not in general correspond to a private-preserving mechanism, especially when the data base is sparse and contains many small counts [1]. By titrating the privacy mechanism we might also be able to apply some form of perturbation to the data and yet also produce statistical useful results.

5 The Differential Privacy Mechanism for Contingency Tables

We represent a set $\alpha \subset \{1, \ldots, k\}$ as a vector in $\{0,1\}^k$ whose positive coordinates are precisely α. In particular, when we speak of α-margin, we are treating α as a point in $\{0,1\}^k$. For vectors $\alpha, \beta \in R^{2^k}$, we will denote the L_1 norm as $\|\alpha\|_1 = \sum_i |\alpha_i|$ and the standard inner product as $\langle \alpha, \beta \rangle = \sum_i \alpha_i \beta_i$. Let $\{f^\alpha, \alpha \in \{0,1\}^k\}$ be the Fourier basis for R^{2^k}, whose α element is the vector $f^\alpha = \{f^\alpha_\beta, \beta \in \{0,1\}^k\}$, where

$$f^\alpha_\beta = \frac{1}{2^{k/2}}(-1)^{\langle \alpha, \beta \rangle}.$$

Barak et al. [1] show that, for every marginal β, the orthonormal Fourier basis yields a basis for $R^{2^{|\beta|}}$, in the sense that

$$C^\beta x = \sum_{\alpha \preceq \beta} \langle f^\alpha, x \rangle C^\beta f^\alpha,$$

where for $\alpha, \beta \in \{0,1\}^k$, $\alpha \preceq \beta$ signifies that every non-zero coordinate of α is also a non-zero coordinate of β. The Fourier basis representation is exactly the traditional u-parametrization of log-linear models e.g., as described in [2]; equivalently, it gives the direct sum decomposition of R^{2^k} in terms of the subspaces of interaction, e.g., see [12, Appendix B]. Based on the Fourier basis representation of the marginal tables, Barak et al. [1] proposed a differentially private mechanism for releasing a prescribed set of margins A from a binary table x, which we reproduce in Table 1. They showed that the algorithm possesses the following properties.

Theorem 1. *The privacy mechanism of Table 1 satisfies differential privacy and, for each* $\delta \in (0,1)$*, with probability at least* $(1-\delta)$*,*

$$\|C^\alpha x - C^\alpha w'\|_1 \leq 2^{|\alpha|} 8 \frac{|B|}{\epsilon} \log\left(\frac{|B|}{\delta}\right) + |B|,$$

uniformly over all $\alpha \in A$*.*

Barak et al. [1] argue that the above mechanism is simultaneously (i) private, since it satisfies the strong requirement of differential privacy, (ii) accurate, as it provides probabilistic guarantees on the maximal L_1 distance between the observed and release margins and (iii) consistent, as it release a margins that can be realized by an integer-valued table (namely w').

Remarks

1. The result is independent of the sample size, and the accuracy guarantees depend only on the model complexity $|B|$ and the differential privacy parameter ϵ.

Table 1. The differentially private mechanism for binary contingency tables

1. Inputs: a differential privacy parameter ϵ, a binary k-dimensional table x and a set of margins A.
2. Let B the downward closure of A with respect to $\preceq$.
3. Generate $\{X_\beta, \beta \in B\}$ as independent random variables with common distribution $\mathrm{Lap}\left(\frac{2|B|}{\epsilon 2^{k/2}}\right)$.
4. For each $\beta \in B$, compute the perturbed β-marginal $\phi^\beta = \langle f^\beta, x\rangle + X_\beta$
5. Solve for $w = \{w_\alpha, \alpha \in \{0,1\}^k\}$ the linear program

$$\begin{gathered} \min b \\ \text{subject to:} \\ w_\alpha \geq 0, \quad \forall \alpha \\ \phi_\beta - \textstyle\sum_\alpha w_\alpha f_\alpha^\beta \leq b, \quad \forall \beta \in B \\ \phi_\beta - \textstyle\sum_\alpha w_\alpha f_\alpha^\beta \geq -b, \quad \forall \beta \in B. \end{gathered}$$

6. Round w to w', where, for each $\alpha \in \{0,1\}^k$, w'_α is the nearest integer to w_α.
7. Return the A-margins of w'.

2. The linear program described above may return a solution for which $b > 0$ (in fact, we have often observed this phenomenon in our computations). This implies that there does not exist any real-valued non-negative table with B-margins given by $\{\phi^\beta, \beta \in B\}$.
3. The linear program has typically many (in fact infinite) solutions.
4. The proof of Theorem 7 in [1] implicitly assume that $b = 0$, which, as we mentioned, does not hold in general.

6 Empirical Evaluation of the Differential Privacy Mechanism

We now analyze the statistical properties of the privacy preserving mechanism of [1] on the three real-life datasets. We study empirically whether the algorithm in Table 1 for producing differentially private results,is also statistically robust, in the sense that the results of statistical analyses of the sanitized margins do not deviate significantly from the results obtained using the original database. In particular, we are interested in the rather basic question of whether a model that fits the original database well will also fit the perturbed data. We work with three well-analyzed examples, the full data for which we provide in the appendix:

1. The data in Table 4 is a sparse 6-dimensional binary contingency table that was obtained from the cross-classification of six dichotomous categorical variables, labeled with the letters A-F, recording the parental alleles corresponding to six

Table 2. Table dimension, sample size, chosen model and likelihood ratio statistic (2) for the three tables analyzed

Table	Dimension	Sample Size	Model	LR
Edwards	$k = 6$	$n = 70$	[AD][AB][BE][CE][CF]	22.96
Czech	$k = 6$	$n = 1841$	[BF][ADE][ABCE]	48.18
Rochdale	$k = 8$	$n = 665$	[ACE][ACG][ADG][BDH] [BF][BE][CEF][CFG]	238.18

Table 3. Variance of the additive noise and bounds for different values of ϵ

	ϵ		
	0.01	1	10
Edwards	Lap(300) $38400\log(12/\delta) + 12$	Lap(3) $384\log(12/\delta) + 12$	Lap(0.3) $38.4\log(12/\delta) + 12$
Czech	Lap(550) $70400\log(22/\delta) + 22$	Lap(5.5) $704\log(22/\delta) + 22$	Lap(0.55) $70.4\log(22/\delta) + 22$
Rochdale	Lap(362.5) $185600\log(29/\delta) + 29$	Lap(3.625) $1856\log(29/\delta) + 29$	Lap(0.3625) $185.6\log(29/\delta) + 29$

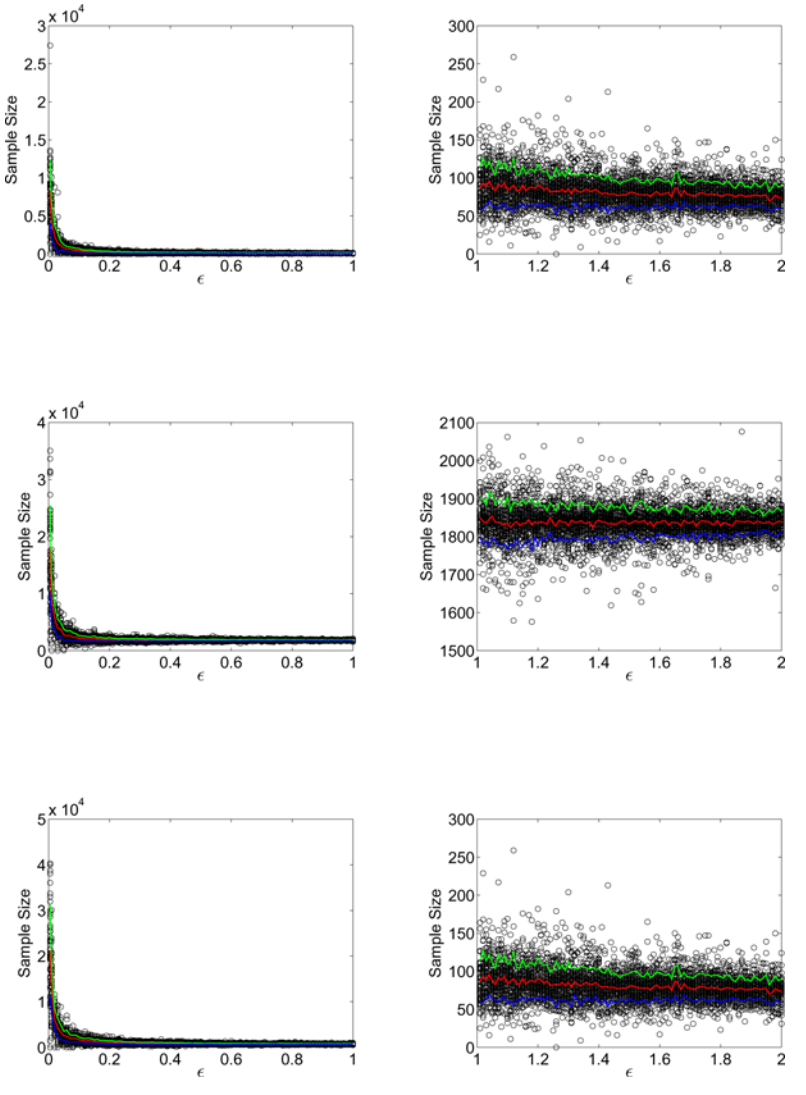


Fig. 1. Sample sizes for the Fungus table (top row), Czech autoworker table (middle) and Rochdale table (bottom). To improve readability, for each table,we split the plot in two parts, for $\epsilon < 1$ (left) and $\epsilon \geq 1$ (right). The three lines represent the mean plus or minus one standard deviation.

loci along a chromosome strand of a barley powder mildew fungus, for a total of 70 offspring. The data were originally described by [3] and further analyzed by [8]. Based on the model selection analysis described in [9], the model $[AD][AB][BE][CE][CF]$ fits the data well and has also a biological foundation. Out of 64 cells, only 22 are non-zero and most the entries are small counts.

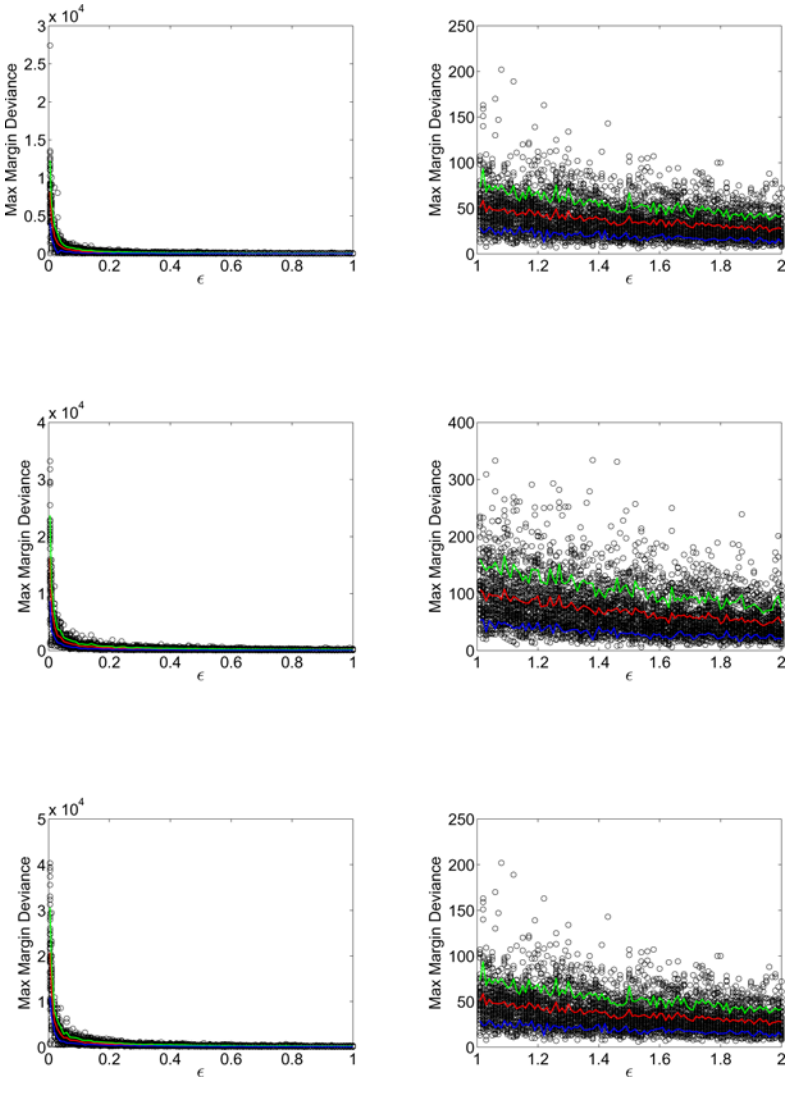


Fig. 2. Maximal L_1 difference between the true and perturbed margins for the Fungus table (top row), Czech autoworker table (middle) and Rochdale table (bottom). To improve readability, for each table, we split the plot in two parts, for $\epsilon < 1$ (left) and $\epsilon \geq 1$ (right). The three lines represent the mean plus or minus one standard deviation.

2. The data in Table 5 were collected in a prospective epidemiological study of 1841 workers in a Czechoslovakian car factory, as part of an investigation of potential risk factors for coronary thrombosis. See [10]. In the left-hand panel of Table 1, A indicates whether or not the worker "smokes", B corresponds to "strenuous mental work", C corresponds to "strenuous physical work", D corresponds to "systolic

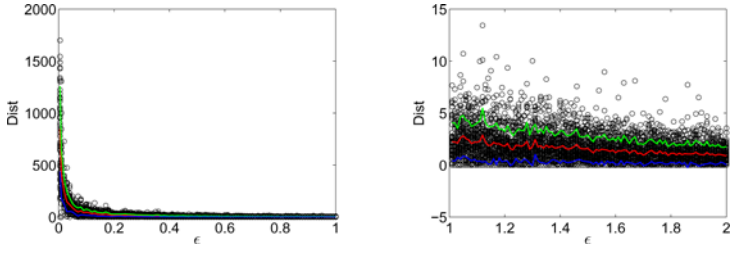


Fig. 3. Optimal values of b for the linear programming part of the algorithm of Table 1 as a function of ϵ for the fungus table

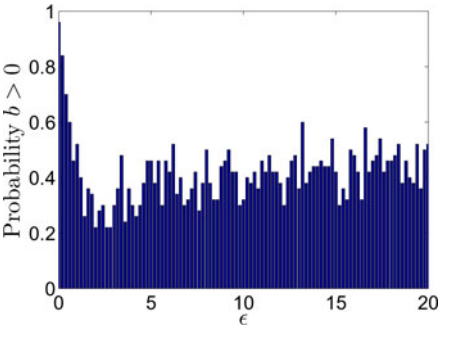


Fig. 4. Fraction of times the optimal value of b in the linear programming part of the algorithm of Table 1 was larger than 0 as a function of ϵ for the fungus table

blood pressure", E corresponds to "ratio of and lipoproteins" and F represents "family anamnesis of coronary heart disease". The model $[BF][ABCE][ADE]$ fits the data well. The cell counts are fairly large, with 14 cells having values of 5 or less.

3. The data in Table 6 involve 8 binary variables (Yes/No) relating women's economic activity and husband's unemployment from a survey of households in Rochdale [14, see page 279]. The 8 variables are: wife economically active (A); wife older than 38 (B); husband unemployed (D); child of age less than 4 (D); wife's education, high-school or higher (E); husband's education, high-school or higher (F); Asian origin (G); other household member working (H). The sample size is 665, and 165 of the 256 cells contain zero counts and 58 cells have positive counts of 4 or less.

For a grid of values of ϵ from 0.005 to 2, we perturbed each of the three tables 50 times using the algorithm of [1]. We summarize key results in a series of figures:

– Figure 1 shows the sample size of the perturbed tables as a function of ϵ. It is easy to see that the smaller ϵ the more variable the sample sizes of the perturbed tables become. In particular, when ϵ is very small, the sample size become unrealistically large, order of magnitudes larger than the true sample sizes. In fact, even for values

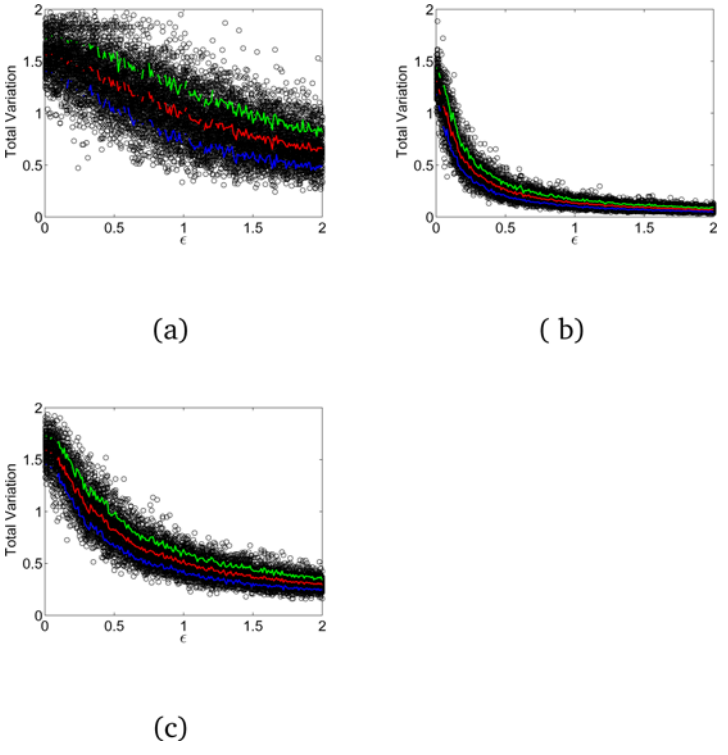


Fig. 5. Total variation distance between the MLE of the chosen model based on the original table and the MLE based on the perturbed tables as a function of ϵ the Fungus table (a), Czech autoworker table (b) and Rochdale table (c). The three lines represent the mean plus or minus one standard deviation.

of ϵ as large as 2 (which is a rather weak privacy guarantee) the sample size is highly variable—we deem this to be a serious problem for statistical analysis.

- Figure 2 shows the maximal L_1 distance between the margins of the true and perturbed tables as a function of ϵ. Once again, for values of ϵ as large as 5, these discrepancies are of the same order of magnitude as the sample size.
- Figure 3 shows the optimal values of b in the linear programming part of the algorithm of Table 1 for the Edward's fungus data as a function of ϵ.
- Figure 4 shows the proportion of times b is larger than 0, which means that there does not exists a real-valued non-negative tables whose margins match the margins of the perturbed table.
- Figure 5 shows the total variation distance between the MLE of the cell probabilities computed using the original distance with the MLE obtained from the perturbed margins, as a function of ϵ. Total variation distance is at most 2. To get a sense of how much the privacy mechanism effects the total variation distance, we computed this distance between the MLE of the cell probabilities based on the original table

and the uniform distribution over the cells for each of our three tables: Edwards–0.83, Czech–0.86, and Rochdale–1.43.

Space precludes a detailed analysis of the information summarized in these figures but we see a clear pattern even for the non-sparse Czech autoworkers example. As the noise level, controlled by the parameter ϵ, increases, the deviance between the generated tables and their MLEs is smaller. This means that if we add too much noise, we get strong privacy guarantees but inadequate and potentially misleading statistical inference. On the other hand, when we add little noise, the statistical inference is better but the differential privacy guarantees have little practical use.

7 Conclusions

We have explored the differential privacy approach to margin protection in contingency tables proposed by Barak et al. [1]. First we analyzed the theoretical claims and we discovered clear shortcomings. Second, we applied the methodology in a systematic fashion to three binary tables (Edwards fungus data, the Czech autoworkers data, and the data from Rochdale), in order to understand how the choice of the key noise parameter, ϵ, situates the methodology from the perspective of the risk-utility trade-off developed in the statistical literature on confidentiality. Through a simulation study for each of the three examples, we demonstrated what we deem to be serious problems with the methodology as originally proposed.

Differential privacy remains an attractive methodology because of its clear definition of privacy and the strong guarantees that it promises. But much is hidden in the noise parameter, ϵ, especially in the context of the proposed methods of Barak et al. Because differential privacy provides guarantees for the method and not for the specific data at hand, we do not believe the methodology is suitable for the type of large sparse tables often produced by statistics agencies and sampling organizations. Our preference remains for the less formal but seemingly effective approach described by Fienberg and Slavkovic [11], Dobra et al. [4] and Winkler [15].

Acknowledgement

This research was partially supported by Army contract DAAD19-02-1-3-0389 to Cylab at Carnegie Mellon University and in part by NSF grant DMS-0631589 and a grant from the Pennsylvania Department of Health through the Commonwealth Universal Research Enhancement Program to the Department of Statistics, Carnegie Mellon University.

References

1. Barak, B., Chaudhuri, K., Dwork, C., Kale, S., McSherry, F., Talwar, K.: Privacy, accuracy, and consistency too: A holistic solution to contingency table release. In: Proceedings of the 26th ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems (2007)

2. Bishop, Y.M., Fienberg, S.E., Holland, P.W.: Discrete Multivariate Analysis: Theory and Practice. MIT Press, Cambridge (1975); reprinted: Springer (2007)
3. Christiansen, S.K., Giese, H.: Genetic analysis of obligate barley powdery mildew fungus based on rfpl and virulence loci. Theoretical and Applied Genetics 79, 705–712 (1991)
4. Dobra, A., Fienberg, S.E., Rinaldo, A., Slavkovic, A.B., Zhou, Y.: Algebraic statistics and contingency table problems: Log-linear models, likelihood estimation, and disclosure limitation. In: Putinar, M., Sullivant, S. (eds.) Emerging Applications of Algebraic Geometry. IMA Series in Applied Mathematics, pp. 63–88. Springer, Heidelberg (2008)
5. Duncan, G.T., Fienberg, S.E., Krishnan, R., Padman, R., Roehrig, S.F.: Disclosure limitation methods and information loss for tabular data. In: Doyle, P., Lane, J., Theeuwes, J., Zayatz, L. (eds.) Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies, pp. 135–166. Elsevier, Amsterdam (2001)
6. Dwork, C.: Differential privacy. In: Bugliesi, M., Preneel, B., Sassone, V., Wegener, I. (eds.) ICALP 2006. LNCS, vol. 4052, pp. 1–12. Springer, Heidelberg (2006)
7. Dwork, C., McSherry, F., Nissim, K., Smith, A.: Calibrating noise to sensitivity in private data analysis. In: Halevi, S., Rabin, T. (eds.) TCC 2006. LNCS, vol. 3876, pp. 265–284. Springer, Heidelberg (2006)
8. Edwards, D.: Linkage analysis using log-linear models. Comp. Statist. and Data Anal. 13, 281–290 (1992)
9. Edwards, D.: Introduction to Graphical Modelling, 2nd edn. Springer, Heidelberg (2000)
10. Edwards, D., Havranek, T.: Fast procedure for model search in multidimensional contingency tables. Biometrika 72, 339–351 (1985)
11. Fienberg, S.E., Slavkovic, A.B.: A survey of statistical approaches to preserving confi- dentiality of contingency table entries. In: Aggarwal, C., Yu, P.S. (eds.) Privacy Preserving Data Mining: Models and Algorithms, pp. 289–310. Springer, Heidelberg (2008)
12. Lauritzen, S.L.: Graphical Models. Oxford University Press, Oxford (1996)
13. Wasserman, L., Shuheng, Z.: A statistical framework for differential privacy. J. Amer. Statist. Assoc. 105, 375–389 (2010)
14. Whittaker, J.: Graphical Models in Applied Multivariate Statistics. Wiley, Chichester (1990)
15. Winkler, W.: General Discret-data Modeling Methods for Producing Synthetic Data with Reduced Re-identification Risk that Preserve Analytic Properties. Research Report Series, Statistics 2010-02 (2008)

Appendix

Table 4. Cell counts 2^6 table involving genetic linkage in barley powder mildew fungus. Source: Edwards [8].

			1				2		D
			1		2		1	2	E
			1	2	1	2	1	2	F
1	1	1	0	0	0	3	0	1	0
		2	0	1	0	0	0	1	0
	2	1	1	0	1	0	7	1	4
		2	0	0	0	2	1	3	0
2	1	1	16	1	4	0	1	0	0
		2	1	4	1	4	0	0	0
	2	1	0	0	0	0	0	0	0
		2	0	0	0	0	0	0	0
A	B	C							

Table 5. Cell counts for Czech autoworker 2^6 table. Source: Edwards and Havranek [10].

			1				2				C
			1		2		1		2		B
			1	2	1	2	1	2	1	2	A
1	1	1	44	40	112	67	129	145	12	23	
		2	35	12	80	33	109	67	7	9	
	2	1	23	32	70	66	50	80	7	13	
		2	24	25	73	57	51	63	7	16	
2	1	1	5	7	21	9	9	17	1	4	
		2	4	3	11	8	14	17	5	2	
	2	1	7	3	14	14	9	16	2	3	
		2	4	0	13	11	5	14	4	4	
F	E	D									

Table 6. Rochdale table. Source: Whittaker [14].

				Y								N								H
				Y				N				Y				N				G
				Y		N		Y		N		Y		N		Y		N		F
				Y	N	Y	N	Y	N	Y	N	Y	N	Y	N	Y	N	Y	N	E
Y	Y	Y	Y	5	0	2	1	5	1	0	0	4	1	0	0	6	0	2	0	
			N	8	0	11	0	13	0	1	0	3	0	1	0	26	0	1	0	
		N	Y	5	0	2	0	0	0	0	0	0	0	0	0	0	0	1	0	
			N	4	0	8	2	6	0	1	0	1	0	1	0	0	0	1	0	
	N	Y	Y	17	10	1	1	16	7	0	0	0	2	0	0	10	6	0	0	
			N	1	0	2	0	0	0	0	0	1	0	0	0	0	0	0	0	
		N	Y	4	7	3	1	1	1	2	0	1	0	0	0	1	0	0	0	
			N	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	
N	Y	Y	Y	18	3	2	0	23	4	0	0	22	2	0	0	57	3	0	0	
			N	5	1	0	0	11	0	1	0	11	0	0	0	29	2	1	1	
		N	Y	3	0	0	0	4	0	0	0	1	0	0	0	0	0	0	0	
			N	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
	N	Y	Y	41	25	0	1	37	26	0	0	15	10	0	0	43	22	0	0	
			N	0	0	0	0	2	0	0	0	0	0	0	0	3	0	0	0	
		N	Y	2	4	0	0	2	1	0	0	0	1	0	0	2	1	0	0	
			N	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
A	B	C	D																	

Does Differential Privacy Protect Terry Gross' Privacy?

Krish Muralidhar [1] and Rathindra Sarathy [2]

[1] University of Kentucky, Lexington KY 40506, USA
[2] Oklahoma State University, Stillwater OK 74078, USA
krishm@uky.edu, rathin.sarathy@okstate.edu

Abstract. The concept of differential privacy was motivated through the example of Terry Gross' height in Dwork (2006). In this paper, we show that when a procedure based on differential privacy is implemented, it neither protects Terry Gross' privacy nor does it provide meaningful responses to queries. We also provide an additional illustration using income data from the US Census. These illustrations raise serious questions regarding the efficacy of using differential privacy based masking mechanism for numerical data.

Keywords: Differential privacy, Laplace noise addition, Numerical data.

1 Introduction

The concept of differential privacy was motivated through the example of Terry Gross' height in Dwork (2006, page 2) as follows:

> Suppose one's exact height were considered a highly sensitive piece of information, and that revealing the exact height of an individual were a privacy breach. Assume that the database yields the average heights of women of different nationalities. An adversary who has access to the statistical database and the auxiliary information "Terry Gross is two inches shorter than the average Lithuanian woman" learns Terry Gross' height, while anyone learning only the auxiliary information, without access to the average heights, learns relatively little.

Dwork (2006) then goes on to provide describe differential privacy and the Laplace based noise addition method to achieve the same. Although never explicitly stated, Dwork (2006) leaves the impression that the Laplace based noise addition would protect Terry Gross. But we never actually see the implications of using Laplace based noise addition and the level of protection it offers Terry Gross. In this paper, we carry this illustration to its natural conclusion to see the impact of using Laplace based noise addition to mask queries relating to the height of Lithuanian women, the extent to which it protects Terry Gross, and the implications of this approach for simple queries. We also provide an additional illustration using the incomes of individuals in the United States. These illustrations show that, for numeric data, the utility of the responses from a masking mechanism based on differential privacy is less than desirable.

J. Domingo-Ferrer and E. Magkos (Eds.): PSD 2010, LNCS 6344, pp. 200–209, 2010

2 Implementing a Differential Privacy Based Procedure

In the context of output perturbation, differential privacy is a standard which requires that the response to any query should be indistinguishable in the presence or absence of a single observation. An alternative description of this requirement would be as follows. Consider two databases D1 and D2 that differ in a single element. The response to any query from D1 should be "indistinguishable" from the response to the same query from D2 in a probabilistic sense if the responses satisfy the following requirement:

$$\frac{P[\kappa_f(D_1)=R]}{P[\kappa_f(D_2)=R]} \leq e^{\varepsilon}. \tag{1}$$

R is the response to the query from the system through the masking mechanisms $\kappa_f(D1)$ and $\kappa_f(D2)$ from D_1 and D_2, respectively (assuming, without loss of generality, that the larger probability is always in the numerator).

Furthermore, if the above requirement can be satisfied in the presence/absence of the most influential observation for a particular query, then this requirement will also be satisfied for any other observation. The impact of the most influential observation for a given query (Δf) can be assessed as follows:

$$\Delta f = \text{Max}\|f(D_1) - f(D_2)\| \tag{2}$$

for all possible realizations of D_1 and D_2, and where $f(D_1)$ and $f(D_2)$ represent the true responses to the query from D_1 and D_2. Dwork (2006) shows that if the response to any query is provided as f(X) + Laplace(0, b) where $b = \Delta f/\varepsilon$ (where X represents a particular realization of the database and f(X) represents the true response to the query), then such a response would satisfy equation (1). It is important to note that since equation (1) must be satisfied in the presence or absence of any observation, the evaluation of Δf must consider all possible realizations of D_1 and D_2; not just a particular realization of a database. In this sense, Δf represents the global sensitivity of the query. Please refer to Dwork (2006) and other papers for a more complete description of differential privacy and Laplace noise addition. We only consider the original definition of differential privacy (Dwork 2006). We do not consider any relaxations, such as found in Nissim et al. (2007), since they do not satisfy the original e^{ε} differential privacy.

3 "Terry Gross is Two Inches Shorter Than the Average Lithuanian Woman"

Lithuania has a population of approximately 3,400,000 with approximately 1,800,000 women. The average height of Lithuanian women is 167.5 cm or 66". Assume that we are Statistics Lithuania who has decided to respond to all queries regarding the height of Lithuanian women using Laplace based noise addition so as to satisfy differential privacy. For the purposes of this illustration, let us assume that Statistics Lithuania considers the height of a woman to be extremely sensitive and has set $\varepsilon = 0.01$.

In implementing Laplace based noise addition, we have to compute the maximum possible difference (Δf) that might possibly occur between two databases that differ in exactly one record for a particular query. Consider the simple Sum query. Let x_{min} represent the smallest possible value for a particular variable and let x_{max} represent the largest possible value for the same variable. For the Sum query, the maximum difference that could occur between two databases that differ in exactly one record $\Delta f = |x_{max} - x_{min}|$. For the purposes of this illustration, we will set $x_{min} = 23$" (the shortest woman according to the Guinness World Records) and $x_{max} = 98$" (the tallest woman according to the Guinness World Records) resulting in the global sensitivity $\Delta f = 75$. Hence, for all Sum queries, the scale parameter (b) of the Laplace noise distribution is set as $\Delta f/\varepsilon = 75/0.01 = 7500$ with the variance of the noise distribution = $(2 \times 7500^2) = 112500000$. Note that the number of records included in a particular Sum query does not affect the parameters of the Laplace noise distribution. Thus, whether the Sum query involves a single record or the entire population, noise is generated from a Laplace(0,7500) distribution. The response to any sum query would equal (The true value of the query + Noise from Laplace(0,7500)).

The Mean query is a simple variation of the Sum query where Mean = Sum/n. We can view this from two different ways. First we can view the Response to the Mean query as simply being the (Response to the Sum query/n). Alternatively, the noise generated for the Mean query would have noise generated from Laplace(0,7500/n) where n represents the number of records in the query. Either approach would result in exactly the same response distribution for the Mean query.

Now consider the response to the query "What is the average height of Lithuanian women?" Suppose the database responds with the true average height of Lithuanian women, namely 66. In Dwork (2006) this information can be used by an intruder to compromise Teryy Gross' height because the intruder has the auxiliary information that Terry Gross is 2 inches shorter than the average Lithuanina woman. To prevent this, the database has implemented Laplace noise addition and the distribution of responses to this query is provided in Figure 1. This figure shows that, even based on an extremely high level of security ($\varepsilon = 0.01$), the response to this query will be within (66" + 0.03") with probability close to 1. Once this response is received, the intruder knows that Terry Gross' height is very close to 64". Thus*, even with a very high level of security, differential privacy offers very little protection to Terry Gross since the intruder is able to estimate her height within 0.03"*.

Unfortunately, although little protection is offered to Terry Gross, the noise addition mechanism has a negative impact on other queries. Consider for instance, the following query: "What is the average height of women living in Smalininkai, Lithuania?" The city of Smalininkai has a total population of 621 with (let us say) 350 women. With a query involving 350 women, one would expect a reasonable answer to be within say + 1". The probability of observing a response within + 1" of the true average height is approximately 5%. The probability of observing a response within + 6" of the true average height is approximately 24% and the probability of being within + 12" is only 43%. In other words, in 57% of the cases, the response from the system would be outside 12" of the true height of 350 women. Clearly, users would consider such a response to be of little or no value.

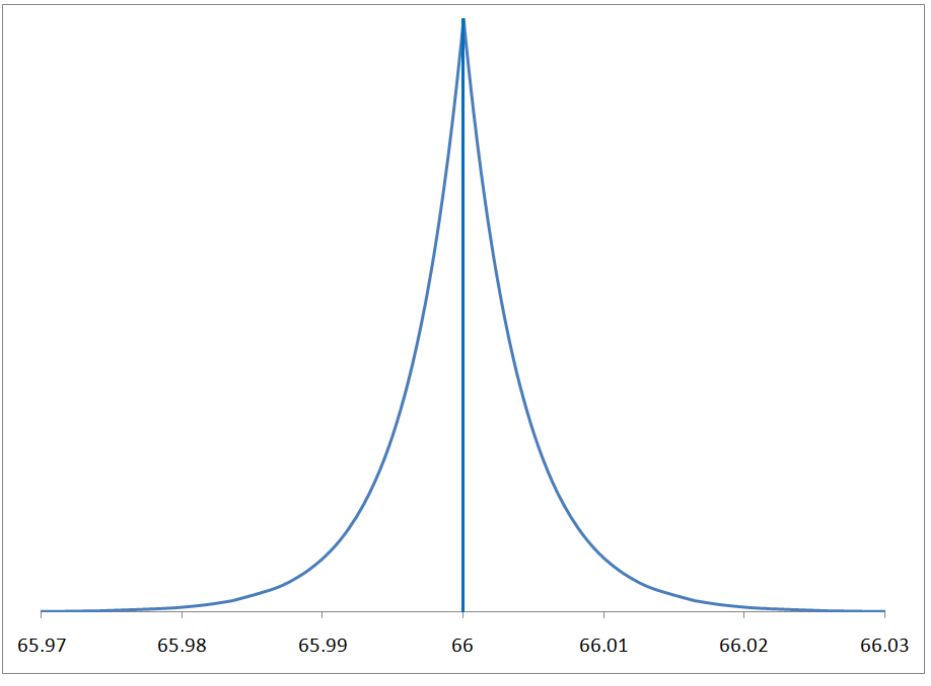


Fig. 1. Response distribution for the average height of Lithuanian women

Now consider the query, "What is the average height of all employed women living in Smalininkai, Lithuania?" Let us say that the number of women satisfying this query is 120. In this case only about 17% of the responses will be within 12" of the true value while the remaining 83% of the responses would fall outside the range. Consider the query "What is the average height of all employed women over age 50 living in Smalininkai, Lithuania?" Assume that the number of women who satisfy this query is 17. Only about 3% of the responses would fall within + 12" of the true value. Thus, in a vast majority of the cases for such queries, the response from the system would be practically useless. The probability of observing responses within specified limits is provided in Table 1 which clearly shows that for small subsets, Laplace based noise addition provides very little utility.

Another interesting aspect of Table 1 is the last column in the table which shows the percentage of cases where the response from the system is within the range of 23" to 98". Even for n = 350, 18% of the responses are either below 23" or greater than 98", which from a practical perspective makes no sense at all. With n = 120, a majority (55%) of the responses are not even within the (23" to 98") range. With n = 17, we get the ridiculous situation where 92% of the responses are outside this meaningful range. To illustrate this further, consider the distribution of the responses to the query "What is the average height of all employed women in Smalininkai?" is provided in Figure 2. For the purposes of this illustration, we have assumed the true height of the 120 women in Smalininkai who are employed to be 66". Note that for most real life

Table 1. An assessment of the utility of the responses

Group	n	Probability that the Response is within ± k" of the True Value							Probability that Response is Reasonable (between 23" and 98")
		0.10"	0.50"	1"	2"	4"	6"	12"	
All Lithuanian women	1800000	100%	100%	100%	100%	100%	100%	100%	100%
All women in Smalininkai	350	0%	2%	5%	9%	17%	24%	43%	82%
All employed women in Smalininkai	120	0%	1%	2%	3%	6%	9%	17%	45%
All employed women in Smalininkai over age 50	17	0%	0%	0%	0%	1%	1%	3%	8%

purposes, a subset of size 120 would be considered large and one would expect a reasonably close estimate of the true height.

The response distribution in Figure 2 shows the lack of utility from the Laplace based noise addition method. As observed earlier, a majority of the responses (55%) fail the simple common sense test since they are not even within the upper and lower limits for height of women. Approximately 17% or the responses will result in average height less than zero, which is clearly unacceptable.

There is a curious contradiction when using Laplace based noise addition to satisfy differential privacy. In order to satisfy differential privacy it is necessary for the variable to be bounded. Yet, when we implement Laplace based noise addition, the resulting responses are unbounded! Thus, we are left with the contradictory situation of having to make the assumption that height of women must be between (23" and 98"), but many of the responses are outside this range.

The irony is that the very high security specification was necessary to protect Terry Gross. Yet, the resulting procedure offers little security to Terry Gross, and the intruder is able to estimate Terry Gross' true height to within 0.03" accuracy. Unfortunately, the resulting response distribution is so poor that for almost, if not all, subsets, the responses have no practical utility as far as the user is concerned. Note that our illustration is a very reasonable one assuming a variable (height of women) that has a natural lower and upper bound. Furthermore, the bounds are reasonably close to the average height which, one would expect, would result in reasonable responses. This, however, is not the case; the responses for even what would be considered as large subsets by most practical standards, are of little value to the user.

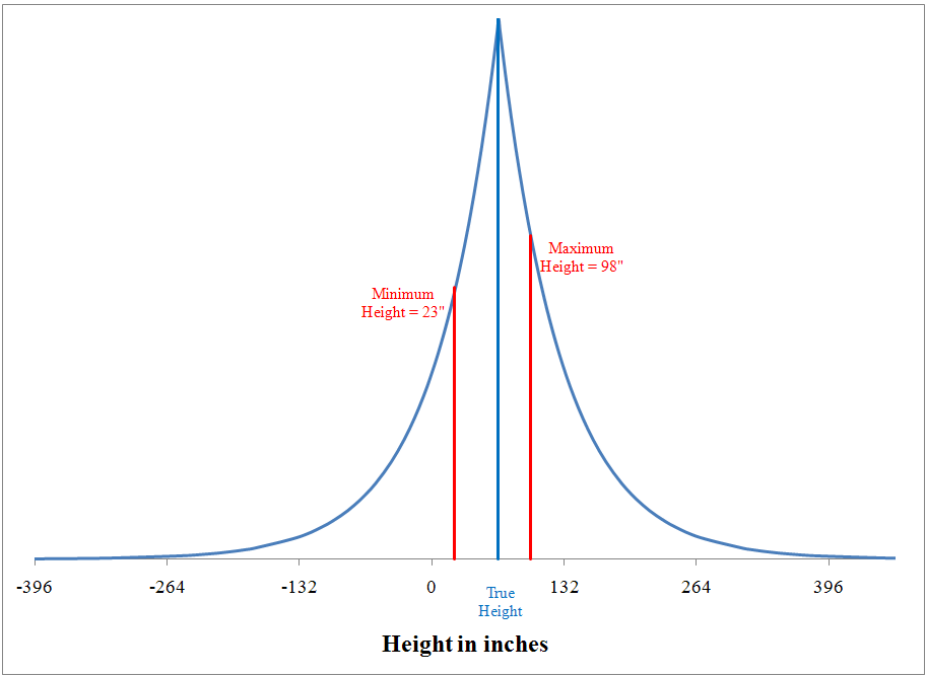


Fig. 2. Distribution of responses to the query "What is the average height of all employed women in Smalininkai?"

4 "Mr. Overy Rich's Income Is $5 Million More Than the Average American"

Consider the situation where the auxiliary information available is the following: "Mr. Overy Rich's income is $5 million more than the average American." Let us also assume that the variable Income is confidential. Let us also assume a security level of say $\varepsilon = 0.25$. It is well known that the income of some hedge fund managers exceed $1 billion[1]. In order to protect such individuals, it is necessary that Δf must be at least 1 billion. Note that, in order to satisfy differential privacy, it is better to be conservative in estimating Δf. For the purposes of this illustration, let us assume that Δf = 1,000,000,000. For this illustration, all information was gathered from the 2006-2008 American Community Survey at the U.S. Census Bureau web site.

There were a total of 97,488,418 individuals employed fulltime in the US with a mean income of $54,698. Assume that Laplace based noise addition is implemented to mask this data. Based on the specifications above, the range of the responses

[1] "James H. Simons, a former math professor who has made billions year after year for the hedge fund Renaissance Technologies, earned $2.5 billion running computer-driven trading strategies. John A. Paulson, who rode to riches by betting against the housing market, came in second with reported gains of $2 billion. And George Soros, also a perennial name on the rich list of secretive moneymakers, pulled in $1.1 billion." (Story, 2009)

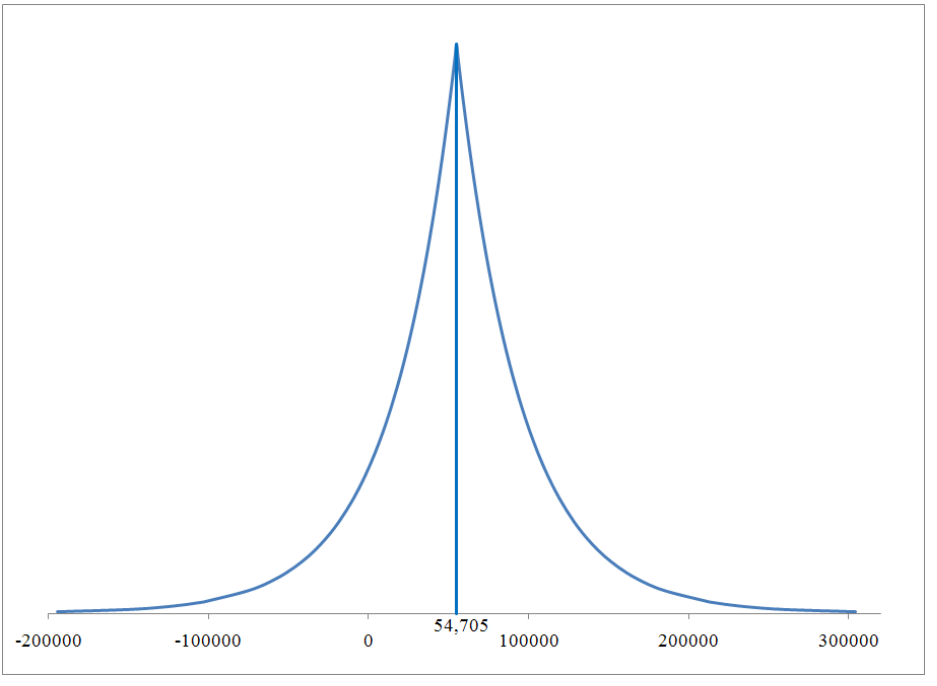


Fig. 3. Distribution of responses for the average income of Fayette County

from 0.1 percentile to the 99.9 percentile would be \$54,443 to \$54,953. In other words, in 99.8% of the cases, the response from the system would range between \$54,698 $\pm$ \$255. Hence an intruder would be able to estimate Mr. Overy Rich's true income as being between \$5,054,443 and \$5,054,953. From a practical perspective, this offers little protection to Mr. Overy Rich since we are able to estimate his income with a margin of error of about \$500 which is very small compared to his income. Thus, with the auxiliary information available to the intruder, differential privacy offers very little protection to Mr. Overy Rich.

By contrast, consider the legitimate query, "What is the average income of adults in Fayette County, Kentucky?" There were a total of 99,683 individuals employed fulltime in Fayette County with an average income of \$54,705. Assume that the response to this query is provided through the Laplace noise addition approach. In this case, the 0.1% of the response distribution would be –\$194,670 and 99.9% of the response distribution would be \$304,080. In other words, the range of the responses would be \$54,705 $\pm$ \$249,375. The distribution of the responses in this case is provided in Figure 3.

For a legitimate user, given that this subset consists of almost 100,000 individuals, it is reasonable to expect the response for the mean income to be very close to the true mean income. At most, one would expect a difference of say \$1000. But with Laplace noise addition, 98% of the responses would fall outside the $\pm$ 1000 range. In fact, 78% of the responses would fall outside the $\pm$ 10000 range. Approximately 12% of the

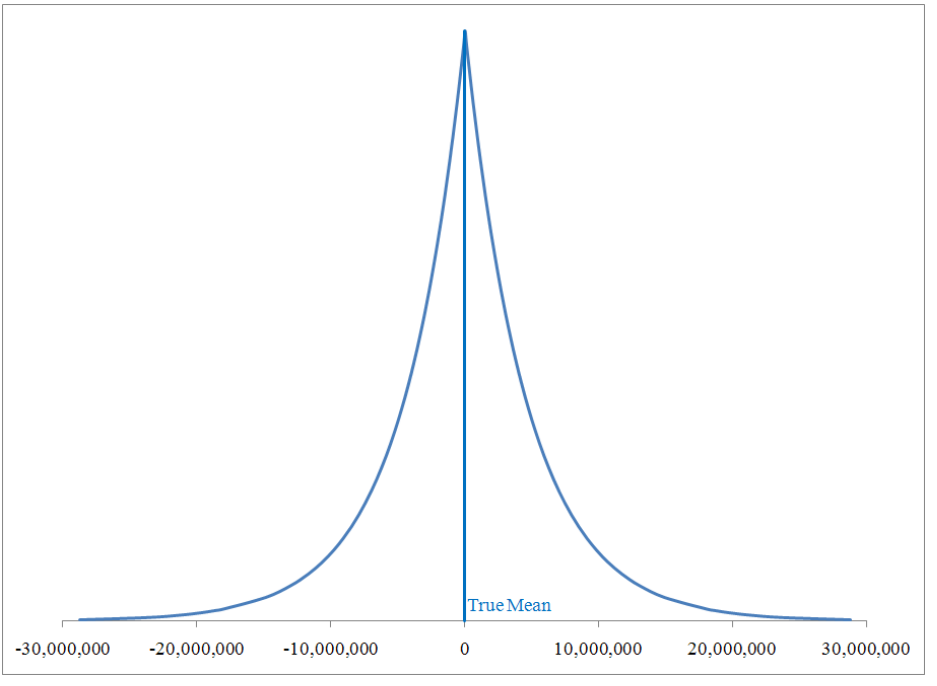


Fig. 4. Distribution of responses for the average income of women in Robertson County ($\varepsilon = 0.25$)

responses would result in average income less than zero. Approximately 17% of the responses would result in average income greater than $100,000, almost twice as much as the true average income. In essence, Laplace based noise addition offers little or no utility to the legitimate user even the size of the subset is relatively large (nearly 100,000).

Consider another similar query "What is the average income of adult women in Robertson County, Kentucky?" In this county, there were a total of about 863 women with an average income of $35,200. For this query, the (0.1%, 99.9%) range of responses would be between (– $28,769,472 and $28,839,872), which from the perspective of a legitimate user, is completely meaningless. Figure 4 provides the range of responses for this query. Only about 1% of the responses are within the range (0 to 100,000). Approximately 49.6% of the observations are below zero. Less than 1% of the responses are in the range $35,200 \pm 35,200$. These types of responses may be justified if the subset is very small. In this case, the size of the subset is, in practical terms, rather large (863). Even for a single record, it makes no sense to provide responses of the order of millions of dollars when the true value is only $35,200. In summary, given the true value is $35,200, when the vast majority responses are in the millions of dollars they offer no utility at all to the legitimate user.

For the purposes of illustration, let us consider the case where $\varepsilon = 10$. Note that this specification offers practically no security at all since $e^{\varepsilon} = e^{10} = 22026$. In other

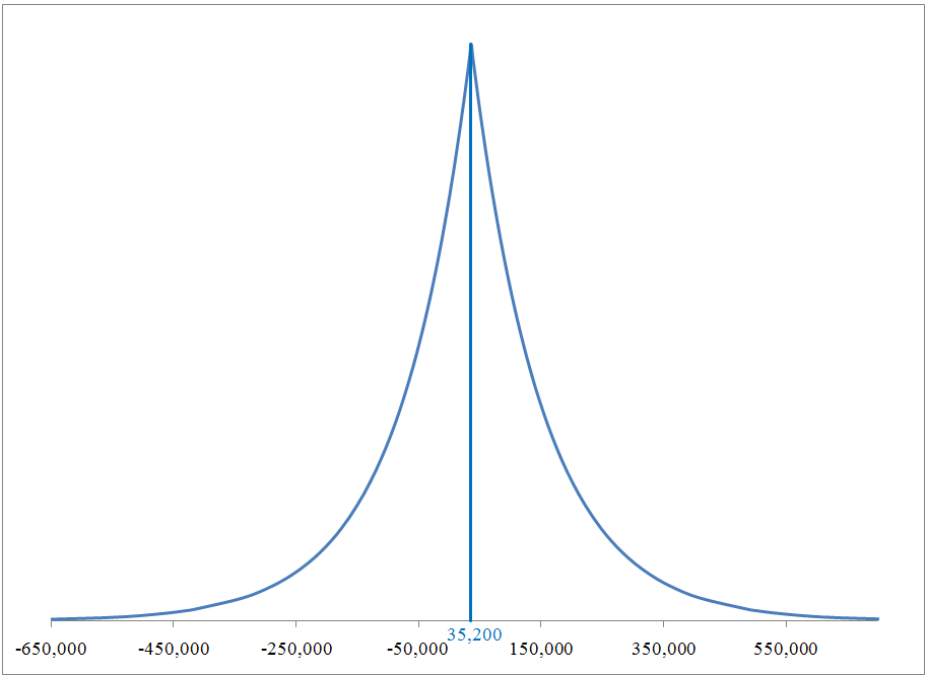


Fig. 5. Distribution of responses for the average income of women in Robertson County ($\varepsilon = 10$)

words, with this specification, the intruder's knowledge gain is of the order of 22026. The implication of this specification is simply that this specification offers practically no security at all. Unfortunately however, even though this specification offers no security, it does not improve the utility of the responses.

Figure 5 shows the distribution of responses for the average income of women in Robertson County with $\varepsilon = 10$. Note that the responses still range between (–658,000 to $755,000). Even with this low level of security, 35% of the responses are negative. More than 70% of the responses are outside the range 35,200 $\pm$ 35,200. Thus, even though ε is very large, the utility based on the responses is still very low and the entire procedure offers practically no security at all. Thus, we are left in the unenviable position of no security and not utility. This example also illustrates the fact that the value of ε makes little difference when Δf is large since the variance of the noise term will be dominated by the value of Δf. In these situations, the Laplace based noise addition will offer little or no utility regardless of the value of ε.

5 Conclusions

Differential privacy is being offered as a procedure for protecting privacy of records for all types of data. Unfortunately, the discussion on differential privacy is almost always limited to a theoretical discussion with a few minor exceptions where

illustrations for well behaved count data are provided (large cell count values, no sparse or empty cells, etc.). There are no practical examples of the application of Laplace based noise addition based on global sensitivity to satisfy differential privacy for numeric data. The real life numerical examples used in this study clearly show that the implementation of the Laplace based noise addition procedure in practice is likely to result in a situation where little disclosure protection is offered for large subsets and little utility is offered for small subsets. The behavior of the Laplace based noise addition procedure for numerical data also leads us to believe that in real life count data (with a large number of cells, sparse cells, zero valued cells, etc.), the Laplace based noise addition is likely to result in similar outcomes (where little security is provided for cells with large counts and little utility is provided for cells with small counts).

It is also important to note that these issues are the result of the inherent requirements of differential privacy. Differential privacy rests on this basic premise: "If I can make the two most extreme values for any query indistinguishable within a factor e^{ε} then all other values will also be indistinguishable." The problem with this approach is that when the magnitude of the difference between the two extreme values (global sensitivity Δf) is very large in relation to the variance of the dataset, the noise added is so large so as to make all responses to the query meaningless. This is clearly illustrated in our examples. Unfortunately, almost all economic data tend to heavily skewed and, in practice, it is likely that very large Δf values are the rule rather than the exception.

In summary, for numerical data, like Dalenius' (1977) definition of privacy before it, differential privacy is an interesting concept, but of little value in practice.

References

1. American Factfinder, U.S. Census Bureau, http://factfinder.census.gov/home/
2. Dalenius, T.: Towards a Methodology for Statistical Disclosure Control. Statistisktidskrift 5, 429–444 (1977)
3. Dwork, C.: Differential Privacy. In: Bugliesi, M., Preneel, B., Sassone, V., Wegener, I. (eds.) ICALP 2006. LNCS, vol. 4052, pp. 1–12. Springer, Heidelberg (2006)
4. Nissim, K., Raskhodnokova, S., Smith, A.: Smooth Sensitivity and Sampling in Private Data Analysis. In: Proceedings of the thirty-ninth annual ACM symposium on Theory of computing, pp. 75–84 (2007)
5. Story, L.: Top Hedge Fund Managers Do Well in a Down Year. New York Times, March 24 (2009), http://www.nytimes.com/2009/03/25/business/25hedge.html

Some Additional Insights on Applying Differential Privacy for Numeric Data

Rathindra Sarathy[1] and Krish Muralidhar[2]

[1] Oklahoma State University, Stillwater OK 74078, USA
[2] University of Kentucky, Lexington KY 40506, USA
rathin.sarathy@okstate.edu, krishm@uky.edu

Abstract. Recently Sarathy and Muralidhar (2009) provided the first attempt at illustrating the implementation of differential privacy for numerical data. In this paper, we attempt to provide further insights on the results that are observed when Laplace based noise addition is used to protect numerical data in order to satisfy differential privacy. Our results raise serious concerns regarding the viability of differential privacy and Laplace noise addition as appropriate procedures for protecting numerical data.

Keywords: Differential privacy, Laplace noise addition, Numerical data.

1 Introduction

Differential privacy was initially proposed by Dwork (2006) as a privacy standard that can be used for any type of data. Laplace noise addition was proposed in the same study as a procedure for satisfying the differential privacy standard. However, most of the illustrations of differential privacy and Laplace noise addition focus almost exclusively on count data. Sarathy and Muralidhar (2009) described a simple implementation of the concept of differential privacy and the associated Laplace noise addition for continuous numerical data. It was subsequently pointed out to us that we had used "local sensitivity" rather than "global sensitivity" in that implementation. We have been criticized for failing to acknowledge this slightly different implementation of differential privacy. While we acknowledge that we used local sensitivity in Sarathy and Muralidhar (2009), in this paper, we explain why that was our only choice. In addition, we also illustrate that the Laplace based noise addition procedure to satisfy differential privacy is vulnerable to a tracker style attack.

2 Implementing a Differential Privacy Based Procedure

Differential privacy is simply a privacy standard whereby the response to any query including or excluding a particular observation is indistinguishable in a probabilistic sense. One mechanism that satisfies differential privacy is the Laplace based noise addition approach which we discuss later. In general terms, the requirement for differential

J. Domingo-Ferrer and E. Magkos (Eds.): PSD 2010, LNCS 6344, pp. 210–219, 2010.

privacy can be described as follows (Dwork 2006). Consider *any two possible datasets D_1 and D_2 that differ by exactly one record*. Let $\kappa_f(D_1)$ and $\kappa_f(D_2)$ be the responses from datasets D_1 and D_2, respectively, where $\kappa_f()$ represents a mechanism used to respond to an arbitrary query f(). For $\kappa_f()$ to satisfy differential privacy, it is necessary that

$$e^{-\varepsilon} \leq \frac{P[\kappa_f(D_1)=R]}{P[\kappa_f(D_2)=R]} \leq e^{\varepsilon}. \quad (1)$$

where R represents the response. A simple interpretation would be that the ratio in the middle represents the "knowledge gain ratio" for an intruder from one version of the database (D_1) over the other (D_2) assuming without loss of generality that the numerator is always larger. Differential privacy requires that the knowledge gain ratio be limited to e^{ε}. Note that differential privacy is predicated on a query/response framework. Since it would be very difficult (if not impossible) to show that differential privacy will be satisfied for every type of analyses that an intruder might employ when microdata is released, adoption of differential privacy precludes the release of masked microdata.

We would also like to briefly address the use of the term "ε differential privacy" when referring to a procedure that satisfies equation (1). In reality, a procedure that satisfies the ratio in equation (1) actually provides "e^{ε} differential privacy" since the intruder's knowledge gain is e^{ε}. In our opinion, a procedure that satisfies "ε differential privacy" should satisfy the requirement specified by Chaudhuri and Monteleoni (2008) which is as follows:

$$\frac{P[\kappa_f(D_1)=R]}{P[\kappa_f(D_2)=R]} \leq 1 + \varepsilon. \quad (2)$$

It is true that when ε is zero or very small, $e^{\varepsilon} \approx (1 + \varepsilon)$. However, the two measures diverge quite quickly. Even when $\varepsilon = 0.20$, e^{ε} exceeds $(1 + \varepsilon)$ by at least a non-negligible 10%. Considering that we see specifications of ε much larger than 0.20[1], special care is needed to see that users do not overestimate the level of security provided by equation (1). In order to avoid any confusion, we suggest the use of the term "e^{ε} differential privacy" when security is evaluated as shown in equation (1) and the term "ε differential privacy" when security is evaluated using equation (2). Since Dwork (2006) definition in as shown in equation (1), for the remainder of the paper, we will be using the term "e^{ε} differential privacy."

3 Laplace Noise Addition to Satisfy Differential Privacy

In order to satisfy differential privacy, Dwork (2006) suggests the use of Laplace based noise addition, again assuming a query/response situation. Assume that the intruder issues the query f(X) on a data set X for which the true response is *a*. Let a

[1] For example when "ε = 2" as in Abowd and Vilhuder (2008), the true knowledge gain ratio equals $e^2 = 7.389$. Similarly, if the "overall ε for the procedure was 8.6" as in Machanavajjhala et al. (2008), the true knowledge gain is $e^{8.6} = 5432$.

differential privacy satisfying mechanism $\kappa_f()$ be implemented for this data set and that the response from the system is R. Dwork (2006) suggests that the masked response $\kappa_f(X) = R = a + y$ where y represents a noise term from a Laplace distribution with mean 0 and scale parameter $b = \Delta f/\varepsilon$ where Δf *represents the maximum difference in the value of f(X) when exactly one input to X is changed. This accounts for the situation, for example, when the intruder's data differs from that of the data set X by exactly one record.*

The value of Δf represents the global sensitivity for the query f(X). To determine Δf one must consider all possible values for D_1 and D_2 in the population of values **D** and not just the specific values that may exist in the current dataset X that is being protected. Hence, in order to implement the globally sensitive version of the Laplace noise addition procedure, it is necessary to determine the value of Δf.

In binary databases where the queries are always assumed to be "count" queries, it is easy to see that the value of $\Delta f = 1$ since the maximum difference in the count between D_1 and D_2 is always 1. For numeric data, it is not even certain that we can determine the global sensitivity for an arbitrary dataset **D** and an arbitrary query f(X). Consider for instance, a numerical variable such as insurance claim that was used by Sarathy and Muralidhar (2009). Let us also assume that a simple sum query was issued. Since we have to protect the universe of possible insurance claims, it would be necessary to determine the global sensitivity of the insurance claim variable for the sum query. But there is no simple approach to do this. Even assuming that insurance claim is a positive variable, we now have to answer the question "What is the largest possible insurance claim that could exist in the universe of insurance claim values?" Without answering this question, it is simply impossible to implement any procedure that satisfies differential privacy (not even Laplace based noise addition).

Thus, for the insurance claim example in Sarathy and Muralidhar (2009), there is no way to determine Δf. Without the value of Δf, a globally sensitive Laplace based noise addition cannot be implemented. There are only two alternatives. The first is to acknowledge that differential privacy simply cannot be used in this situation, but that serves no purpose in a paper intended to illustrate the application of differential privacy. The second is to use local sensitivity to illustrate the application, an implementation choice made in Sarathy and Muralidhar (2009)[2]. Wasserman and Zhou (2009) also note this issue when they state that "In particular, it is difficult to extend differential privacy to unbounded domains."

The conclusion that we can reach from this discussion is that, in order to satisfy differential privacy, *it is necessary that the upper and lower bounds on the values in the database exist and are known. Without this knowledge, it is impossible to compute* Δf *for an arbitrary function f() and impossible to implement the Laplace based noise addition procedure to satisfy differential privacy.* One other approach is to arbitrarily impose lower and upper bounds (bottom and top coding). This is a convenient solution, but in our opinion, defeats the very purpose of differential privacy whose primary objective is to ensure that even the extreme values are protected while

[2] In evaluating whether differential privacy is satisfied, we only consider the original definition of differential privacy (Dwork 2006). We do not consider any relaxations such as found in Nissim et al. (2007) as satisfying true e^{ε} differential privacy.

simultaneously providing meaningful responses. Further, top/bottom coding may not even satisfy differential privacy.

As discussed earlier, practically every illustration of differential privacy deals with count data where computing Δf is not a problem. However, when we are dealing with numeric variables, it is more than likely that the variable is unbounded (or at least the true bounds are not known). This is particularly true for most economic variables such as insurance claims, income, revenue, cost, profit, and the list goes on. For these variables, Sarathy and Muralidhar (2009) have shown that implementing Laplace noise addition based on local sensitivity does not satisfy differential privacy. In summary, *differential privacy with global sensitivity can be implemented using Laplace noise addition if and only if the variable under consideration has known bounds*.

4 Vulnerability to Tracker Attack

In investigating differential privacy in more detail, we find that even when the numerical data has known bounds, Laplace noise addition has limitations. For the purpose of this illustration, let us assume that the data set consists of real numbers between the limits 0 and 1. For simplicity and without loss of generality, we will limit our discussion to the Sum query and $\varepsilon = 1$. For this data set the global sensitivity $\Delta f = 1$. In order to satisfy e^{ε}-differential privacy, the noise term for the Sum queries must be generated from a Laplace distribution with mean 0, scale parameter $b = \Delta f/\varepsilon = 1$ (and resulting noise variance = $2b^2 = 2$). Consistent with the description on Dwork (2006), we assume that the details of the perturbation (ε, Δf, etc.) are provided to the users. We generated a dataset **X** with 50 observations and we assume that the intruder has 49 observations (2, 3, …, 50) in the dataset (Table 1). The intruder's objective is to estimate the unknown observation x_1.

Note that the variance of the entire data set provided in Table 1 is 0.09339. However, in order to satisfy differential privacy, it is necessary that the variance of the noise distribution be 2. The key issue here is that unlike traditional noise addition where the noise added is proportional to the variance in the data set, the noise added to satisfy differential privacy is independent of the actual data set and is based only on the value of Δf. Consequently, the level of noise added to satisfy differential privacy can be of orders of magnitude greater than the variance in the data set, reducing the utility of the responses. The problem can be further magnified when the data set is skewed resulting in very large Δf and very large noise variance compared to the variance of the data set. In addition, since the range of the variable **X** is between (0, 1), the user knows that the sum of $(x_i + x_j)$ must be in the range (0, 2). However, due to the large noise variance and the fact that the Laplace distribution is unbounded, a large proportion of the responses (21 out of 49) are outside the meaningful range of (0, 2). This is a problem for most datasets.

The intruder issues the following series of queries $(x_1 + x_2)$, $(x_1 + x_3)$, …, $(x_1 + x_{50})$ resulting in a total of 49 queries. Let $a_2, a_3, \ldots, a_{50}$ represent the true response to the queries, respectively. Let $y_2, y_3, \ldots, y_{50}$ represent the noise terms generated from a

Table 1. Data and computations for the example

Individual	x	$(x_1 + x_i)$	Random #	y_i	$R_i = (x_1 + x_i) + y_i$	Estimate of x_1
1	0.97032					
2	0.85490	1.82522	0.52770	0.05700	1.88222	1.02732
3	0.72936	1.69968	0.07910	-1.84387	-0.14419	-0.87355
4	0.06435	1.03467	0.43122	-0.14799	0.88668	0.82233
5	0.42397	1.39429	0.79758	0.90427	2.29856	1.87459
6	0.75934	1.72966	0.89064	1.52000	3.24966	2.49032
7	0.67422	1.64454	0.52139	0.04372	1.68826	1.01404
8	0.62075	1.59107	0.93254	2.00312	3.59419	2.97344
9	0.66039	1.63071	0.26885	-0.62044	1.01027	0.34988
10	0.54600	1.51632	0.70928	0.54225	2.05857	1.51257
11	0.22039	1.19071	0.50533	0.01072	1.20143	0.98104
12	0.98132	1.95164	0.56455	0.13822	2.08986	1.10854
13	0.22174	1.19206	0.76173	0.74119	1.93325	1.71151
14	0.88548	1.85580	0.92787	1.93619	3.79199	2.90651
15	0.95191	1.92223	0.18962	-0.96957	0.95266	0.00075
16	0.65780	1.62812	0.80940	0.96445	2.59257	1.93477
17	0.88826	1.85858	0.63536	0.31569	2.17427	1.28601
18	0.74429	1.71461	0.73987	0.65344	2.36805	1.62376
19	0.12368	1.09400	0.35452	-0.34386	0.75014	0.62646
20	0.59708	1.56740	0.10678	-1.54385	0.02355	-0.57353
21	0.03746	1.00778	0.76936	0.77376	1.78154	1.74408
22	0.82311	1.79343	0.14070	-1.26801	0.52542	-0.29769
23	0.03147	1.00179	0.81132	0.97456	1.97635	1.94488
24	0.32822	1.29854	0.40197	-0.21823	1.08031	0.75209
25	0.20763	1.17795	0.87744	1.40601	2.58396	2.37633
26	0.57210	1.54242	0.50602	0.01211	1.55453	0.98243
27	0.66724	1.63756	0.07235	-1.93313	-0.29557	-0.96281
28	0.36904	1.33936	0.26919	-0.61921	0.72015	0.35111
29	0.83805	1.80837	0.79157	0.87498	2.68335	1.84530
30	0.72112	1.69144	0.85190	1.21672	2.90816	2.18704
31	0.98357	1.95389	0.83221	1.09189	3.04578	2.06221
32	0.23028	1.20060	0.56917	0.14890	1.34950	1.11922
33	0.09613	1.06645	0.04275	-2.45934	-1.39289	-1.48902
34	0.00538	0.97570	0.13981	-1.27435	-0.29865	-0.30403
35	0.46984	1.44016	0.56482	0.13886	1.57902	1.10918
36	0.96043	1.93075	0.31687	-0.45610	1.47465	0.51422
37	0.20283	1.17315	0.22134	-0.81491	0.35824	0.15541
38	0.60845	1.57877	0.30995	-0.47820	1.10057	0.49212
39	0.45104	1.42136	0.13421	-1.31520	0.10616	-0.34488
40	0.63254	1.60286	0.95061	2.31488	3.91774	3.28520
41	0.49287	1.46319	0.62664	0.29205	1.75524	1.26237
42	0.47326	1.44358	0.87987	1.42603	2.86961	2.39635
43	0.87437	1.84469	0.17758	-1.03516	0.80953	-0.06484
44	0.01190	0.98222	0.86866	1.33681	2.31903	2.30713
45	0.89823	1.86855	0.53223	0.06663	1.93518	1.03695
46	0.54942	1.51974	0.42995	-0.15095	1.36879	0.81937
47	0.03274	1.00306	0.03394	-2.68988	-1.68682	-1.71956
48	0.61882	1.58914	0.36316	-0.31976	1.26938	0.65056
49	0.30366	1.27398	0.17065	-1.07496	0.19902	-0.10464
50	0.33108	1.30140	0.40338	-0.21474	1.08666	0.75558
					Average Of Estimates	0.97262

Laplace(0,b) to the queries. Let $R_2 = (x_1 + x_2) + y_2$, $R_3 = (x_1 + x_3) + y_3$, …, $R_{50} = (x_1 + x_{50}) + y_{50}$ represent the responses from the system to the queries. These values are provided in Table 1 as well. Since the intruder knows the true values of x_2, x_3, …, x_{50}, the intruder can simply subtract the respective value from the response to result in an estimate of x_1 as follows:

$$\hat{x}_1^i = R_i - x_i, \quad i=2, 3, ..., 50.$$

For example, $R_2 = 1.88222$, $x_2 = 0.85490$, and hence $\hat{x}_1^2 = 1.02732$. Table 1 provides the results of all 49 queries. The resulting variable from this process $\hat{x}_1$ is an iid Laplace(x_1,b) random variable.

Now consider $\bar{x}_1$the mean of ($\hat{x}_1$). Let q represent the number of queries issued. In this illustration q = 49. From the central limit theorem, when q is large, we know that $\bar{x}_1 \sim$ Normal(x_1, $2b^2/q$). Even when q is small, since $\hat{x}_1$is IID Laplace(0,b), we know that the mean and variance of $\bar{x}_1$ are x_1 and $2b^2/q$, respectively, although we do not know the exact distribution of $\bar{x}_1$. Thus, even when q is small, the variance of the resulting estimate is smaller by a factor of q compared to the variance of the original Laplace distribution.

The intruder now simply estimates the true value of x_1 as the mean of $\hat{x}_1^i, i = 2, 3, \ldots, 50$. From the data in Table 1, we know that $\bar{x}_1^{Est} = \sum \frac{\hat{x}_1^i}{49} = 0.97262$. Now consider the following probabilities:

$$P[\bar{x}_1^{Est} \geq 0.97262 \mid x_1 = 1] = P\left[\bar{x}_1^{Est} \geq 0.97262 \mid \text{Normal}\left(1, \frac{2}{49}\right)\right] = 0.575992308, \text{and}$$

$$P[\bar{x}_1^{Est} \geq 0.97262 \mid x_1 = 0] = P\left[\bar{x}_1^{Est} \geq 0.97262 \mid \text{Normal}\left(0, \frac{2}{49}\right)\right] = (5 \times 10^{-12}).$$

If we now consider the ratio of the two probabilities above, we get $\frac{0.575992308}{(5 \times 10^{-12})} = (1.17 \times 10^{11}) \gg e^1 = 2.7182$. Thus, the ratio of the two probabilities does not satisfy the requirements of e^{ε} differential privacy (but does satisfy $e^{q\varepsilon}$ differential privacy as we discuss later).

Note that 0.97262 is an excellent point estimate of x_1 (0.97032) for such a relatively small sample size of 49. For more realistic situations such estimates are expected to be very close to the true value and would result in extremely sharp interval estimates with high confidence. By most standards of statistical disclosure control, this would be considered an unacceptable breach of confidentiality and privacy. Even with only 50 observations, if we assume that the intruder has 49 observations, the intruder is not limited to 49 queries. The intruder can also issue all possible combinations of queries involving x_1 and the remaining known observations. One such possible query is (x_1 + x_2 + x_3). From the response to this query, the intruder can get the estimate of x_1 simply as the Response – (x_2 + x_3). Even when n is relatively small, the intruder can issue a very large number of queries to the system in this manner, resulting in very large q. When q is very large, $(2b^2/q) \rightarrow 0$ and $\bar{x}_1^{Est} \rightarrow x_1$. Thus, with increasing q, the intruder gets a very accurate estimate of the true value of x_1. Since this result is true for x_1, similarly we can show it to be true for any value x_i.

In summary, when the intruder has information regarding the (n – 1) observations, they can use this information to issue a series of (tracker) queries in order to estimate the value of the missing observation with a great deal of accuracy. This type of phenomena has been addressed previously in the statistical disclosure limitation literature by Denning et al. (1979), Duncan and Mukerjee (2001), and others. It is interesting that this result is also consistent with the observations of Dinur and Nissim (2004) who showed that an intruder, *with no prior information*, given an unlimited number of queries, can reconstruct the value of the entire database. What these results indicate is that, when we assume that the intruder has (n – 1) of the n observations, then *only the first query will provide the desired level of security. All subsequent queries will result in a reduction in security*. For the q^{th} query, the level of security provided is only $e^{q\varepsilon}$ and not e^{ε} (Dwork and Smith 2009).

It can be argued that the variance of the noise term can be increased with the number of queries to compensate for the reduction in security. Since the security provided for the q^{th} query is $e^{q\varepsilon}$, in order to achieve the same level of security for the q^{th} query as for the first query, it would be necessary that the scale parameter of the Laplace distribution for the q^{th} query should equal $(q \times b)$, with resulting variance equal to $(q^2 \times 2b^2)$. In other words, in our current example, for the 10^{th} query, the variance of the noise added would be 200 units; and for the 50^{th} query, the noise variance would be 5000. Adding noise with variance of 5000 (or even 200) when the variance of the actual data set is only 0.09339 makes the query responses practically meaningless. Table 2 shows the impact of increasing the noise variance to account for the reduction in the level of security. In generating this data, we have used exactly the same random numbers to generate the noise in this table as we did in Table 1. For instance, for the query sum of $(x_1 + x_{47})$ where the true sum is 1.00306, we get a masked response of -122.7312. Similarly, for the sum of $(x_1 + x_{40})$ where the true sum is 1.60286, the response from the system is 91.8833. As observed earlier, since $\mathbf{X}$ is in the range (0, 1), the sum of $(x_i + x_j)$ must be in the range (0, 2). However, we observe in Table 2 that as the number of queries increases practically none of the responses fall in the meaningful range of (0, 2). Out of the total of 49 queries, only 5 fall in the meaningful range. For any intelligent user who knows that the sum of two observations must be in the range (0, 2), practically all the responses from the system after the first few queries are useless. Hence, as shown in Table 2, increasing the variance as the number queries increases may maintain security, but makes the responses practically useless, and hence is simply not a feasible approach.

In summary, with the original specifications, the data administrator can only be certain that e^{ε}-differential privacy is satisfied only for the first query. For all subsequent queries, the value of ε increases and the level of security decreases. If we attempt to increase the noise variance to compensate for the reduction in security, the resulting responses to queries are practically useless. Finally, while we have illustrated this approach for a single data set, the intruder can adopt the tracker approach for any data set of any size. Hence, the results in this section can be generalized to any data set.

Table 2. Responses from the system when noise variance increases with queries

Individual	x	$(x_1 + x_i)$	Random #	y_i	R = $(x_1 + x_i) + y_i$
1	0.97032				
2	0.85490	1.82522	0.52770	0.0570	1.8822
3	0.72936	1.69968	0.07910	-3.6877	-1.9881
4	0.06435	1.03467	0.43122	-0.4440	0.5907
5	0.42397	1.39429	0.79758	3.6171	5.0114
6	0.75934	1.72966	0.89064	7.6000	9.3297
7	0.67422	1.64454	0.52139	0.2623	1.9069
8	0.62075	1.59107	0.93254	14.0218	15.6129
9	0.66039	1.63071	0.26885	-4.9635	-3.3328
10	0.54600	1.51632	0.70928	4.8802	6.3966
11	0.22039	1.19071	0.50533	0.1072	1.2979
12	0.98132	1.95164	0.56455	1.5204	3.4721
13	0.22174	1.19206	0.76173	8.8943	10.0863
14	0.88548	1.85580	0.92787	25.1705	27.0263
15	0.95191	1.92223	0.18962	-13.5740	-11.6517
16	0.65780	1.62812	0.80940	14.4668	16.0949
17	0.88826	1.85858	0.63536	5.0510	6.9096
18	0.74429	1.71461	0.73987	11.1084	12.8230
19	0.12368	1.09400	0.35452	-6.1894	-5.0954
20	0.59708	1.56740	0.10678	-29.3332	-27.7658
21	0.03746	1.00778	0.76936	15.4752	16.4830
22	0.82311	1.79343	0.14070	-26.6283	-24.8348
23	0.03147	1.00179	0.81132	21.4403	22.4421
24	0.32822	1.29854	0.40197	-5.0193	-3.7208
25	0.20763	1.17795	0.87744	33.7443	34.9223
26	0.57210	1.54242	0.50602	0.3027	1.8451
27	0.66724	1.63756	0.07235	-50.2613	-48.6237
28	0.36904	1.33936	0.26919	-16.7185	-15.3792
29	0.83805	1.80837	0.79157	24.4995	26.3078
30	0.72112	1.69144	0.85190	35.2849	36.9763
31	0.98357	1.95389	0.83221	32.7567	34.7106
32	0.23028	1.20060	0.56917	4.6158	5.8164
33	0.09613	1.06645	0.04275	-78.6988	-77.6323
34	0.00538	0.97570	0.13981	-42.0535	-41.0778
35	0.46984	1.44016	0.56482	4.7211	6.1613
36	0.96043	1.93075	0.31687	-15.9635	-14.0328
37	0.20283	1.17315	0.22134	-29.3368	-28.1636
38	0.60845	1.57877	0.30995	-17.6934	-16.1147
39	0.45104	1.42136	0.13421	-49.9775	-48.5562
40	0.63254	1.60286	0.95061	90.2804	91.8833
41	0.49287	1.46319	0.62664	11.6822	13.1454
42	0.47326	1.44358	0.87987	58.4673	59.9109
43	0.87437	1.84469	0.17758	-43.4769	-41.6322
44	0.01190	0.98222	0.86866	57.4829	58.4651
45	0.89823	1.86855	0.53223	2.9319	4.8004
46	0.54942	1.51974	0.42995	-6.7928	-5.2730
47	0.03274	1.00306	0.03394	-123.7343	-122.7312
48	0.61882	1.58914	0.36316	-15.0285	-13.4394
49	0.30366	1.27398	0.17065	-51.5983	-50.3243
50	0.33108	1.30140	0.40338	-10.5222	-9.2208

In a recent paper, Dwork and Smith (2009, page 139) acknowledge the issue with multiple queries when they observe that:

> Differential privacy applies equally well to an interactive process, in which an adversary adaptively questions the curator about the data. The probability K(S) then depends on the adversary's strategy, so the

> definition becomes more delicate. However, one can prove that if the algorithm used to answer each question is ε-differentially private, and the adversary asks q questions, then the resulting process is qε-differentially private, no matter what the adversary's strategy is.

This is precisely the result that was illustrated in this section. The implications of this statement are far reaching than what it seems at first glance. Assume that differential privacy based Laplace noise addition has been implemented on a data set and some ε has been specified. The above statement implies that the intruder's knowledge gain for the very first query is e^{ε}; for the second query, it is $e^{2\varepsilon}$; … for the q^{th} query, it is $e^{q\varepsilon}$. In other words, the intruder's knowledge gain increases exponentially with the number of queries. *Consequently, after just a few queries, the intruder's knowledge gain is so large that differential privacy based Laplace noise addition procedure offers no security at all*.

The only solution to alleviate the above problem is to increase the variance of the noise added with the number of queries. Unfortunately, as we have shown, this has the consequence of making the responses useless after just a few queries. Thus, after just a few queries, Laplace noise addition results in either no security or no utility.

5 Conclusions

Differential privacy is often characterized by its proponents along the following lines (Dwork and Smith 2009, page 137):

> For appropriate ε, a mechanism κ satisfying this definition addresses all concerns that any participant might have about the leakage of her personal information: even if the participant were to remove her data from the data set, no outputs (and thus no consequences of outputs) would become significantly more or less likely. For example, if the database were to be consulted by an insurance provider before deciding whether or not to insure a given individual, then the presence or absence of that individual's data in the database would not significantly affect her chance of receiving coverage. Differential privacy is therefore an *ad omnia* guarantee, as opposed to an *ad hoc* definition that provides guarantees only against a specific set of attacks or concerns.

This characterization ignores the following very practical issues highlighted in this paper:

(1) In many situations, it may not even be possible to implement differential privacy since the numerical variable in question may not have known natural lower and upper bounds.
(2) Even when upper and lower bounds are known, because of global sensitivity, the level of noise added may be so large as to make responses from such a system meaningless for many queries.

(3) Even if the above two requirements are satisfied, the intruder's knowledge gain will be limited to e^{ε} only for the first query. For subsequent queries, the intruder's knowledge gain increases exponentially, resulting in practically no security after just a few queries.
(4) If we attempt to address the issue in (3) above by increasing the noise variance, after just a few queries, the resulting noise variance is so large as to make all responses to all queries meaningless.

In summary, differential privacy and the associated Laplace noise addition procedure may sound like a good idea in theory. However, when we actually examine the applicability of this approach to numerical data as we do in this paper, we find that it has very limited applicability offering either very little security or very little utility or neither.

References

1. Abowd, J.M., Vilhuber, L.: How Protective Are Synthetic Data? In: Domingo-Ferrer, J., Saygın, Y. (eds.) PSD 2008. LNCS, vol. 5262, pp. 239–246. Springer, Heidelberg (2008)
2. Chaudhuri, K., Monteleoni, C.: Privacy-preserving logistic regression. In: Proceedings of the Twenty-Second Annual Conference on Neural Information Processing Systems (NIPS), pp. 289–296 (2008)
3. Denning, D.E., Denning, P.J., Schwartz, M.D.: The tracker: A threat to statistical database security. ACM Transactions on Database Systems 4, 76–96 (1979)
4. Dinur, I., Nissim, K.: Revealing Information while Preserving Privacy. In: PODS 2003, San Diego, CA, pp. 202–210 (2003)
5. Duncan, G.T., Mukherjee, S.: Optimal Disclosure Limitation Strategy in Statistical Databases: Deterring Tracker Attacks Through Additive Noise. Journal of the American Statistical Association 95, 720–729 (2000)
6. Dwork, C., Smith, A.: Differential Privacy for Statistics: What we Know and What we Want to Learn. Journal of Privacy and Confidentiality 1, 135–154 (2009)
7. Dwork, C.: Differential Privacy. In: Bugliesi, M., Preneel, B., Sassone, V., Wegener, I. (eds.) ICALP 2006. LNCS, vol. 4052, pp. 1–12. Springer, Heidelberg (2006)
8. Machanavajjhala, A., Kifer, D., Abowd, J., Gehrke, J., Vilhuber, L.: Privacy: From Theory to Practice on the Map. In: ICDE, pp. 277–286. IEEE Computer Society, Los Alamitos (2008)
9. Nissim, K., Raskhodnokova, S., Smith, A.: Smooth Sensitivity and Sampling in Private Data Analysis. In: Proceedings of the thirty-ninth annual ACM symposium on Theory of computing, pp. 75–84 (2007)
10. Sarathy, R., Muralidhar, K.: Differential Privacy for Numeric Data. In: Joint UNECE/Eurostat work session on statistical data confidentiality, Bilbao, Spain (2009)
11. Wasserman, L., Zhou, S.: A Statistical Framework for Differential Privacy. Journal of the American Statistical Association 105, 375–389 (2009)

Remote Data Access and the Risk of Disclosure from Linear Regression: An Empirical Study

Philipp Bleninger[1], Jörg Drechsler[1], and Gerd Ronning[2]

[1] Institute for Employment Research, Regensburger Str. 104, 90478 Nuremberg, Germany
[2] Department of Economics, Tübingen University, Mohlstrasse 36, 72074 Tübingen, Germany

Abstract. In the endeavor of finding ways for easy data access for researchers not employed at a statistical agency remote data access seems to be an attractive alternative to the current standard of either altering the data substantially before release or allowing access only at designated data archives or research data centers. Data perturbation is often not accepted by the researchers since they do not trust the results from the altered data sets. But on-site access puts some heavy burdens on the researcher and the data providing agency both in terms of time and money. Remote data access or remote analysis servers that allow to submit queries without actually seeing the microdata have the potential of overcoming both these disadvantages. However, even if the microdata is not available to the researcher directly, disclosure of sensitive information for individual survey respondents is still possible.

In this paper we illustrate how an intruder could use some commonly available background information to reveal sensitive information using simple linear regression. We demonstrate the real risks from this approach with an empirical evaluation based on a German establishment survey, the IAB Establishment Panel. Although these kind of attacks can easily be prevented once the agency is aware of the problem, this small simulation aims to emphasize that there might be many ways to obtain sensitive information using multivariate analysis and not all of them are obvious. Thus, agencies thinking about actually implementing some form of remote data access should consider carefully which queries could be allowed by the system.

Keywords: Artificial outlier, disclosure risk, IAB Establishment Panel, linear regression, remote access, strategic dummy.

1 Introduction

Data collecting agencies generally have two options if they are willing to provide access to their data for external researchers. They can release data sets to the public if they can guarantee that the dissemination will not harm the privacy of any survey respondent or they can allow external researchers on-site access to the data in research data centers (RDC) or data enclaves. Since most data

J. Domingo-Ferrer and E. Magkos (Eds.): PSD 2010, LNCS 6344, pp. 220–233, 2010.

have to be altered in some way to allow data dissemination, many researchers prefer the direct access to the unaltered data at the RDC, especially if the data dissemination requires perturbation of the microdata. For this reason more and more agencies deposit their data at data enclaves or set up their own research data centers. However, the use of these facilities comes at a high price both for the researcher and the providing agency. Researchers have to travel to the agency before they ever get in touch with the original data. Although some agencies provide dummy data sets to give the researcher an idea of the real data, these dummy data sets often are of very low quality and the researcher might not realize that the data collected by the agency is not suitable for her analysis before traveling to the agency. Furthermore, researchers can request a certain time slot at the RDC in which they expect to finish their research. It is very difficult for the researcher to anticipate how long the data preparation will take without access to the data, and unexpected problems might require more days than the admitted time slot will allow. Besides, if the researcher wants to extend her research maybe using more variables than she asked for in the original proposal, she might have to go through the complete reviewing process again before she can actually add the variables to her analysis. On the other hand, the agency has to check every output from the analysis for potential disclosure violations. Only cleared outputs may leave the RDC and may be used by the researcher for publication. At present, this output checking is still carried out manually. With the growing popularity of the RDCs the capacity of handling all this output checking is at the limit.

Given these drawbacks remote data access seems to be the panacea for data access for external researchers. In an ideal world full remote access would enable the external researcher to connect to a host server from her desktop machine. She would see the microdata on the screen and would be allowed to manipulate them in any way but the actual data would never leave the server and it would not be possible to store the microdata on the desktop computer. Requested queries would be automatically scanned for possible confidentiality violations and only those queries that pass the confidentiality check would be answered by the server. Remote access would free the researcher from the burden of traveling to the RDC and it would render the cost intensive and time consuming manual output checking unnecessary. However, there are many obstacles with this approach making the full implementation of a remote data access more than questionable. Apart from the technical issues of guaranteeing a safe connection between the desktop computer of the external user and the microdata server at the agency, direct access to the unchanged microdata is prohibited by law in many countries. For example in Germany, the data accessible for external researchers is required to be *de facto* anonymised[1], which is still a privilege compared to the *absolute* anonymity that is required for all published results. One solution in this context could be that the researcher would only see an anonymised version of the

[1] *De facto* anonymous means that the effort that is necessary to identify a single unit in the data set is higher than the actual benefit the potential intruder would achieve by this identification.

microdata on her screen but the queries she submits to the server would actually be run on the original data. However, this would still require the server to identify all queries that might lead to a breach of confidentiality. Some of these queries are easy to identify. For example queries that ask for the maximum or minimum of a variable should never be allowed. For tabulation queries potentially identifying small cells could be suppressed using standard rules from the cell suppression literature[2]. However, there are other analyses for which it is not that obvious that they actually might impose an increased risk of disclosure and illustrating this for a specific set of queries is the main aim of this paper.

We focus on the risks from simple linear regression analysis under the assumption that the user will never see the true microdata. Given the legal restrictions in many countries (see discussion above), we believe that even under remote access the user will only see an anonymised version of the true microdata. In this sense our notion of remote access is located somewhere in the middle between the dream of a full remote access and the idea of a remote analysis server that can only answer specified queries without providing access to any microdata at all. We note that our findings are also relevant in the context of a plain remote analysis server.

Often regression analysis is considered as safe in the sense that it is assumed that no output checking is required. Following the discussion in [3] we illustrate that an intruder with background knowledge on some of the variables contained in the data set can get accurate estimates for any sensitive variable she is interested in using only the results from a linear regression analysis. We use the IAB Establishment Panel to demonstrate empirically that at least for business data very limited and easily available background information can be sufficient to allow the intruder to obtain sensitive information with this approach.

The remainder of the paper is organized as follows. Chapter 2 will recapitulate the basic concept that allows the intruder to retrieve sensitive information for a single respondent based on the background information she has about that respondent. In this chapter we follow the outline described in [3]. In the next chapter we shortly introduce the data set we used for the empirical simulations: the IAB Establishment Panel. This data set is used in the following chapter to illustrate that only very limited background information is required to learn sensitive information about a survey respondent in this setting. The paper concludes with some final remarks.

2 The Formal Approach

In the following we assume that the intruder has - at least approximate - knowledge about some of the variables contained in the survey for a certain survey respondent m. It is important to note that this knowledge may refer to any set

[2] Even cell suppression can quickly become problematic, if we allow dynamic queries. In this case, the server would have to keep track of all earlier queries and would have to guarantee that requests submitted at a later point in time would not allow the calculation of cell entries that are being suppressed now.

of variables in the data set, no matter if the variables are sensitive or not. For example in a business survey, the external information available to the intruder might be the energy consumption or the total production time. The intruder would then use these variables for obtaining information on sensitive variables such as investment, sales, or research expenditures. In the following we denote the variable for which information is at hand by x and the true value for this variable provided by the survey respondent m by x_m^0. Finally, let $\hat{x}_m$ be the external information the intruder obtained about the survey respondent m for this variable. This information may be exact, that is, $\hat{x}_m = x_m^0$ is known with certainty. In other situations their might only be information available about an interval in which the true value x_m^0 must fall. This range might be formulated in additive or multiplicative terms., that is

$$x_m^0 - \gamma < \hat{x}_m < x_m^0 + \gamma \text{ or } (1-\delta)x_m^0 < \hat{x}_m < (1+\delta)x_m^0 \quad .$$

Let the sensitive variable of interest be denoted by y and its value for the specific respondent by y_m.

[3] pointed out that the knowledge of $\hat{x}_m$ may be used to obtain information for any other variable (contained in the microdata set) for this respondent by making the variable of interest the dependent variable in a simple linear regression analysis. The authors propose two approaches: (i) One could generate an "artificial outlier" which is obtained by transformation. (ii) Alternatively, one could employ a "strategic dummy variable" which uses the background information for identifying the respondent m.

For the artificial outlier approach the intruder may use her (exact) knowledge by defining a new regressor variable

$$z = \frac{1}{|x - \hat{x}_m| + \varepsilon} \tag{1}$$

where ε is arbitrarily small. If we include this regressor variable in a linear regression with the variable of interest specified as the dependent variable, the regressor z will become extremely large for the respondent m and therefore generates a leverage point such that the predicted value of the dependent variable tends towards the true value y_m^0 for this respondent. A formal proof that

$$\lim_{z_m \to \infty} \hat{y}_m \quad = \quad y_m$$

holds, is given in Appendix A.1 for the case of a simple regression and rests on the assumption that no other respondent reports a value for x that is equal to x_m^0.

Alternatively, one could define a dummy that exploits the knowledge regarding the variable x. In case of exact knowledge the dummy would be given by

$$\Im_{x=x_m} \quad = \begin{cases} 1 \text{ if } x = \hat{x}_m \\ 0 \quad \text{else} \end{cases} \tag{2}$$

whereas assuming only approximate knowledge one would use

$$\Im_{x \simeq x_m} \quad = \begin{cases} 1 \text{ if } x - \gamma < \hat{x}_m < x + \gamma \\ 0 \quad \text{else} \end{cases} \tag{3}$$

or the corresponding multiplicative specification mentioned above. It is shown in Appendix A.2 that a simple regression which uses just this dummy variable and any variable of interest as the dependent variable will result in

$$\hat{y}_m \quad = \quad y_m^0 \quad .$$

The result remains valid also in the case of other regressors added. See Appendix A.2.

However, the proof again is based on the assumption that only a single respondent is identified using the knowledge regarding x. If x is a categorical variable, this is an unrealistic assumption and even for continuous variables many respondents may report the same value. Still, with the dummy variable approach the constructed dummy can easily be based on more than one variable exploiting all the information the intruder has about the survey respondent. In our business survey example this could mean that the intruder also uses her information about the industry, an approximate number of employees, and regional information about the establishment she is looking for. In this case we could define an indicator dummy for each variable for which the intruder has background information. Let $x_1, ..., x_p$ be the variables for which background information is available and let $\Im_1, ...\Im_p$ be the corresponding indicators defined as in (2) or (3). Now the final indicator can be defined as follows:

$$\Im \quad = \begin{cases} 1 \text{ if } \Im_1 = 1 \wedge \Im_2 = 1 \wedge \cdots \wedge \Im_p = 1 \\ 0 \qquad\qquad\qquad\qquad \text{else} \end{cases} \tag{4}$$

It is important to note that both approaches critically rely on the assumption that a single record can be identified with the external information the intruder has about m. However, the artificial outlier approach requires that a single variable is sufficient to identify the respondent and also that the intruder knows x_m^0 exactly. This is often unrealistic in reality. With the dummy variable approach several variables can be combined to identify the target uniquely and it can be sufficient to have a rough estimate of x_m^0.

3 The IAB Establishment Panel

Since our empirical evaluations in the next section are based on the wave 2007 of the IAB Establishment Panel a short introduction of the data set should prelude our illustrations. The IAB Establishment Panel is based on the German employment register aggregated via the establishment number as of 30 June of each year. The basis of the register, the German Social Security Data (GSSD) is the integrated notification procedure for the health, pension and unemployment insurances, which was introduced in January 1973. This procedure requires employers to notify the social security agencies about all employees covered by social security. As by definition the German Social Security Data only include employees covered by social security - civil servants and unpaid family workers for example are not included - approx. 80% of the German workforce are

represented. However, the degree of coverage varies considerably across the occupations and the industries.

Since the register only contains information on employees covered by social security, the panel includes establishments with at least one employee covered by social security. The sample is drawn using a stratified sampling design. The stratification cells are defined by ten classes for the size of the establishment, 16 classes for the region, and 17 classes for the industry. These cells are also used for weighting and extrapolation of the sample. The survey is conducted by interviewers from TNS Infratest Sozialforschung. For the first wave, 4,265 establishments were interviewed in West Germany in the third quarter of 1993. Since then the Establishment Panel has been conducted annually - since 1996 with over 4,700 establishments in East Germany in addition. In the wave 2007 more than 15,000 establishments participated in the survey. Each year, the panel is accompanied by supplementary samples and follow-up samples to include new or reviving establishments and to compensate for panel mortality. The list of questions contains detailed information about the firms' personnel structure, development and personnel policy. For a detailed description of the data set we refer to [2] or [5]. For the simulations we use one data set with all missing values imputed. We treat all imputed values like originally observed values for simplicity. See [1] for a description of the multiple imputation of the missing values in the survey.

4 Empirical Evidence

For our empirical evaluations, we use the wave 2007 of the establishment survey and treat the turnover of an establishment as the sensitive variable to be disclosed. Thus, we exclude all entities from the survey that do not report turnover such as non-industrial organizations, regional and local authorities and administrations, financial institutions, and insurance companies. The remaining data set includes 12,814 completely observed establishments.

We start with the artificial outlier approach. Using the number of employees as the available background information we construct a variable z according to (1) setting $\epsilon = 0.0001$. To evaluate the risks for the complete data set we run an artificial outlier regression for each distinct value of establishment size and predict the turnover. Table 1 summarizes the results of these regressions for different subsets of the data.

The first column defines the subset of the data. For example, the results for the 90% quantile represent only the largest 10% of establishments. The second column presents the number of records that are contained in the subset. Column 3 contains the percentage of records that are uniquely identified based on an artificial outlier derived from the establishment size, i. e. it contains the percentage of unique high leverage points. If there is more than one high leverage point, additional establishments reduce the prediction accuracy of the target's turnover.

Table 1. Disclosure risk evaluations using the artificial outlier approach

quantile	N	prop. exactly identified	mean rel. error of all	mean rel. error of exactly identified
all	12814	0.034	13773.795	0.001
0.5	6516	0.066	26936.156	0.001
0.75	3217	0.134	51952.053	0.001
0.9	1282	0.335	1.957	0.001
0.99	129	0.969	0.011	0.0001
0.999	13	1	$1.195 * 10^{-6}$	$1.195 * 10^{-6}$

Column 4 presents the average absolute relative error between the predicted and the observed value for turnover for the target record m, i.e.

$$\Delta = \frac{1}{N} \sum_{j=1}^{N} \frac{|\hat{y}_{m=j} - y_{m=j}|}{y_{m=j}} \tag{5}$$

for all records in the subset. Finally, column 5 presents the same quantity only for the records that are uniquely identified and therefore generate a unique high leverage point for z. As expected the disclosure risk clearly increases with establishment size. Under the (in most cases unrealistic) assumption that the intruder would know the exact reported establishment size, we observe a substantial increase in the risk when going from the largest 10% of the establishments (33.5% correctly identified) to the largest 1% of establishments (96.9% correctly identified). Below these thresholds identification risks are relatively low since establishment size alone will not uniquely identify a single record. The results in column 4 illustrate that generally risks are low as long as a unique identification is not possible. The average absolute relative error is very large (often far more than 100%) indicating that the predicted value on average differs substantially from the reported value. Finally, all the values close to zero in the last column are by no means surprising. This is a direct result of the proof given in the appendix. We only include these results to emphasize that once a record is uniquely identified, the intruder does not have to have direct access to the microdata. Instead she can use the artificial outlier approach (or the dummy variable approach discussed below) to exactly reveal any sensitive information about the identified record.

We also checked whether adding a second artificial outlier z^* generated from another variable would increase the risks. Using the total number of traineeships as additional background information available to the intruder we found that the prediction accuracy from the disclosure regression actually decreased. One reason for this results might be that all records with the same number of employees as the target record will have a high leverage point for z but a different subset of records, namely all the records that have the same number of traineeships, will have a high leverage point for z^*. Thus, overall the number of leverage points increases and as a result the prediction accuracy for y_m decreases. We

Table 2. Disclosure risk evaluations using the strategic dummy approach

quantile	N	indicators $\Im_k$	prop. exactly identified	ave. rel. error of all	ave. rel. error of exactly identified
all	12814	exact size	0.034	13801.825	0
		approx. size	0.0009	11025.450	0
		+ federal state	0.023	11739.574	0
		+ legal form	0.116	13633.345	0
		+ branch	0.658	1.478	0
0.5	6516	exact size	0.066	26985.871	0
		approx. size	0.002	21526.008	0
		+ federal state	0.046	21945.190	0
		+ legal form	0.200	26774.728	0
		+ branch	0.846	0.323	0
0.75	3217	exact size	0.134	52023.417	0
		approx. size	0.003	40965.983	0
		+ federal state	0.085	39651.427	0
		+ legal form	0.228	48390.483	0
		+ branch	0.868	0.147	0
0.9	1282	exact size	0.335	1.956	0
		approx. size	0.009	4.296	0
		+ federal state	0.186	1.944	0
		+ legal form	0.352	1.499	0
		+ branch	0.895	0.070	0
0.99	129	exact size	0.969	0.011	0
		approx. size	0.085	1.311	0
		+ federal state	0.682	0.136	0
		+ legal form	0.806	0.055	0
		+ branch	0.953	0.021	0
0.999	13	exact size	1	0	0
		approx. size	0.310	0.093	0
		+ federal state	1	0	0
		+ legal form	1	0	0
		+ branch	1	0	0

are currently working on evaluating the general effects of using more than one artificial outlier in the disclosure regression.

For the strategic dummy approach we evaluate for each record if a unique identification is possible using a varying amount of background information. For the background information we chose four variables that we believe are easy to obtain for an intruder from public records, namely the (approximate) size of the establishment, i.e. its (approximate) total number of employees, the German Federal State the establishment is located in, its legal form and its industrial sector (recorded in 40 categories). We evaluate the increase in risk if these variables are added successively to the strategic dummy. The results are summarized in Table 2.

Not surprisingly the same percentage of records as in Table 1 are identified, if the exact establishment size is used as a dummy. Relaxing the unrealistic assumption of exactly knowing the size of the establishment we use an indicator for the approximate total number of employees that identifies all records that lie within ±2.5% of the reported establishment size. Using only this information almost never uniquely identifies a record in the data set. Even for the top 0.1% of establishments only 31% are uniquely identified. However, adding more information significantly increases the risk. When all four background variables are used, more than 65% of the establishments are identified uniquely in the entire data set. Since arguably intruders will only be interested in the larger establishments and not in small family businesses, the fact that almost 90% of the records can be uniquely identified for the largest 10% of the establishments based on very little background information is an alarming result. Again, we only include the results in the last column of the table to emphasize that once a record is uniquely identified all information in the data set for that record can be revealed easily without access to the actual microdata.

This leads to the question how the intruder will know that she has indeed uniquely identified the m-th respondent. Of course, the natural way would be to check the residuals of the regression for zeroes. However, residuals usually are not reported in remote access. Alternatively, for the dummy variable approach the intruder could check the mean of the generated dummy variable which should be $1/n$ in case of unique identification. If the agency decides to suppress means for binary variables with few positive (or negative) outcomes, the intruder could compute the variance of the dummy variable. Given a unique identification it should be equal to $Var(\Im) = 1/n^2$. Both approaches are of course not possible when generating an artificial outlier since z would just be a new continuous variable with unknown mean and variance. In this case, the intruder might check, if a unique maximum exists for z. Only if the maximum is unique, a single record has been identified. However, such requests will likely be suppressed by the remote server. This can be seen as an additional argument in favor of the strategic dummy approach.

5 Conclusion

It is obvious that agencies – once they are aware of the risks described in the previous sections – can easily prevent this type of disclosure, e.g. by prohibiting regressions that contain dichotomous regressors with less than say 3 positive outcomes. But it is important that the agency must be aware of the problem to prevent it. The point that we are trying to make is that there are many constellations that might lead to a risk of disclosure. Some are obvious others are more difficult to detect in advance. Full remote access without any intervenience of the agency would require that all possible constellations are considered and ruled out before data access is provided. The risk from linear regressions that is the main topic of this paper is only one example of a disclosure risk that might not be obvious at first glance. We believe there are many other situations that

might be equally harmful. For example it is well known that saturated models can reveal the exact information for small cell table entries that would have been protected by cell suppression or any other SDL technique if the table would have been requested directly. We believe that more research in the area is needed to detect other user queries that might impose some risk of disclosure. Whether it will be possible to rule out all potential disclosure risks in advance remains more than questionable.

Acknowledgments. This research was supported by a grant from the German Federal Ministry of Education and Research.

References

1. Drechsler, J.: Multiple imputation of missing values in the wave 2007 of the IAB Establishment Panel. IAB Discussion Paper No. 6 (2010)
2. Fischer, G., Janik, F., Müller, D., Schmucker, A.: The IAB Establishment Panel – from sample to survey to projection. Tech. rep., FDZ-Methodenreport No. 1 (2008)
3. Gomatam, S., Karr, A.F., Reiter, J.P., Sanil, A.P.: Data dissemination and disclosure limitation in a world without microdata: A risk-utility framework for remote access servers. Statistical Science 20, 163–177 (2005)
4. Hoaglin, D.C., Welsh, R.E.: The Hat Matrix in Regression and ANOVA. The American Statistician 32, 17–22 (1978)
5. Kölling, A.: The IAB-Establishment Panel. Journal of Applied Social Science Studies 120, 291–300 (2000)

A Artificial Outliers and Strategic Dummies

We consider the linear model

$$\mathbf{y} \quad = \quad \mathbf{X}\,\boldsymbol{\beta} \;+\; \mathbf{u}\,, \tag{6}$$

where $\mathbf{y}$ and $\mathbf{u}$ are n-dimensional vectors, $\boldsymbol{\beta}$ is a K-dimensional vector and $\mathbf{X}$ a $(n \times K)$ matrix with $\boldsymbol{\iota}' = (1, 1, \ldots, 1)$ as the first column. The vector of predicted values is given by

$$\hat{\mathbf{y}} = \mathbf{X}\,(\mathbf{X}'\mathbf{X})^{-1}\,\mathbf{X}'\,\mathbf{y} = \mathbf{H}\,\mathbf{y} \tag{7}$$

where $\mathbf{H}$, called the hat matrix, measures the "leverage" of a certain regressor.[3]

A.1 Artificial Outliers

In the following we assume that the observations are ordered such that observations for survey respondent m are in the first row of the data matrix. Therefore z_1 contains the artificial outlier which tends towards infinity; compare the definition (1) of artificial outliers in the main text.

[3] See, e.g. [4].

In the special case of a simple regression ($K = 2$) with $\mathbf{X} = \left(\boldsymbol{\iota}\ \mathbf{z}\right)$ the elements of the hat matrix are given by

$$h_{jk} = \frac{1}{n\sum z_i^2 - \left(\sum z_i\right)^2}\left(\sum_{i=1}^{n} z_i^2 - z_j\sum_{i=1}^{n} z_i - z_k\sum_{i=1}^{n} z_i + nz_jz_k\right) ,$$

with $j = 1, ..., n$ and k=1,...,n. Therefore the jth element of the vector of predicted values $\hat{\mathbf{y}}$ is given by

$$\begin{aligned}\hat{y}_j &= \textstyle\sum_{k=1}^{n} h_{jk}y_k \\ &= \frac{1}{n\sum z_i^2 - \left(\sum z_i\right)^2}\sum_{k=1}^{n}\left(\sum_{i=1}^{n} z_i^2 - z_j\sum_{i=1}^{n} z_i - z_k\sum_{i=1}^{n} z_i + nz_jz_k\right) y_k\end{aligned}$$

and in particular for $j = 1$ we have

$$\begin{aligned}\hat{y}_1 &= \tfrac{1}{n\sum z_i^2-(\sum z_i)^2}\ \textstyle\sum_{k=1}^{n}\left[\sum_{i=1}^{n} z_i^2 - z_1\sum_{i=1}^{n} z_i - z_k\sum_{i=1}^{n} z_i + nz_1z_k\right] y_k \\ &= \tfrac{1}{n\sum z_i^2-(\sum z_i)^2}\left[\left(\textstyle\sum_{i=1}^{n} z_i^2 - z_1\sum_{i=1}^{n} z_i\right)\sum_{k=1}^{n} y_k - \left(\sum_{i=1}^{n} z_i - nz_1\right)\sum_{k=1}^{n} z_ky_k\right] \\ &= \frac{\left(\sum_{i=1}^{n} z_i^2 - z_1\sum_{i=1}^{n} z_i\right)\sum_{k=1}^{n} y_k}{n\sum z_i^2-(\sum z_i)^2} - \frac{\left(\sum_{i=1}^{n} z_i - nz_1\right)\sum_{k=1}^{n} z_ky_k}{n\sum z_i^2-(\sum z_i)^2} \\ &= \frac{\left(z_1^2+\sum_{i>1} z_i^2 - z_1(z_1+\sum_{i>1} z_i)\right)\sum_{k=1}^{n} y_k}{n(z_1^2+\sum_{i>1} z_i^2)-(z_1+\sum_{i>1} z_i)^2} - \frac{\left(z_1+\sum_{i>1} z_i - nz_1\right)(z_1y_1+\sum_{k>1} z_ky_k)}{n(z_1^2+\sum_{i>1} z_i^2)-(z_1+\sum_{i>1} z_i)^2} \\ &= A - B .\end{aligned}$$

In order to obtain results for $z_1 \to \infty$ we write the two terms as follows:

$$A = \frac{\left[\left(1 + \frac{\sum_{i>1} z_i^2}{z_1^2}\right) - \left(1 + \frac{\sum_{i>1} z_i}{z_1}\right)\right]\sum_{k=1}^{n} y_k}{n\left(1 + \frac{\sum_{i>1} z_i^2}{z_1^2}\right) - \left(1 + \frac{\sum_{i>1} z_i}{z_1}\right)^2}$$

and

$$B = \frac{\left(1 + \frac{\sum_{i>1} z_i}{z_1} - n\right)\left(y_1 + \frac{\sum_{k>1} z_ky_k}{z_1}\right)}{n\left(1 + \frac{\sum_{i>1} z_i^2}{z_1^2}\right) - \left(1 + \frac{\sum_{i>1} z_i}{z_1}\right)^2}$$

from which we obtain

$$\lim_{z_1\to\infty} \hat{y}_1 = \lim_{z_1\to\infty} (A - B) = \frac{0}{n-1} - \frac{(1-n)\,y_1}{n-1} = y_1 .$$

Therefore for a sufficiently large z_1 we can approximate y_1 by its predicted value $\hat{y}_1$.

A.2 Strategic Dummy Variables

Simple regression. In case of unique identification by (2), (3) or (4) the regressor matrix is given by

$$\mathbf{X} \quad = \quad (\boldsymbol{\iota}\, \mathbf{e}_1)\,,$$

where $\mathbf{e}_1$ is an n-dimensional vector with 1 as the first element and 0 for the remaining $n-1$ elements. Therefore

$$(\boldsymbol{X}'\boldsymbol{X})^{-1} = \frac{1}{(n-1)} \begin{pmatrix} 1 & -1 \\ -1 & n \end{pmatrix}$$

and

$$\mathbf{H} \quad = \quad \frac{1}{n-1} \begin{pmatrix} n-1 & \mathbf{0}' \\ \mathbf{0} & \boldsymbol{\iota}_{n-1}\boldsymbol{\iota}'_{n-1} \end{pmatrix},$$

where $\mathbf{0}$ is the $(n-1)$-dimensional null vector and $\boldsymbol{\iota}_{n-1}$ a $(n-1)$-dimensional vector of ones. Note that $h_{11} = 1$ and $h_{1j} = 0$, $j > 1$, so that the predicted value for y_1 is given by

$$\hat{y}_1 \quad = \quad \sum_{k=1}^{n} h_{1k} y_k \quad = \quad \frac{1}{n-1} \left((n-1)\, y_1 + \sum_{k>1} 0 \cdot y_k \right) \quad = \quad y_1\,.$$

The case of additional regressors. We now consider the case that other regressors are added to the regression which might be motivated by the idea that the use of a strategic dummy is not so easily detected by the agency if other regressors are also included in the model. We write the model in partitioned form as

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u} = (\,\boldsymbol{X}_1\ \boldsymbol{X}_2\,) \begin{pmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{pmatrix} + \boldsymbol{u} = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{X}_2\boldsymbol{\beta}_2 + \boldsymbol{u}\,.$$

with

$$\mathbf{X}_2 \quad = \quad \mathbf{e}_1$$

so that this submatrix contains only the information regarding the strategic dummy. Then the vector of predicted values can be written as

$$\begin{aligned} \hat{\mathbf{y}} &= \mathbf{X}_1\,\hat{\boldsymbol{\beta}}_1 \ + \ \mathbf{X}_2\,\hat{\boldsymbol{\beta}}_2 \\ &= \mathbf{X}_1\,\hat{\boldsymbol{\beta}}_1 \ + \ \mathbf{X}_2\,(\boldsymbol{X}'_2\boldsymbol{X}_2)^{-1}\boldsymbol{X}'_2 \left(\boldsymbol{y} - \boldsymbol{X}_1\hat{\boldsymbol{\beta}}_1 \right) \end{aligned} \tag{8}$$

Since

$$\mathbf{X}_2\,(\boldsymbol{X}'_2\boldsymbol{X}_2)^{-1}\boldsymbol{X}'_2 \quad = \quad \begin{pmatrix} 1 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}$$

and

$$\mathbf{X}_2\,(\boldsymbol{X}_2'\boldsymbol{X}_2)^{-1}\boldsymbol{X}_2'\left(\boldsymbol{y}-\boldsymbol{X}_1\hat{\boldsymbol{\beta}}_1\right)=\begin{pmatrix}1&\mathbf{0}\\ \mathbf{0}&\mathbf{0}\end{pmatrix}\left(\boldsymbol{y}-\boldsymbol{X}_1\hat{\boldsymbol{\beta}}_1\right)$$

$$=\begin{pmatrix}y_1\\0\\\vdots\\0\end{pmatrix}-\begin{pmatrix}1&x_{12}&\dots&x_{1K}\\0&0&0&0\\\vdots&\vdots&\vdots&\vdots\\0&0&0&0\end{pmatrix}\hat{\boldsymbol{\beta}}_1\ ,$$

we obtain for the vector of predicted values in (8):

$$\hat{\mathbf{y}}=\mathbf{X}_1\,\hat{\boldsymbol{\beta}}_1\ +\ \begin{pmatrix}y_1\\0\\\vdots\\0\end{pmatrix}\ -\ \begin{pmatrix}1&x_{12}&\dots&x_{1K}\\0&0&0&0\\\vdots&\vdots&\vdots&\vdots\\0&0&0&0\end{pmatrix}\hat{\boldsymbol{\beta}}_1 \tag{9}$$

and in particular for the first element we get

$$\hat{y}_1\ \ =\ \ \left(1\,x_{12}\,\dots\,x_{1K}\right)\hat{\boldsymbol{\beta}}_1\ +\ y_1\ -\ \left(1\,x_{12}\,\dots\,x_{1K}\right)\hat{\boldsymbol{\beta}}_1\ \ =\ \ y_1\,. \tag{10}$$

Non-unique identification and multiple regressors. The empirical example in Section 4 shows that y_m and $\hat{y}_m$ may differ substantially if more than one respondent is identified using the background information available for x_m. In this section we evaluate the fitted value $\hat{y}_m$ in this case.

If more than one respondent is "picked" by the strategic dummy (with alternative specifications given in Section 2) then the submatrix $\mathbf{X}_2$ (which actually is a vector) has the form

$$\mathbf{X}_2\ \ =\ \ \begin{pmatrix}\boldsymbol{\iota}_q\\ \mathbf{0}\end{pmatrix}$$

where we assume that q units in the data set have the same reported value for the available background information as the target record x_m and that they are placed in the first q rows of the data matrix. $\boldsymbol{\iota}_q$ is a vector of ones and $\mathbf{0}$ denotes a $n-q$ dimensional vector of zeroes. Moreover, we have

$$\mathbf{X}_2\,(\boldsymbol{X}_2'\boldsymbol{X}_2)^{-1}\boldsymbol{X}_2'\ \ =\ \ \frac{1}{q}\begin{pmatrix}\boldsymbol{\iota}_q\boldsymbol{\iota}_q'&\mathbf{0}\\ \mathbf{0}&\mathbf{0}\end{pmatrix}$$

and

$$\mathbf{X}_2 \left(\boldsymbol{X}_2'\boldsymbol{X}_2\right)^{-1}\boldsymbol{X}_2' \left(\boldsymbol{y} - \boldsymbol{X}_1\hat{\boldsymbol{\beta}}_1\right) = \tfrac{1}{q}\begin{pmatrix} \boldsymbol{\iota}_q\boldsymbol{\iota}_q' & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}\left(\boldsymbol{y} - \boldsymbol{X}_1\hat{\boldsymbol{\beta}}_1\right)$$

$$= \begin{pmatrix} \overline{y}_q \\ \overline{y}_q \\ \vdots \\ \overline{y}_q \\ 0 \\ \vdots \\ 0 \end{pmatrix} - \begin{pmatrix} 1 & \overline{x}_q^{(2)} & \dots & \overline{x}_q^{(K)} \\ 1 & \overline{x}_q^{(2)} & \dots & \overline{x}_q^{(K)} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & \overline{x}_q^{(2)} & \dots & \overline{x}_q^{(K)} \\ 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 \end{pmatrix} \hat{\boldsymbol{\beta}}_1 \ ,$$

where we use

$$\overline{y}_q = \frac{1}{q}\sum_{i=1}^{q} y_i \quad \text{und} \quad \overline{x}_q^{(k)} = \frac{1}{q}\sum_{i=1}^{q} x_{ik} \ , \ k = 2, \dots, K \ .$$

Comparing this with (9) we note that for the first q elements of the vector $\hat{\mathbf{y}}$ we obtain

$$\hat{y}_i \quad = \quad \overline{y}_q \ + \ \left(1, x_{i2} - \overline{x}_q^{(2)}, \dots, x_{iK} - \overline{x}_q^{(K)}\right)\hat{\boldsymbol{\beta}}_1 \ , i = 1, 2, \dots, q \quad . \tag{11}$$

which implies the following:

- If $q = 1$ and therefore a single unit is identified, the above result is equivalent with (10) because then $\overline{y}_q = y_1$ and for all regressors $\overline{x}_q^{(k)} = x_{1k}$.
- If the strategic dummy is used as a single regressor then for all q units

  $$\hat{y}_i \quad = \quad \overline{y}_q$$

  holds, that is, the estimated value of y equals the arithmetic mean of all q units.
- If more regressors are added to the model, no clear-cut statement regarding the difference between y_m and $\hat{y}_m$ can be made.

The Microdata Analysis System at the U.S. Census Bureau*

Jason Lucero and Laura Zayatz

U.S. Census Bureau, Statistical Research Division,
4600 Silver Hill Road, Washington, DC 20233-9100, United States
jason.lucero@census.gov

Abstract. The U.S. Census Bureau collects its survey and census data under Title 13 of the U. S. code, which promises to protect the confidentiality of our respondents. The agency has the responsibility to release high quality data products without violating the confidentiality of our respondents. This paper discuses a Microdata Analysis System (MAS) that is currently under development at the Census Bureau. We begin by discussing the reason for developing a MAS, and answer some questions about the MAS. We next give a brief overview of the MAS and the confidentiality rules within the system. The rest of this paper gives an overview of the evaluation of the universe subsampling routine in the MAS known as the *Drop Q Rule*. We conclude with some remarks on future research.

Keywords: Data Confidentiality, Remote Access Servers, Universe Uubsampling, Schur-Convexity.

1 Introduction

The U.S. Census Bureau collects its survey and census data under Title 13 of the U.S. Code. This prevents the Census Bureau from releasing any data "…whereby the data furnished by any particular establishment or individual under this title can be identified." In addition to Title 13, the Confidential Information Protection and Statistical Efficiency Act of 2002 (CIPSEA) requires the protection of information collected or acquired for exclusively statistical purposes under a pledge of confidentiality. In addition, the agency has the responsibility of releasing data for the purpose of statistical analysis. In common with most national statistical institutes, our goal is to release as much high quality data as possible without violating the pledge of confidentiality [1], [2].

This paper discusses a Microdata Analysis System (MAS) that is under development at the Census Bureau. The system is designed to allow data users to perform various statistical analyses (for example, regressions, cross-tabulations,

* This report is released to inform interested parties of ongoing research and to encourage discussion of work in progress. The views expressed are those of the author and not necessarily those of the U.S. Census Bureau.

J. Domingo-Ferrer and E. Magkos (Eds.): PSD 2010, LNCS 6344, pp. 234–248, 2010.

generation of correlation coefficients, etc.) of confidential survey and census microdata without seeing or downloading the underlying microdata. We begin by answering some frequently asked questions about the MAS. We then discuss the current state of the system, including an overview of the types of data sets and statistical analyses that will be available in the system, as well as a brief outline of the confidentiality rules used to protect the data products generated from the MAS. We next give a brief overview of a recent evaluation of a particular confidentiality rule called the *Drop q Rule*. We end with remarks on future work.

2 Frequently Asked Questions about the MAS

2.1 Why Do We Need a MAS?

The Census Bureau conducts reindentification studies on our public use microdata files. In these studies, we attempt to link outside files that have identifiers on them to our public use files. We have found and fixed a few problems, but there is a growing concern that more problems will arise in the future because more and more data is becoming publicly available on the internet, and more people are using record linkage software and data mining in an effort to increase the amount of information they can work with. We are worried that we might have to cut back on the detail in our files and use more data perturbation techniques to protect them.

Another reason for developing a MAS is to allow data users to access more detailed and accurate information than what is currently available in our public use microdata files. For example, the data that can be accessed through the MAS could identify smaller geographic areas or show more detail in variable categories that are normally not shown in our public use files. Our goal for the MAS is to allow access to as much high quality data as possible [3], [4].

2.2 What Data Sets and What Types of Statistical Analyses Will Be Available on the MAS?

We will begin with data from demographic surveys and decennial censuses. Eventually, we would like to add establishment survey and census data as well as linked data sets. We will initially begin with regression analyses, cross-tabulations and correlation matrices. We will add other analyses within the MAS in the future.

2.3 Who Will Use the MAS and Will It Cost Anything?

The MAS will be used by people with needs for fairly simple statistical analyses, for example: news media, some policy makers, teachers, and students. Some users may feel the need to use the underlying confidential microdata for more exploratory data analysis. For that, they will have to continue to use our public use files. A final decision on cost has not yet been made. However, the current plan is to offer this as a free service through the Census Bureau's DataFERRETT service [5].

2.4 Will the Census Bureau Keep Track of Who Uses the MAS and What Queries Have Been Submitted?

We must mention that we are not sure if we are allowed to legally do this. We are investigating this. There are two possible reasons why we would want to do this. First, we would like to see how people are using the system, so we can make modifications and enhancements to improve the user's experience. The second reason would be for disclosure avoidance purposes, as these data may be used to help identify disclosure risks arising from multiple queries to the system.

3 A Brief Overview of the MAS and the Confidentiality Rules within the System

In 2005, the Census Bureau contracted with Synectics to develop an alpha prototype of the MAS, which was written in SAS. We also contracted with Jerry Reiter of Duke University to help us to develop the confidentiality rules within the system, and Steve Roehrig of Carnegie Mellon University to help us test these confidentiality rules. Some rules were developed and modified as a result of the testing. The alpha prototype used the Current Population Survey (CPS) March 2000 Demographic Supplement and the 2005 American Community Survey (ACS) as initial test data sets. The alpha prototype allowed users to perform cross-tabulations, ordinary least squares regression, logistic regression, and generate matrices of correlation coefficients. A beta prototype of the MAS is now being developed as part of DataFERRET [5]. Unlike the alpha prototype, this prototype is being written in R.

The MAS software is programmed with several confidentiality rules and procedures that uphold disclosure avoidance standards. The purpose of these rules and procedures is to prevent data intruders from reconstructing the microdata records of individuals within the underlying confidential data through submitting multiple queries. The confidentiality rules discussed within the next few sections are quite complex. This paper gives a brief overview of them. Much more detail can be found in [6] and [7]. In addition, we will only discuss the confidentiality rules for universe formations and for regression models. The confidentiality rules for cross-tabulations and correlation coefficients are still under development.

3.1 Confidentiality Rules for Universe Formation

On the MAS, users are allowed to limit their statistical analysis on a universe, or subpopulation, of interest. To form a universe on the MAS, users first select conditions on a subset of recoded variables, presented to the user in the form of metadata. These recoded variables within the metadata are categorical recodes of the raw categorical and numerical variables found in the microdata.

The category levels of the raw categorical variables within the original microdata set are coded directly into the metadata. To define a universe using a categorical variable, a user simply selects the categorical variable name and observed category level bin they see in the metadata. For example, if the user selects *sex* = 2 (female)

from the metadata, they have defined their universe to be the subpopulation of all females.

Raw numerical variables are presented to the user as categorical recodes based on output from a separate cutpoint program. This cutpoint program generates buckets or bins of numerical values, and ensures that there is a pre-specified minimum number of observations between any two given cutpoint values [8]. To define universe using a numerical variable, users must a range of numerical variable values from the pre-specified bins they see from the metadata. For example, if the user selects *income* = 4 ($45,000 to $53,000) from the metadata, they have defined their universe to be the subpopulation of all individuals whose income is between $45,000 and $53,000. This furthers the confidentiality protection by preventing users from forming universes bases on a single raw numerical value. That is, users cannot define their universe to be *income* = $45,000, they must choose a range of values.

To define a universe on the MAS, users would first select m recoded variables from the metadata, then select up to j bin levels for each of the m recoded variables. Universe formation on the MAS is performed using an implicit table server. For example, suppose a data user defines their universe as:

$$[\text{gender} = \text{female and } \$45{,}000 < \text{income} \leq \$53{,}000] \quad (1)$$

OR

$$[\text{gender} = \text{male and } \$28{,}000 < \text{income} \leq \$45{,}000] \quad (2)$$

This universe is represented as a two-way table of counts for *sex* by *income*, as shown in Table 1. Piece (1) is represented by the outlined cell in Table 1, while piece (2) is represented by the set of shaded cells. Note that there $n_{24} + n_{12} + n_{13}$ total observations in this universe. For convenience, we will use the notation U(n) to indicate a universe that contains n total observations. For example, the universe defined from pieces (1) and (2) above will be referred to as U($n_{24} + n_{12} + n_{13}$).

Table 1. Table representation of the universe defined from (1) and (2)

	income				
gender	$0 to $28,000	$28,000 to $39,000	$39,000 to $45,000	$45,000 to $53,000	Total
male	n_{11}	n_{12}	n_{13}	n_{14}	$n_{1.}$
female	n_{21}	n_{22}	n_{23}	n_{24}	$n_{2.}$
Total	$n_{.1}$	$n_{.2}$	$n_{.3}$	$n_{.4}$	$n_{..}$

All universes formed on the MAS must past both of the following two confidentiality rules. If at least one of these two rules fail, the MAS will reject the universe query and prompt the universe to modify his or her selections. Note that these rules are tested prior to performing the user's selected statistical analysis on their defined universe.

The first rule is the *No Marginal 1 or 2 Rule*. No universe defined with exactly m variables on the MAS may be defined from an m-way table that contains at least one m-1 dimensional marginal total equal to 1 or 2. For example, to check the No

Marginal 1 or 2 Rule for U($n_{24} + n_{12} + n_{13}$) defined above, the following equations must be satisfied:

$$n_{i.} \neq 1 \text{ or } 2, \text{ for i} = 1,2 \text{ and } n_{.j} \neq 1 \text{ or } 2, \text{ for j} = 1,\ldots,4$$

The second rule is the Minimum Number of Observations Rule. In general, a universe must contain at least Γ observations before a user can perform a statistical analysis on this universe. The value of Γ is not given here since it is Census Confidential. Cutpoint bins are always combined to check this rule. In addition, the way this rule is checked is dependent on whether or not the universe pieces are disjoint or joint. A universe is classified as *disjoint* if its individual pieces do not share cell counts in common. For example, pieces (1) and (2) for the universe U($n_{24} + n_{12} + n_{13}$) are disjoint since they do not share any cell counts in common. Since U($n_{24} + n_{12} + n_{13}$) is a disjoint universe, the MAS would check that both piece (1) and piece (2) contain at least Γ observations. That is, both of the following equations must be satisfied. Note that the cutpoint bins of *income* are combined within piece (2) prior to performing the test.

$$n_{24} \geq \Gamma \text{ and } (n_{12} + n_{13}) \geq \Gamma$$

A universe is classified as *joint* if as least one of its individual pieces shares cell counts in common with at least one other piece. For example, suppose the user defines the following universe, U($n_{2.} + n_{.3} + n_{.4}$) = (3) OR (4), where pieces (3) and (4) are defined as:

$$[\text{gender} = \text{female}] \tag{3}$$

$$[\$39{,}000 < \text{income} \leq \$53{,}000] \tag{4}$$

U($n_{2.} + n_{.3} + n_{.4}$) is derived from the set of outlined and shaded cells in Table 2, where the outlined cells represent piece (3) and the shaded cells represent piece (4). Note that the cell counts n_{23} and n_{24} are shared among pieces (3) and (4).

Table 2. Table representation of the universe defined from (3) and (4)

	income				
gender	\$0 to \$28,000	\$28,000 to \$39,000	\$39,000 to \$45,000	\$45,000 to \$53,000	Total
male	n_{11}	n_{12}	n_{13}	n_{14}	$n_{1.}$
female	n_{21}	n_{22}	n_{23}	n_{24}	$n_{2.}$
Total	$n_{.1}$	$n_{.2}$	$n_{.3}$	$n_{.4}$	$n_{..}$

Since U($n_{2.} + n_{.3} + n_{.4}$) is a joint universe, to test the *Minimum Number of Observations Rule*, the MAS would first check that both pieces (3) and (4) contain at least Γ observations each, then check that the non-empty intersection I = (3) $\cap$ (4)

contains at least Γ* observation, where Γ* < Γ. That is, the following three equations must be satistified:

$$n_{2.} \geq \Gamma \text{ and } (n_{.3} + n_{.4}) \geq \Gamma \text{ and } (n_{23} + n_{24}) \geq \Gamma^*$$

If at least one piece (3) or (4), or if the intersection *I*, fails to pass the tests above, then the MAS will reject the entire universe. Once again, the cutpoint bins of income are first combined within piece (4) and within $I = (3) \cap (4)$ prior to the testing of the *Minimum Number of Observations* Rule for (4) or *I*.

While the above rules ensures that a universe U(*n*) meets a minimum size requirement, it does not protect against differencing attack disclosures. A *differencing attack disclosure* occurs when a data intruder attempts to rebuild a confidential microdata record by subtracting the statistical analysis results obtained through two separate and similar queries. For example, suppose a data intruder first creates two universes on the MAS: U(*n*) and U(*n*-1), where U(*n*-1) contains the exact same *n* observations as U(*n*), less one unique observation. The difference U(*n*) – U(*n*-1) = U(1), where U(1) is a manipulated universe that contains the one unique observation. Suppose further that the data intruder then requests two separate cross-tabulations for *sex* by *employment status*, T[U(*n*)] and T[U(*n*-1)], fitted on U(*n*) and U(*n*-1), respectively, as shown in Figure 1. Since U(*n*) and U(*n*-1) only differ by one unique observation, T[U(*n*-1)] will be exactly the same as T[U(*n*)], less one unique cell count.

The matrix subtraction T[U(*n*)] – T[U(*n*-1)] = T[U(1)], where T[U(1)] is a two-way table of *sex* by *employment status* built upon the one unique observation contained in U(1) = U(*n*) – U(*n*-1). As shown in Figure 1, T[U(1)] contains a cell count in the male non-employed cell with zeros in the remaining cell, which tells the data intruder that the one unique observation contained in U(1) is a non-employed male. By performing similar differencing attacks like the shown above, a data intruder can successfully rebuild the confidential microdata record for the one unique observation contained in U(1).

T[U(*n*)]	*employment status*	
sex	employed	non-employed
Male	n_1	n_2
female	n_3	n_4

-

T[U(*n*-1)]	*employment status*	
sex	employed	non-employed
male	n_1	n_2-1
female	n_3	n_4

=

T[U(1)]	*employment status*	
sex	employed	non-employed
male	0	1
female	0	0

Fig. 1. An Example of a Differencing Attack Disclosure

To help protect against differencing attack disclosures, the MAS implements a universe subsampling routine called the *Drop Q Rule*. Once a universe data set passes the universe formation rules, the MAS will first draw a random value of $Q_v = q_v \in \{2,\ldots,k\}$ from a Discrete Uniform distribution with probability mass function

$P(Q_v = q_v) = 1/(k\text{-}1)$. Then, given $Q_v = q_v$, the MAS will subsample the universe data set U(n) by removing q_v records at random from U(n) to yield a new subsampled universe data set, U(n-q_v).

On the MAS, all statistical analyses are performed on the subsampled U(n-q_v) data set and not on the original U(n) data set. Each unique universe U(n) that is defined on the MAS will be subsampled independently according to the *Drop Q Rule*. In addition, to prevent an "averaging of results" attack, the MAS will produce only one subsampled U(n-q_v) data set for each unique U(n) data set, and will fix the subsampled U(n-q_v) data set to each unique U(n) data set for the lifetime of the system. That is, if the same user, or a different user, selects the same unique U(n) data set as before, then the MAS would use the exact same subsampled U(n-q_v) data set as before for the statistical analysis.

Therefore, if the data intruder attempts the differencing attack T[U(n)] – T[U(n-1)] = T[U(1)] as shown in Figure 1, he would actually be performing the differencing attack T[U(n-q_1)] – T[U(n-1-q_2)] = T[U(1)] as shown in Figure 2, where T[U(n-q_1)] and T[U(n-1-q_2)] are two-way tables of *sex* by *employment status* based on the two independently subsampled universes, U(n-q_1) and U(n-1-q_2), where the random vectors $\mathbf{X} = \mathbf{x} = (x_1,\ldots,x_4)$ and $\mathbf{Y} = \mathbf{y} = (y_1,\ldots,y_4)$ give the number of counts that were specifically removed from each cell in T[U(n-q_1)] and T[U(n-1-q_2)], respectively, and $\Sigma_j\, x_j = q_1$,$\Sigma_j\, y_j = q_2$, $0 \le x_j \le q_1$, and $0 \le y_j \le q_2$, for $j = 1,\ldots,4$.

Since each $Q_1 = q_1$ and $Q_2 = q_2$ are independently drawn from a Discrete Uniform distribution, q_1 will not necessarily be equal to q_2, and the resulting table T[U(1)] = T[U(n-q_1)] – T[U(n-1-q_2)] may or may not yield a successful disclosure of *sex* = male and *employment status* = non-employed for the one unique observation contained in U(1). A brief overview of the effectiveness of the *Drop Q Rule* against differencing attack disclosures will be discussed in Section 4.

T[U(n-q_1)]	employment status	
Sex	employed	non-employed
Male	n_1-x_1	n_2-x_2
Female	n_3-x_3	n_4-x_4

-

T[U(n-1-q_2)]	*employment status*	
sex	employed	non-employed
male	n_1-y_1	n_2-1-y_2
female	n_3-y_3	n_4-y_4

?
=

T[U(1)]	*employment status*	
sex	employed	non-employed
male	0	1
female	0	0

Fig. 2.

3.2 Confidentiality Rules for Regression Models

The MAS also implements a series of confidentiality rules for regression models. For example, users may only select up to 20 independent variables for any single regression equation. Users are allowed to transform numerical variables only, and they must select their transformations from a pre-approved list. This prevents the user from using transformations that deliberately over emphasize outliers.

Any fully interacted regression model that contains only dummy variables as predictors can pose a potential disclosure risk [9], [10]. Therefore, users are allowed to include only two-way and three-way interaction terms within any specified regression model. No regression model that contains more than three variables can be fully interacted. Predictor dummy variables must each contain a minimum of α observations each. Any dummy variable that fails this requirement gets dropped from the regression model equation, along with the dummy variable that represents the reference category level for that particular categorical predictor variable. That is, dummy variables (or equivalently, category levels) that contain less than α observations are combined with the reference category level to prevent data intruders from fitting regression models with categorical predictors that contain sparse categories. The value of α is not given here since it is Census confidential.

Prior to passing back any regression output back to the user, the MAS checks and ensures that R^2 is not too close to 1. If R^2 is too close to 1, then the MAS will withhold from outputting any regression analysis results back to the user. If R^2 is not too close to 1, then the MAS will pass the estimated regression coefficients and the Analysis of Variance (or Deviance) table to the user without restrictions.

Actual residual values can pose a potential disclosure risk, since a data intruder can obtain the actual real values of the dependent variable by simply adding the residual to the fitted values obtained from the regression model. Therefore, the MAS never passes back real residual values back to the user. To help data users assess the fit of their Ordinary Least Squares regression models, all diagnostic plots on the MAS are based on synthetic residuals and synthetic real values. These plots are designed to mimic the actual patterns seen in the scatter plots of the real residuals vs. the real fitted values [11].

4 Evaluation of the Effectiveness of the Drop Q Rule

We will only present a brief overview of this evaluation here. Full details about this evaluation can be found in [12]. Given a pair of similar universes that only differ by one unique observation, $U(n)$ and $U(n\text{-}1)$, we investigated the effectiveness of the *Drop Q Rule* in preventing contingency table differencing attack disclosures of the form $T[U(1)] = T[U(n\text{-}q_1)] - T[U(n\text{-}1\text{-}q_2)]$, as was shown in Figure 2.

Using the same example as was shown in Section 3.1, since $U(n\text{-}q_1)$ and $U(n\text{-}1\text{-}q_2)$ are two independently subsampled universes, the resulting table $T[U(1)] = T[U(n\text{-}q_1)] - T[U(n\text{-}1\text{-}q_2)]$ (in Figure 2) may or may not a count of 1 in the shaded cell that represents the *sex* and *employment status* categories of the one unique observation contained in U(1), with zeros within the remaining three cells.

We wanted to find the probability of obtaining such a table T[U(1)] that contains a 1 in the shaded cell for s*ex* = male and *employment status* = non-employed, with zeros within the remaining three cells, from the differencing attack $T[U(n\text{-}q_1)] - T[U(n\text{-}1\text{-}q_2)] = T[U(1)]$. That is, we wanted to find the probability that the resulting table $T[U(1)] = T[U(n\text{-}q_1)] - T[U(n\text{-}1\text{-}q_2)]$ yielded a successful disclosure of *sex* and *employment status* for the one unique observation contained in U(1).

For a given value $Q_1 = q_1$ the observed vector $(X_1 = x_1,\ldots,X_4 = x_4)$ of counts that are actually removed from each cell in $T[U(n\text{-}q_1)]$ are dependent on the distribution of

cell proportions $\boldsymbol{\pi} = (\pi_1,\ldots,\pi_4)$ within the original two-way table, T[U(n)]. Similarly, for a given value $Q_2 = q_2$ the observed vector $(Y_1 = y_1,\ldots,Y_4 = y_4)$ of counts that are actually removed from each cell in T[U(n-1-q_2)] are dependent on the distribution of cell proportions $\boldsymbol{\psi} = (\psi_1,\ldots,\psi_4)$ within the original two-way table T[U(n-1)]. However, if n is large, $\boldsymbol{\pi} \approx \boldsymbol{\psi}$. Therefore, for large values of n, $\mathbf{X} \mid \boldsymbol{\pi}, Q = q_1 \dot{\sim}$ Multinomial($\boldsymbol{\pi}$, q_1) and $\mathbf{Y} \mid \boldsymbol{\pi}, Q = q_2 \dot{\sim}$ Multinomial($\boldsymbol{\pi}$, q_2). Since U(n-q_1) and U(n-1-q_2) are subsampled independently, the random tables T[U(n-q_1)] and T[U(n-1-q_2)] are also subsampled independently and the random vectors $\mathbf{X}$ and $\mathbf{Y}$ are independent. Therefore, the approximate joint probability of $\mathbf{X}|\boldsymbol{\pi}$, $Q_1 = q_1$ and $\mathbf{Y}|\boldsymbol{\pi}$, $Q = q_2$ is:

$$\begin{aligned}&P(\mathbf{X} = \mathbf{x} \mid \boldsymbol{\pi}, Q = q_1 \cap \mathbf{Y} = \mathbf{y} \mid \boldsymbol{\pi}, Q = q_2) = \\ &P(\mathbf{X} = \mathbf{x} \mid \boldsymbol{\pi}, Q = q_1)\, P(\mathbf{Y} = \mathbf{y} \mid \boldsymbol{\pi}, Q = q_2) = \\ &\left(\frac{q_1!}{x_1!\cdots x_4!}\right)\left(\frac{q_2!}{y_1!\cdots y_4!}\right)\pi_1^{x_1+y_1}\cdots\pi_4^{x_4+y_4}\end{aligned} \tag{5}$$

However, $P(Q_v = q_v) = 1/(k\text{-}1)$, therefore

$$\begin{aligned}&P([\mathbf{X} = \mathbf{x} \cap Q_1 = q_1 \mid \boldsymbol{\pi}\,] \cap [\mathbf{Y} = \mathbf{y} \cap Q_2 = q_2 \mid \boldsymbol{\pi}]) = \\ &P(\mathbf{X} = \mathbf{x} \cap \boldsymbol{\pi}, Q_1 = q_1)\, P(\mathbf{Y} = \mathbf{y} \cap \boldsymbol{\pi}, Q_2 = q_2) = \\ &P(\mathbf{X} = \mathbf{x} \mid \boldsymbol{\pi}, Q_1 = q_1)\, P(Q = q_1) P(\mathbf{Y} = \mathbf{y} \mid \boldsymbol{\pi}, Q_2 = q_2)\, P(Q_2 = q_2) = \\ &\left(\frac{1}{k-1}\right)^2\left(\frac{q_1!}{x_1!\cdots x_4!}\right)\left(\frac{q_2!}{y_1!\cdots y_4!}\right)\pi_1^{x_1+y_1}\cdots\pi_4^{x_4+y_4}\end{aligned} \tag{6}$$

Equation (6) gives us the approximate joint probability of observing a T[U(n-q_1)] with exactly $(x_1,\ldots,x_4)$ counts removed from each cell, given $Q = q_1$, and observing a T[U(n-1-q_2)] with exactly $(y_1,\ldots,y_4)$ counts removed from each cell, given $Q = q_2$, where $\boldsymbol{\pi} = (\pi_1,\ldots,\pi_4)$ are the observed cell proportions of counts within the original two-way table of *sex* by *employment status*, T[U(n)], and $\Sigma_j\, \pi_j = 1$.

T[U(n-q_1)]	employment status	
sex	employed	non-employed
male	n_1-x_1	n_2-x_2
female	n_3-x_3	n_4-x_4

-

T[U(n-1-q_1)]	*employment status*	
sex	employed	non-employed
male	n_1-x_1	n_2-1-x_2
female	n_3-x_3	n_4-x_4

=

T[U(1)]	*employment status*	
sex	employed	non-employed
male	0	1
female	0	0

Fig. 3.

We found that the resulting table, T[U(1)], would yield a successful disclosure if and only if $q_1 = q_2$ and if and only if $(x_1,\ldots,x_4) = (y_1,\ldots,y_4)$, as shown in Figure 3.Therefore, equation (9) becomes:

$$\left(\frac{1}{k-1}\right)^2\left(\frac{q_1!}{x_1!\cdots x_4!}\right)^2\pi_1^{2x_1}\cdots\pi_4^{2x_4} \tag{7}$$

Equation (7) gives us the approximate joint probability of obtaining any one pair of subsampled two-way tables, T[U(n-q_1)] and T[U(n-1-q_1)] (as shown in Figure 3), such that the same value $Q_1 = q_1$ was drawn for both subsampled universes U(n-q_1) and U(n-1-q_1), and the exact same observed vector of counts (x_1,…,x_4) were removed at random among the four cells of both T[U(n-q_1)] and T[U(n-1-q_1)], where $\Sigma_j\, x_j = q_1$. For any given value of $Q = q_1$, there are exactly C(4+q_1-1, q_1) sequences of vectors (x_1,…,x_4), such that $\Sigma_j\, x_j = q_1$. Therefore, if we sum (7) over all possible C(4+q_1-1, q_1) sequences of (x_1,…,x_4), we obtain:

$$\sum_{x_1,\ldots,x_4 \geq 0}^{x_1+\cdots+x_4=q_1} \left(\frac{1}{k-1}\right)^2 \left(\frac{q_1!}{x_1!\cdots x_4!}\right)^2 \pi_1^{2x_1} \cdots \pi_4^{2x_4} \tag{8}$$

(8) gives us the approximate joint probability of obtaining all possible pairs of subsampled tables, T[U(n-q_1)] and T[U(n-1-q_1)] (from Figure 3), for a single given observed value of $Q_1 = q_1$. However, since $Q_1 = q_1 \in \{2,\ldots,k\}$, if we sum (8) over all possible observed values q_1 can take, we obtain (9), the total approximate joint probability of observing all pairs of subsampled two-way tables, T[U(n-q_1)] and T[U(n-1-q_1)], , for all possible values of Q = $q_1 \in \{2,\ldots,k\}$. As a result, (9) gives us the approximate total probability of obtaining a successful disclosure of gender and employment status, for the one observation contained in U(1) = U(n) – U(n-1), from the differencing attack T[U(n-q_1)] – T[U(n-1-q_1)] = T[U(1)]. Proposition 1 summarizes these results to differencing attacks performed on m-way tables.

$$\sum_{q_1=2}^{k} \sum_{x_1,\ldots,x_4 \geq 0}^{x_1+\cdots+x_4=q_1} \left(\frac{1}{k-1}\right)^2 \left(\frac{q_1!}{x_1!\cdots x_4!}\right)^2 \pi_1^{2x_1} \cdots \pi_4^{2x_4} \tag{9}$$

Proposition 1: Suppose T[U(n-q_1)] and T[U(n-1-q_2)] are any two pairs of similar m-way tables that both contain J total cells each, fitted on two independently subsampled universe data sets, U(n-q_1) and U(n-1-q_2), where both U(n-q_1) and U(n-1-q_2) were subsampled according to the *Drop Q Rule*. Let (x_1,…,x_J) and (y_1,…,y_J) be the observed vector of counts that were randomly removed from each cell in T[U(n-q_1)] and T[U(n-1-q_2)], where $\Sigma_j\, x_j = q_1$, $\Sigma_j\, y_j = q_2$, $0 \leq x_j \leq q_1$, and $0 \leq y_j \leq q_2$. Then the differencing attack of T[U(n-q_1)] – T[U(n-1-q_2)] will yield an m-way table, T[U(1)], that will successfully discloses all observed category levels for all m variables for the one unique observation contained in U(1) if and only if $q_1 = q_2$, and if and only if (x_1,…,x_J) = (y_1,…,y_J) in T[U(n-q_1)] and T[U(n-1-q_1)], respectively. Let $\boldsymbol{\pi}$ = (π_1,…,π_J) be the observed cell proportions of the original m-way table T[U(n)], fitted on the full universe U(n). Then, if n is large, the approximate probability of obtaining a successful disclosure from T[U(1)] = T[U(n-q_1)] - T[U(n-1-q_2)] is:

$$\xi_{J,k}(\pi_1,\ldots,\pi_J) = \sum_{q_1=2}^{k} \sum_{x_1,\ldots,x_J \geq 0}^{x_1+\cdots+x_J=q_1} \left(\frac{1}{k-1}\right)^2 \left(\frac{q_1!}{x_1!\cdots x_J!}\right)^2 \pi_1^{2x_1} \cdots \pi_J^{2x_J} \tag{10}$$

where $\Sigma_j\, \pi_j = 1$. Note that if at least one $\pi_j = 0$, then we define $0^{2x_j} = 1$ ■

Theorem 1: The approximate probability function (10) is a Schur-convex function of $\boldsymbol{\pi} = (\pi_1,\ldots,\pi_J)$, where $\Sigma_j \pi_j = 1$ ■

The proof of Theorem 1 relies on the concepts of majorization and Schur-convexity ([11],[12] and [13]) and can be found in the appendix. Using a series of simulated differencing attacks on one-way, two-way, and three-way tables, we found that, on the average, (10) approximates the probability of obtaining a successful disclosure from the resulting table T[U(1)] = T[U(n-q_1)] - T[U(n-1-q_2)], within two decimal places.

Theorem 2: The approximate probability function $\xi_{J,k}(\boldsymbol{\pi})$ (13), subject to the linear constraint $\Sigma_j \pi_j = 1$, achieves a minimum value when $\pi_1 = \cdots = \pi_J = 1/J$ and achieves a maximum when at one $\pi_j = 1$ with the remaining $\pi_i = 0$, for $i \neq j$. Futhermore, (10) satisfies the following bounds:

$$\left(\tfrac{1}{k-1}\right)^2 \sum_{q_1=2}^{k} \sum_{x_1,\ldots,x_J \geq 0}^{x_1+\cdots+x_J=q_1} \left(\tfrac{q_1!}{x_1!\cdots x_J!}\right)^2 \left(\tfrac{1}{J}\right)^{2q_1} \leq \xi_{J,k}(\pi_1,\ldots,\pi_J) \leq \tfrac{1}{k-1}. \quad \blacksquare \quad (11)$$

The proof of Theorem 2 relies on the fact that (11) is a Schur-convex function, can be found in the appendix.

5 Future Work

The MAS will continue to be developed within Data FERRET. We will soon be testing the software itself and the confidentiality rules within the MAS beta prototype to ensure they properly uphold disclosure avoidance standards. We will draft up a set of confidentiality rules for cross-tabulations, and add different types of statistical analyses within the system. We will explore other types of differencing attack disclosures, and explore ways to prevent such differencing attacks.

References

[1] Duncan, G.T., Keller-McNulty, S., Stokes, S.L.: Disclosure Risk vs. Data Utility: The R-U Confidentiality Map. Technical Report 2003-6, Heinz School of Public Policy and Management, Carnegie Mellon University (2003)

[2] Kaufman, S., Seastrom, M., Roey, S.: Do Disclosure Controls to Protect Confidentiality Degrade the Quality of the Data? In: Proceedings of the Section on Survey Research, American Statistical Association (2005)

[3] Weinberg, D., Abowd, J., Rowland, S., Steel, P., Zayatz, L.: Access Methods for United States Microdata. In: Proceedings of the Workshop on Data Access to Microdata, Nurembourg, Germany, August 20-21 (2007); Also found on the Social Science Research Network `http://hq.ssrn.com` and US Census Bureau Center for Economic Studies Paper No. CES-WP-07-25

[4] Rowland, S., Zayatz, L.: Automating Access with Confidentiality Protection: The American FactFinder. In: Proceedings of the Section on Government Statistics, American Statistical Association (2001)
[5] Chaudhry, M.: Overview of the Microdata Analysis System. Statistical Research Division Internal Report, Census Bureau (2007)
[6] Lucero, J.: Confidentiality Rules for Universe Formation and Geographies for the Microdata Analysis System. Statistical Research Division Confidential Research Report Series No. CCRR-2009/01, Census Bureau (2009)
[7] Lucero, J.: Confidentiality Rule Specifications for Performing Regression Analysis on the Microdata Analysis System. Statistical Research Division Confidential Research Report Series #????, U.S. Census Bureau (2010) (in Progress)
[8] Lucero, J., Zayatz, L., Singh, L.: The Current State of the Microdata Analysis System at the Census Bureau. In: Proceedings of the American Statistical Association, Government Statistics Section (2009)
[9] Reznek, A.P.: Disclosure Risks in Cross Section Regression Models. In: Proceedings of the American Statistical Association, Government Statistics Section (CD-ROM). American Statistical Association, Alexandria (2003)
[10] Reznek, A.P., Riggs, T.L.: Disclosure Risks in Regression Models: Some Further Results. In: Proceedings of the American Statistical Association, Government Statistics Section (CD-ROM). American Statistical Association, Alexandria (2004)
[11] Reiter, J.P.: Model Diagnostics for Remote-Access Regression Servers. Statistics and Computing 13, 371–380 (2003)
[12] Lucero, J.: Evaluation of the Effectiveness of the Drop q Rule Against Differencing Attack Disclosures. Statistical Research Division Confidential Research Report Series No.CCRR-2010/03, U.S. Census Bureau (2010)
[13] Marshall, A.W., Olkin, I.: Inequalities: Theory of Majorization and Its Applications. Mathematics in Science and Engineering Series, vol. 143. Academic Press, London (1979)
[14] Zhang, X.: Schur-Convex Functions and Isoperimetric Inequalities. Proceedings of the American Mathematical Society 126(2), 461–470 (1998)
[15] Ben-Haim, Z., Dvorkind, T.: Majorization and Applications to Optimization (2004), unknown publication found from the web at
http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.4.9813&rep=rep1&type=pdf

Appendix A

We will use concepts from the theory of majorization and Schur-convexity to prove Theorems 1 and 2. Majorization and Schur-Convexity are useful tools that can be sometimes used to prove certain properties of functions, such as the solution to an optimization problem. The following definitions and theorems were taken from [13], [14] and [15].

Definition 1: For a given vector $\mathbf{z} = (z_1,\ldots,z_J)$, let $z_{(1)}$ denote the smallest element of $\mathbf{z}$, let $z_{(2)}$ denote the second smallest element of z, and so on. A vector z is said to majorize a vector y (denoted $z \succ y$) if

$$\sum_{i=1}^{I} z_{(i)} \geq \sum_{i=1}^{I} y_{(i)}, \text{ for I = 1,..., J – 1, and } \sum_{j=1}^{J} z_j = \sum_{j=1}^{J} y_j.$$

Lemma 1: Let $\mathbf{z} = (z_1,\ldots,z_J)$ be any arbitrary vector with $\Sigma_j z_j = s$ where $z_j \geq 0$ for $j = 1,\ldots,J$. Define the uniform vector $\mathbf{u} = \left(\frac{s}{n},\ldots,\frac{s}{n}\right)$. Then for any given vector $\mathbf{z}$, $\mathbf{z} \succ \mathbf{u}$.

Majorization is a partial ordering among any two vectors of equal dimensions, and applies only to vectors having the same sum. It is a measure of the degree to which the vector elements differ. Intuitively, the uniform vector is the vector with the minimal difference between its elements. Therefore, all other vectors z, whose elements add up to the same sum s, will majorize u.

Definition 2: For $J > 1$, a function g: $R^J \rightarrow R$ is called a symmetric function if for every permutation matrix Π, $g(\Pi(z_1,\ldots,z_J)) = g(z_1,\ldots,z_J)$.

Definition 3: For $J > 1$, a function $g(\mathbf{z})$: $R^J \rightarrow R$ is called a Schur-convex function if $\mathbf{z} \succ \mathbf{y}$ implies $g(\mathbf{z}) \succ g(\mathbf{y})$.

Schur-convex functions are functions that preserve the ordering of majorization. That is, a Schur-convex function translates the ordering of vectors to a standard scaler ordering. Symmetry is a necessary condition for a function to be Schur-convex. In addition, any function that is both symmetric and convex is a Schur-convex function. We will use the following useful lemmas to prove Theorem 1.

Lemma 2: The complete homogenous symmetric function

$$H(\pi_1,\ldots,\pi_J) = \sum_{x_1,\ldots,x_J}^{x_1+\cdots x_J=q} \pi_1^{x_1} \cdots \pi_J^{x_J} \tag{12}$$

Is a Schur-convex function of $\boldsymbol{\pi} = (\pi_1,\ldots,\pi_J)$.

Lemma 3: Let $\varphi(\mathbf{z})$: $R^J \rightarrow R$ be a Schur-convex function of $\mathbf{z} = (z_1,\ldots,z_J)$ and let $g(z_j)$: $R \rightarrow R$ be a convex function of z_j, for $j = 1,\ldots,J$. Then the composition $\Phi(\mathbf{z}) = \varphi(g(z_1),\ldots,g(z_J))$, $\Phi(\mathbf{z})$: $R^J \rightarrow R$ is a Schur-Convex function of $\mathbf{z}$.

Lemma 4: Let the function $\psi(z_1,\ldots,z_J)$ be an increasing, real valued function defined on R^J, and let $\varphi_1,\ldots, \varphi_k$ be real valued Schur-Convex functions, each with common domain $A \subset R^J$. Then the composition: $\Phi(z_1,\ldots,z_J) = \psi(\varphi_1(z_1,\ldots,z_J),\ldots, \varphi_k(z_1,\ldots,z_J))$ is a Schur-convex function of $\mathbf{z} = (z_1,\ldots,z_J)$.

Lemma 5: Let $c > 0$ be any constant, and let $z_k \geq 0$ for all $k = 1,\ldots,K$. Then the function $\psi(z_1,\ldots,z_K) = c(\Sigma_k z_k)$ is an increasing function on R^K.

The following Lemma will be useful to prove Theorem 1:

Lemma 6: The function

$$F_{J,q}(\pi_1,\ldots,\pi_J) = \sum_{x_1,\ldots,x_J}^{x_1+\cdots x_J=q} \left(\frac{q!}{x_1!,\ldots,x_J!}\right)^2 \pi_1^{2x_1}\cdots\pi_J^{2x_J} \tag{13}$$

is a Schur-convex function of $\boldsymbol{\pi} = (\pi_1,\ldots,\pi_J)$, where $\Sigma_j\, \pi_j = 1$, $0 \le \pi_j \le 1$ for all j, and the summation in (13) is taken over all possible sequences of $\Sigma_j\, x_j = q$, for any given value of q. (Note: if one $\pi_j = 0$, then we define $0^{2x_j} = 1$)

Proof: (13) is a symmetric function on the set R^J. To apply Lemma 3 to (13), define $g(\pi_j) = a_j\pi_j^2$, where $0 \le \pi_j \le 1$ for all j, and

$$a_j = \begin{cases} \left(\frac{q!^{\frac{1}{J}}}{x_j!}\right)^{\frac{2}{x_j}} & if\ x_j \ge 1 \\ 1 & if\ x_j = 0 \end{cases}$$

and set $\varphi(\pi_1,\ldots,\pi_J) = H(\pi_1,\ldots,\pi_J)$ (12) from Lemma 2. Since (12) is a Schur-convex function of $\boldsymbol{\pi} = (\pi_1,\ldots,\pi_J)$, and $g(\pi_j) = a_j\pi_j^2$ is convex for each π_j. Then, by Lemma 3 the compostition

$$\Phi(\boldsymbol{\pi}) = H(g(\pi_1),\ldots,g(\pi_J)) = \sum_{x_1,\ldots,x_J\ge 0}^{x_1+\cdots+x_J=q} \left[\left(\frac{q!^{\frac{1}{J}}}{x_1!}\right)^{\frac{2}{x_1}}\pi_1^2\right]^{x_1}\cdots\left[\left(\frac{q!^{\frac{1}{J}}}{x_J!}\right)^{\frac{2}{x_J}}\pi_J^2\right]^{x_J}$$

$$= \sum_{x_1,\ldots,x_J\ge 0}^{x_1+\cdots+x_J=q} \left(\frac{q!^{\frac{1}{J}}}{x_1!}\right)^2 \pi_1^{2x_1}\cdots\left(\frac{q!^{\frac{1}{J}}}{x_J!}\right)^2 \pi_J^{2x_J} = \sum_{x_1,\ldots,x_J\ge 0}^{x_1+\cdots+x_J=q} \left(\frac{q!^{\frac{1}{J}}\cdots q!^{\frac{1}{J}}}{x_1!\cdots x_J!}\right)^2 \pi_1^{2x_1}\cdots\pi_J^{2x_J}$$

$$= \sum_{x_1,\ldots,x_J\ge 0}^{x_1+\cdots+x_J=q} \left(\frac{q!}{x_1!\cdots x_J!}\right)^2 \pi_1^{2x_1}\cdots\pi_J^{2x_J}$$

is a Schur-convex function of $\boldsymbol{\pi} = (\pi_1,\ldots,\pi_J)$. ■

Proof of Theorem 1: (10) is a symmetric function of $\boldsymbol{\pi} = (\pi_1,\ldots,\pi_J)$ on the set R^J. To apply Lemma (4) to (10), set

$$\psi(z_2,\ldots,z_k) = \left(\frac{1}{k-1}\right)^2 \sum_{q1=2}^{k} z_{q_1} \text{ for } z_2,\ldots,z_k \ge 0 \tag{14}$$

$$\text{and set } \varphi_{q_1} = \sum_{x_1,\ldots,x_J}^{x_1+\cdots x_J=q_1} \left(\frac{q_1!}{x_1!,\ldots,x_J!}\right)^2 \pi_1^{2x_1}\cdots\pi_J^{2x_J} \tag{15}$$

By Lemma 5, (14) is an increasing function of $z_2,\ldots,z_k$ on the set R^{k-1}, and (15) is a Schur-convex function by Lemma 6, for a given value of $Q_1 = q_1$. Therefore, by Lemma (4), the composition:

$$\xi_{J,k}(\pi_1,\ldots,\pi_J) = \psi(\varphi_2(\pi_1,\ldots,\pi_J),\ldots,\varphi_k(\pi_1,\ldots,\pi_J)) = \left(\tfrac{1}{k-1}\right)^2 \sum_{q_1=2}^{k} \varphi_{q_1}(\pi_1,\ldots,\pi_J)$$

$$= \left(\tfrac{1}{k-1}\right)^2 \sum_{q_1=2}^{k} \sum_{x_1,\ldots,x_J\geq 0}^{x_1+\cdots+x_J=q_1} \left(\frac{q_1!}{x_1!\cdots x_J!}\right)^2 \pi_1^{2x_1}\cdots\pi_J^{2x_J}$$

is a Schur-convex function of $\boldsymbol{\pi} = (\pi_1,\ldots,\pi_J)$. ■

Lemma 7 will be used to help prove Theorem 2.

Lemma 7: Let $\varphi(z_1,\ldots,z_n)$ be a Schur-convex function. Suppose we wish to find the minimum of $\varphi(z_1,\ldots,z_n)$ given the linear constraint $\Sigma_j z_j = s$. Then, since $\varphi(z_1,\ldots,z_n)$ is Schur-convex, the minimum of φ is achieved when $\mathbf{z} = \mathbf{u} = \left(\frac{s}{n},\ldots,\frac{s}{n}\right)$.

Proof: By Lemma 1, the uniform vector $\mathbf{u}$ is majorized by any other vector $\mathbf{z}$ that has the same sum s. Since $\varphi(\mathbf{z})$ is a Schur-convex function, by Definition 3, $\mathbf{z} \succ \mathbf{u}$ implies $\varphi(\mathbf{z}) \geq \varphi(\mathbf{u})$. Therefore, the minimum of $\varphi(z_1,\ldots,z_n)$, subject to the constraint $\Sigma_j z_j = s$, is achieved at $\mathbf{z} = \mathbf{u}$. ■

Proof of Theorem 2: Since (10) is a Schur-convex function, by Lemma 7, the minimum of (10) subject to the constraint $\Sigma_j \pi_j = 1$ is achieved when $\boldsymbol{\pi} = \mathbf{u} = \left(\frac{1}{J},\ldots,\frac{1}{J}\right)$, and the minimum of (10) is:

$$\left(\tfrac{1}{k-1}\right)^2 \sum_{q_1=2}^{k} \sum_{x_1,..,x_J\geq 0}^{x_1+\cdots+x_J=q_1} \left(\frac{q_1!}{x_1!\cdots x_J!}\right)^2 \left(\tfrac{1}{J}\right)^{2q_1} \tag{16}$$

In addition, since (10) is a Schur-convex function, it is symmetric. Therefore, for $j = 1,\ldots,J$, for any given permutation matrix Π, $\xi_{J,k}(\Pi\ (\pi_1,\ldots,\pi_J)) = \xi_{J,k}(\pi_1,\ldots,\pi_J)$. Furthermore, for any given vector $\boldsymbol{\pi} = (\pi_1,\ldots,\pi_J)$ whose elements sum to 1 and $0 \leq \pi_j \leq 1$ for all J, it is easy to check that $(1,0,\ldots,0) \succ \boldsymbol{\pi} \succ \mathbf{u}$, which implies

$$\tfrac{1}{k-1} = \xi_{J,k}(\pi_1,\ldots,\pi_J) \succ \xi_{J,k}(\boldsymbol{\pi}) \succ \xi_{J,k}(\mathbf{u}). \tag{17}$$

Combining (16) and (17) gives (11). ■

Establishing an Infrastructure for Remote Access to Microdata at Eurostat

Wolf Heinrich Reuter[1] and Jean-Marc Museux[2]

[1] Department of Economics, Vienna University of Economics and Business, Althanstrasse 39-45, A-1090 Vienna, Austria
Wolf.Reuter@wu.ac.at
[2] Methodology and Research Unit, Eurostat, L-2920 Luxembourg
Jean-Marc.Museux@ec.europa.eu

Abstract. Eurostat is pursuing the establishment of an infrastructure for remote access for researchers in order to satisfy the growing demand for microdata. Some European countries already implemented such solutions. This paper compares the systems which can be categorized in (1) terminal server, (2) distance network and (3) job submission systems. They differ in IT infrastructure, workstation control, user management and authentication, file systems and disclosure control activities. The second part of the paper describes the efforts and outlook as well as options and challenges for Eurostat when building such a system.

Keywords: remote access, infrastructure, microdata, Eurostat.

1 Introduction

Researchers are asking for ever more and easier access to microdata in order to provide the policy makers with in-depth analysis of policies defined at EU level. Competing on the open market for information, Eurostat tends to consider the release of microdata for scientific purpose as a core tasks for National Statistical Organizations. Today access to EU wide unprotected microdata is only available by traveling to Luxembourg (Eurostat). Though some NSIs already implemented remote access solutions to provide researchers with the requested data. Through new technical opportunities and a changing European legal framework Eurostat is also starting to implement a remote access system.

We define remote access to microdata as follows: A properly identified person can directly work with the un-anonymised but de-identified microdata from a safe place. During the whole process the microdata stays at the access point. The physical security of the data and the risk of data leaking are ensured by strong IT requirements all along the process.

The intention is to build a common remote access infrastructure used by several Eurostat projects which in future could make use of accessing confidential data from a location outside of Eurostat. It is expected that the Eurostat remote access infrastructure would serve as a central node for a shared infrastructure which has a positive impact on the release of confidential data for scientific purpose but also for

J. Domingo-Ferrer and E. Magkos (Eds.): PSD 2010, LNCS 6344, pp. 249–257, 2010.

the production of European Statistical Systems (ESS) statistics. Currently such an infrastructure is defined and set-up. The test phase of the remote access for researchers along various organizational models is planned for the spring of 2012. In the long-term a decentralized system with data stored at the national authorities' level and accessible directly by researchers could be envisaged. In the future, synergies can be developed with the data archives specializing in data dissemination and in providing services around the access to confidential microdata.

In chapter 2 we summarize the existing remote access to microdata infrastructures in Europe. In Chapter 3 we provide a review of the efforts taken by Eurostat in the past and an outlook to the following steps. Chapter 4 tackles the challenges and decisions to be taken in order to establish such an infrastructure at Eurostat. Chapter 5 concludes.

2 Remote Access to Microdata in Europe

2.1 Overview

The European countries Denmark, France, Luxembourg, The Netherlands, and Sweden already implemented a remote access to microdata solution. In principle there are three different approaches: a terminal server solution using the CITRIX or Microsoft remote desktop technology (implemented by The Netherlands, Denmark, Sweden[1]), a distance network solution in combination with taking control over the client workstation (France) and a submission system automatically processing statistical batch code (Luxembourg[2]). We will focus on the first two remote access solutions (direct access to microdata), rather than the remote execution solution (no direct access to microdata).

Terminal server solution: The required remote access software can be installed on any workstation having access to the internet. The software connects to the remote server using a secure (e.g. HTTPS) channel. Within this remote client applications can be started which behave as if they were installed on the researcher's computer. But in fact only the screen output is sent to the client and the application is actually run completely at the server independently of the client.

Distance network solution: The researcher's workstation is completely integrated into the remote network through a secure Virtual Private Network (VPN) tunnel. The local network just has to provide an internet connection. The workstation behaves as if the computer is allocated and connected at the remote location. All network methods can be applied to this workstation, e.g. authentication via central directory, domain controller, etc. The applications are run either locally on the workstation or using virtual desktops at a server in the network.

The Netherlands and Denmark implemented the terminal server solution relying on the CITRIX and Sweden on the Microsoft Terminal Services technology. France has developed a distance network solution together with controlling the workstation of the researcher by limiting the access to the network to so called SD-Boxes©. The following chapters will describe various aspects of the solutions in more detail.

[1] Similar to the solution used e.g. at the University of Chicago (NORC).
[2] Similar to the solution used e.g. at the ABS in Australia.

2.2 IT Infrastructure

The three terminal server solutions in principle are based on a similar IT infrastructure (see annex 1 for graphical presentation) consisting of (1) a web/proxy-server, acting as a gateway / portal and tunneling all traffic to the remote access environment, (2) a central user directory including the access rights and credentials[3], (3) a domain controller, (4) a file server holding the working files of the users and the provided databases and (5) a server farm handling the terminal services and running the statistical applications. Some or all of the servers can be virtualized reducing the demand for physical servers.

The Dutch and Swedish network is based on a Windows Server and the Danish on a UNIX system. The remote access environment is always strictly separated from the production environment. Data exchange between these environments is handled either manually or via an automatic process using SecureFTP.

The network infrastructure of the distance network solutions in France is similar to a standard Windows network consisting of a SSTP gateway, a domain controller, an active directory and a file server. The maintenance of this solution thus is also similar to the maintenance of the standard corporate Windows network.

All systems use automatic disk-to-disk or band back-ups to secure a copy of the working and management files in case of a break-down. It should be noticed that also the back-ups have to be secured against disclosure.

The applications provided for the researcher to run on the server farms once in the remote access environment are statistical (e.g. SAS, SPSS and STATA), mathematical (e.g. R, GAMS, GAUSS) and editing (e.g. Office, Acrobat, LaTeX) software. Taking advantage of the applications running independently of the client on the servers all systems also allow the researchers to start a calculation process, disconnect from their sessions (application is still running on the server) and reconnect later to obtain the results.

2.3 Client Workstation

Denmark and Sweden do not control the workstations of the researcher and let them install the required remote access software on their own. In the Netherlands staff of the statistical institute visits the working place of the researcher in order to install the necessary soft- (remote client application, server root certificate) and hardware (smartcard and fingerprint reader) and to check the computer and environment[4] the researcher's workstation is placed in.

The French system can only be accessed using a SD-Box©. This box in principle is a standard workstation assembled from native components. Though the functionality is controlled by hard- and software restrictions which are implemented in the hardware or imposed by the remote network. The box checks its integrity and the

[3] In France no biometrics (fingerprints) are stored at the central servers due to the countries law situation. The fingerprints are stored and validated against the information on the smartcard.

[4] Room has to be lockable, workstation is separated from other researchers, etc.

connection to the remote network[5] before booting and is blocked otherwise. Only the remote access functionality is available for the researcher, all other functionalities of the workstation[6] are hidden.

2.4 Users and Authentication

In all countries access to the system is subject to the authorization of the research institution and the individual research project. In the Netherlands and France each research project has to pursue a statistical or scientific / historical purpose and to make (all) its results publicly available in order to get accepted.

In the Netherlands only research institutes mentioned by Dutch law or approved by the Central Statistics Committee are allowed to get access. In France the approval is given by the "comité du secret" and in case of data from administrative processes also the agreement of the original administration will be necessary. Due to the different legal framework in Denmark and Sweden also other research organizations and private companies[7] are granted access.

The enrollment of the user and configuration of the authentication credentials in the Netherlands is done during the visit of a statistical staff at the researcher's institution to install the hard- and software. In France the researcher has to travel to the statistical institute in order to get the required hardware (SD-Box© and keyboard with fingerprint and smartcard reader) and get enrolled.

Sweden and Denmark are using a combination of username / password and RSA token to identify the researcher. In the Netherlands and France a combination of biometric (fingerprint) and smart card authentication is used. In Denmark, the Netherlands and Sweden one researcher (using its authentication credentials) has access to several virtual accounts depending on the number of contracts / projects the researcher is working on. In France a new smart-card is needed for every project.

2.5 Database and File Systems

In the Netherlands the dataset are stored as flat files (CSV) in the remote access environment. In Denmark and France as SAS files[8]. All databases are not anonymised but de-identified, meaning that all direct identifiers[9] have been deleted or changed beyond recognition.

No custom-build datasets are produced in the Netherlands. In Denmark all data files are custom-build and reduced for every project following the "need-to-know principle"[10]. Afterwards the researcher can work with the data freely and make new datasets from the original data sets as well as on request let databases be linked by the statistical department. Also in France data can be matched by the statistical staff after approval of the "comité du secret".

In all systems the working files of the researchers are strictly divided between different projects. In the Dutch, Swedish and Danish system all user accounts working

[5] Using the SSTP protocol to establish a VPN.
[6] In principle a standard operating system and BIOS is running on the box.
[7] Excluding media organizations and restricting access to at least one year old business-data.
[8] In principle also CSV files.
[9] Names, addresses, personal numbers, etc.
[10] Data should not be more comprehensive than necessary for carrying out the project.

under the same contract can share data through a central shared folder. In the French system all projects are encapsulated in virtual machines on the server with no data exchange.

2.6 Disclosure Control

All described systems prevent the researcher from getting data out of the remote access system. All data stays at the server environment at all time. The solutions prevent the researchers from printing, copying[11] or transferring[12] data. Furthermore all important activities are stored in the log-files of the various systems.

A contract is signed between the respective statistical institute and the researcher to hold the researcher responsible for any breach of confidentiality. In France and the Netherlands the contract is also signed with the research institution. The penalty for breaking the contract is ranging from the denial of any access to statistics for this institution (The Netherlands) to imprisonment (Denmark).

Due to the legal framework in Denmark and Sweden the output files of the researchers can be directly sent to them via encrypted email. The output is stored at the servers though and checked randomly by the statistical department. In Netherlands and France all output intended to leave the secure environment is checked by respective staff on disclosure issues and afterwards forwarded via encrypted email.

3 Eurostat Efforts

In order to foster the establishment of an ESS remote access system Eurostat has organized several workshops[13] and took part in various international initiatives[14]. Additionally three European collaboration projects related to the subject have been initiated or completed:

- “ESS collaborative network (ESSnet) on Statistical Disclosure Control - SDC I” (2006-2007) & “SDC II” (2008-2009) comprised various tasks in the SDC domain among which the output checking guidelines are of high relevance for the current issue.
- “ESSnet on decentralized access to EU microdata sets” (2009-2010) studied the feasibility of different scenarios of setting up a network of safe centers allowing researchers to access confidential EU microdata sets.

[11] The "copy and paste" function is deactivated in the terminal server solutions and useless in the French system because there is no place the data could be copied to.

[12] All ports and outbound internet connections are blocked in the remote access environments.

[13] Every 2 years since 2006 Eurostat is organizing workshops on microdata access. In 2006 the main subject of the seminar was discussion about UNECE guidelines on "Managing Statistical Confidentiality and microdata access, Principles and guidelines of good practice" and risk management approach. In 2008 the workshop was organized together with CESSDA to investigate the possible collaborations.

[14] Eurostat follows international initiatives in the domain of microdata access. Since many years regular work sessions are organized with UNECE on statistical data confidentiality. Eurostat is also involved in the OECD/Australian Bureau of Statistics initiative investigating the possibility of international remote access to microdata.

Various aspects related to such access, be they legal, administrative, methodological or technical were studied.

- FP7 project "Data without Boundaries" (European Social Science Data Archives and remote access to Official Statistics) (upcoming) aims at fostering the integration of data archives and NSIs and to develop the corresponding infrastructure (methodological and logistic). If the project is successful, it will be necessary to prepare for the integration and implementation of the project outcome into the ESS.

Since 2008 Eurostat is actively pursuing the Establishment of an infrastructure for remote access to confidential data. The infrastructure will be set-up in a way to allow a three steps development. Step (1) allows the researchers to get access to the Eurostat data from the safe centers located at the NSIs – researcher do not have to travel to Luxembourg anymore but to their nearest safe centre, implementing the recommendations of the ESSnet on decentralized access, (2) researchers also get access to the system from their workplaces and (3) a decentralized remote access infrastructure is established including the infrastructures of the NSIs and possibly the national data archives.

Step (1) is feasible under the current legislative framework (Regulation EC No 831/ 2002). The further developments will depend on the changes in the legal framework (new implementing regulation replacing Regulation (EC) No 831/2002[15]) and on the results of the above mentioned projects.

4 Way Forward at Eurostat

A series of decisions have to be taken by Eurostat often in collaboration with the NSIs. Most important about the security levels needed and the involvement in the process of the NSIs.

The security level will become more important in step (2) when researchers also get access from their workplaces. There are various technical solutions to identify a researcher (fingerprints, facial recognition, smart card, RSA token, username / password and any combination). If there is a need to have full control over the soft- and hardware of the workstation connecting to the remote access environment a solution like the French one with allowing only specific computers[16] which secure their integrity to connect to the system, should be preferred. It has to be noted that the costs of providing the hardware rises from username / password authentication, through the biometrics / smartcard reader, to the secure workstation.

First cost model estimates predict that typical infrastructure costs of between 300.000 and 350.000 EUR depending on the options selected (biometrics, custom-build applications, complete workstation deployment, etc.). Additionally yearly costs of approximately 60.000 EUR are estimated for the maintenance of the infrastructure and 2.000 - 2.500 EUR per project for the surrounding tasks, e.g. disclosure control, administrative control, project applications, etc.

[15] The high-level Working Group on Statistical Confidentiality (WGSC) has appointed a Task Force already working and proposing a first draft of the new regulation until September 2010.

[16] In France this would be the SD-Box©.

In the long term perspective, it is considered that some standardized tasks could be split between NSIs and Eurostat or delegated to selected partners: (1) Managing the application process of a research project and consequently the user and permissions management, (2) storage and maintenance of the microdata, (3) installation of hard- and software and enrollment for researcher, (4) giving support and answering requests (infrastructure, data, methodology related) and (5) being in charge of the output checking process and consequently answering to the researcher respectively forwarding the cleared output.

IT Infrastructure: Currently only few researchers are making use of the safe centre located at Eurostat to access the EU wide microdata. With the establishment of the remote access this number will surely increase significantly. Therefore in the long run an infrastructure is needed to cope with a large number of concurrent users. Though in step (1) when remote access is only granted from within the safe centers of the NSIs approximately 27 or little more[17] concurrent users will be possible. Thus depending on the decision concerning the security a standard sized (average of four physical servers) terminal services or distance network infrastructure will be build. The statistical software available in the remote access environment will depend on the level of support intended to offer by Eurostat or the NSIs and their respective capability to support the software.

Client Workstation: Also depending on the security level decision either the complete workstation (similar to the SD-Box©) will be distributed to the NSIs and researchers or staff of Eurostat / NSIs will install the hard- and software at the client location. The complete workstation solution will improve the security by having control over all software running on the computer and thus also preventing screen prints, automatic data detection or transfer applications. An important challenge is the integration of the respective solution in the network infrastructures of the client (NSIs or research institution). Both solutions will only need one open port in the firewalls but nevertheless foreign software or even hardware (complete workstation solution) granting outside access into the corporate networks will have to be installed.

Users and Authentication: The current process to gain access to EU wide microdata will stay the same until the new implementing regulation will come into effect and possibly change the procedure. Also the contracts and obligations of research institution, researcher and project are regulated there. In step (1) no sophisticated biometrics / smartcard / token authentication would be needed due to the fact that the researcher will be identified in person when entering the safe centre. Afterwards security aspects, costs and compatibility with the workstations and networks will be crucial to the decision which identification credentials are necessary.

Database and Filesystem: The data is stored as flat files at Eurostat and no custom-build or matching of databases would be allowed due to the current legal framework and practice.

Disclosure Control: The microdata in the system will be de facto deidentified data. But there might be an option, that for very sensitive data, on the request of MSs some standard disclosure methods will be applied directly to the microdata. As today all data / output taken out of the secure environment, i.e. the remote access environment,

[17] The safe centers in 27 member states, but some safe centres might offer more than one workstation.

has to be checked on disclosure issues manually. This should be done decentralized using for example the output checking guidelines available on the CASC website (neon.vb.cbs.nl/casc/). The punishment for breaking the confidentiality is regulated in member states private laws rather than European law.

5 Conclusion and Outlook

Eurostat is pursuing the establishment of a remote access infrastructure to keep up with the leading statistical institutes and provide requested data for the researchers. This procedure depends on various decisions, regulatory changes and discussions with member states and stakeholders.

In the meantime new technological and methodological possibilities will emerge and improve, which should steadily be taken into account in the decision and design process. For example automatic output checking supporting the manual procedure is already implemented at the remote access solution of the Australian Bureau of Statistics. In this system only unclear or difficult cases have to be taken up by the statistical staff. Or for example new statistical software running in the web browser and allowing access to the data without additional soft- or hardware.

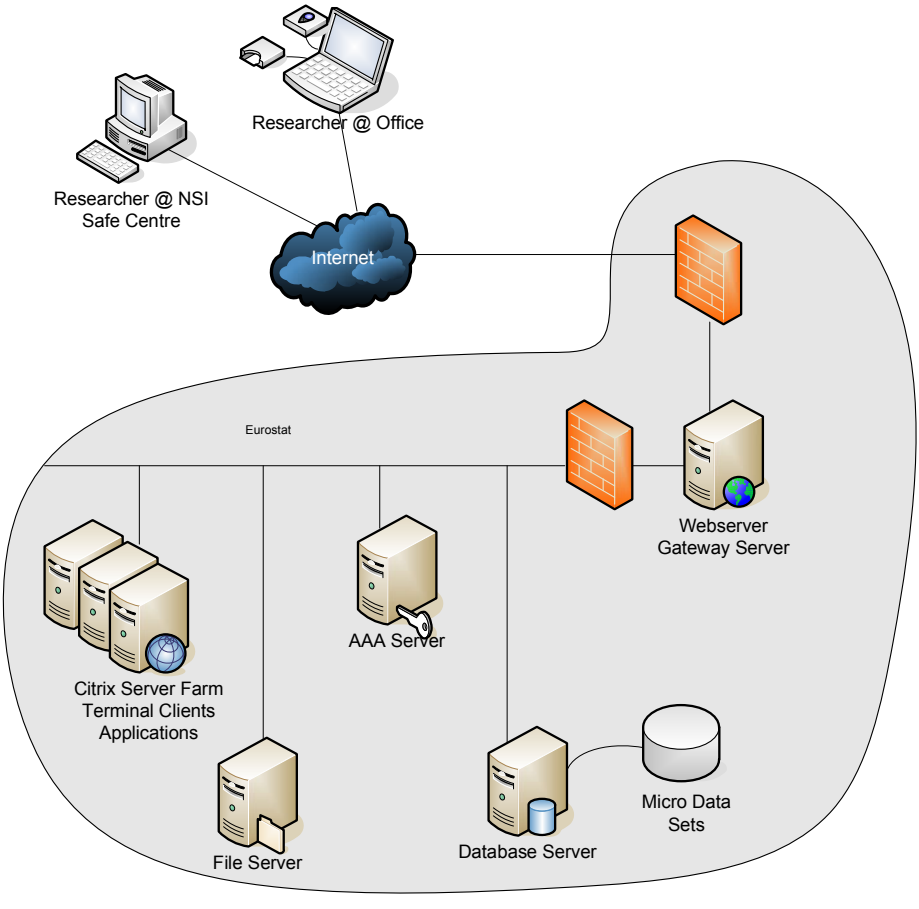
The expanded access in general also requires expanded procedural and specific legal protections, as well as various standards. Furthermore everyone working with confidential records requires education and training in disclosure issues and practices.

References

1. Andersen, O.: From on-site to remote data access - the revolution of the Danish system for access to microdata. In: Joint ECE/Eurostat work session on statistical data confidentiality, Luxembourg, April 7-9 (2003)
2. Borchsenius, L.: Recent developments in the Danish system for access to microdata. In: OECD Conference: Assessing the feasibility of micro-data, Luxembourg, October 26-27 (2006)
3. Borchsenius, L.: New developments in the Danish system for Access to Microdata. In: Joint UNECE/Eurostat work session on statistical data confidentiality, Geneva, November 9-11 (2005)
4. Coder, J., Cigrang, M.: LISSY Remote Access System. In: Joint ECE/Eurostat work session on statistical data confidentiality, Luxembourg, April 7-9 (2003)
5. ESSnet Project: Decentralised Access to EU-Microdata Sets. Final Report (January 2010)
6. Flemming, P., Petersen, J.K., Schnor, O., Husted, L.: Microdata for research and analysis - potential and problems. Paper for the Siena Group on Social Statistics meeting, Helsinki, February 9-11 (2005)
7. Gadouche, K., Donnay, P.: Presentation of CASD – Remoteaccess to microdata in France, Luxembourg, May 10 (2010)
8. Grim, R., Heus, P., Mulcahy, T., Ryssevik, J.: Secure Remote Access system for an upgraded CESSDA RI (September 2009)
9. Hjelm, C.G.: Remote access solution in Sweden. In: Workshop on microdata access, Luxembourg, April 12 (2008)
10. Kruten, T.: A remote access solution - The Lissy System at LIS Asbl. In: Workshop on Microdata Access, New developments and a way forward, Luxembourg, December 3-4 (2008)

11. Leth-Sørensen, S.: The Danish System for Access to Microdata, Nürnberg (2007)
12. Royer, J.F.: Microdata access for researchers in INSEE-France, Recent developments, open questions. In: 2nd workshop on data access, Cardiff, February 13-14 (2009)
13. Schillings, M., Franken, R.: Remote Access Micro Data facility - Technical Design. Statistics Netherlands at Heerlen, March 19 (2010)
14. Söderberg, L.J.: MONA – Microdata on-line access at Statistics Sweden. In: Joint UNECE/Eurostat work session on statistical data confidentiality, Geneva, November 9-11 (2005)
15. Thygesen, L., Andersen, O., Schnor, O.: The Danish System for Access to Microdata – From on-site to remote access. In: CAED Conference 2003, London, September 15-16 (2003)
16. United Nations: Managing Statistical Confidentiality & Microdata Access, Principles and Guidelines of good practice, United Nations Publication, Sales No. E.07.II.E.7, New York and Geneva (2007)

Annex 1 – IT Infrastructure Schematic Overview



Coprivacy: Towards a Theory of Sustainable Privacy

Josep Domingo-Ferrer

Universitat Rovira i Virgili
UNESCO Chair in Data Privacy
Department of Computer Engineering and Mathematics
Av. Països Catalans 26, E-43007 Tarragona, Catalonia
josep.domingo@urv.cat

Abstract. We introduce the novel concept of coprivacy or co-operative privacy to make privacy preservation attractive. A protocol is coprivate if the best option for a player to preserve her privacy is to help another player in preserving his privacy. Coprivacy makes an individual's privacy preservation a goal that rationally interests other individuals: it is a matter of helping oneself by helping someone else. We formally define coprivacy in terms of Nash equilibria. We then extend the concept to: i) general coprivacy, where a helping player's utility (*i.e.* interest) may include earning functionality and security in addition to privacy; ii) mixed coprivacy, where mixed strategies and mixed Nash equilibria are allowed with some restrictions; iii) correlated coprivacy, in which Nash equilibria are replaced by correlated equilibria. Coprivacy can be applied to any peer-to-peer (P2P) protocol. We illustrate coprivacy in P2P user-private information retrieval, and also in content privacy in on-line social networking.

Keywords: Coprivacy, Data privacy, User-private information retrieval, Content privacy in social networks, Game theory.

1 Introduction

The motivation of the coprivacy concept and its incipient theory presented in this paper is one of double sustainability in the information society:

1. *Privacy preservation is essential to make the information society sustainable.* This idea, which we already introduced in [6], should lead to clean information and communications technologies (ICT) offering functionality with minimum invasion of the privacy of individuals. Such an invasion can be regarded as a virtual pollution as harmful in the long run to the moral welfare of individuals as physical pollution is to their physical welfare.
2. *Privacy preservation itself should be sustainable, and be achieved as effortlessly as possible as the result of rational co-operation rather than as an expensive legal requirement.* Indeed, even if privacy was acclaimed as a fundamental right by the United Nations in article 12 of the Universal Declaration

J. Domingo-Ferrer and E. Magkos (Eds.): PSD 2010, LNCS 6344, pp. 258–268, 2010.

of Human Rights (1948), relying on worldwide legal enforcement of privacy is nowadays quite unrealistic and is likely to stay so in the next decades: as noted in [18], privacy needs a strong democratic society. However, unlike law, technology is global and can enforce privacy worldwide, provided that privacy is achieved as the result of rational cooperation. This is the objective of the coprivacy concept and theory presented in this paper.

Two major pollutants of privacy are privacy-unfriendly security and privacy-unaware functionality. *Privacy-unfriendly security* refers to the tendency of sacrificing privacy with the excuse of security: this is done by governments (*e.g.* the former UK security and intelligence co-ordinator asserted in 2009 that anti-terror fight will need privacy sacrifice) and by corporations (*e.g.* biometrics enforced on customers with the argument of fighting identity theft). *Privacy-unaware* (let alone privacy-unfriendly) *functionality* is illustrated by search engines (Google, Yahoo, etc.), social networking web services, Web 2.0 services (*e.g.* Google Calendar, Streetview, Latitude) and so on, which concentrate on offering enticing functionality for users while completely disregarding their privacy. At most, privacy vs third parties is mentioned, but not privacy of the user vs the service provider itself, who becomes a big brother in the purest Orwellian sense.

1.1 Contribution and Plan of This Paper

The environmental analogy above can be pushed further by drawing inspiration on the three "R" of environment: reducing, reusing and recycling.

Reducing. Re-identifiable information must be reduced. This is the idea behind database anonymization: *e.g.* k-anonymization [17] by means of microdata masking methods (*e.g.*, [4]) reduces the informational content of quasi-identifiers. Reduction is also the idea behind ring and group signatures [3,11], which attempt to conciliate message authentication with signer privacy by reducing signer identifiability: the larger the group, the more private is the signer. Just as in the environment there are physical limits to the amount of waste reduction, in the privacy scenario there are functionality and security limits to reduction: completely eliminating quasi-identifiers dramatically reduces the utility of a data set (functionality problem); deleting the signature in a message suppresses authentication (security problem). A useful lesson that can be extracted from reduction is *privacy graduality*: privacy preservation is not all-or-nothing, it is a continuous magnitude from no privacy to full privacy preservation.

Reusing. The idea of reusing is certainly in the mind of impersonators mounting replay attacks, but it can also be used by data protectors to gain privacy. Such is the case of re-sampling techniques for database privacy: an original data set with N records is re-sampled M times with replacement (where M can be even greater than N) and the resulting data set with M records is released instead of the original one. This is the idea behind synthetic data generation via multiple imputation [16]. However, as it happened for

reduction there are functionality limitations to data reuse: the more reuse, the less data utility.

Recycling. The idea of recycling is probably more intriguing and far less explored than reducing and reusing. Adapted to the privacy context, recycling can be regarded as leveraging other people's efforts to preserve their privacy to preserve one's own privacy. Of course, there is a functionality toll to privacy recycling: one must adjust to other people's ways. Nonetheless, we believe that *recycling has an enormous potential in privacy preservation, as it renders privacy an attractive and shared goal, thereby making it easier to achieve and thus more sustainable*. In this spirit, we next introduce a new recycling concept, called *coprivacy*, around which this proposal is centered.

Section 2 defines coprivacy and some of its generalizations. Section 3 illustrates coprivacy in the context of peer-to-peer (P2P) user-private information retrieval. Section 4 illustrates correlated coprivacy applied to attribute disclosure in social networks. Section 5 lists conclusions and open research issues.

2 Coprivacy and Its Generalizations

We introduce in this section the novel concept of coprivacy in a community of peers, whereby one peer recycles to her privacy's benefit the efforts of other peers to maintain their own privacy. Informally, there is coprivacy when the best option for a peer to preserve her privacy is to help another peer in preserving his privacy. The great advantage that *coprivacy makes privacy preservation of each specific individual a goal that interests other individuals*: therefore, privacy preservation becomes *more attractive* and hence *easier to achieve and more sustainable*. A formal definition of coprivacy follows.

Let $P_1, \cdots, P_N$ be the players in a game. Denote by S^i the set of player P^i's possible strategies. For each strategy $s^i_j \in S^i$, let $u_i(s^i_j)$ be the privacy utility of s^i_j for P^i, where a higher utility means higher overall privacy preservation for P^i vs the other players. Further, let

$$s^{i*}_{u_i} = \arg\max_{s^i_j \in S^i} u_i(s^i_j)$$

be the best strategy for P^i.

Definition 1 (Coprivacy). *Let Π be a game with peer players $P^1, \cdots, P^N$, and an optional system player P^0. Each player may have leaked a different amount of private information to the rest of players before the game starts. The game is as follows: i) P^1 selects one player P^k with $k \in \{0\} \cup \{2, \cdots, N\}$ and submits a request to P^k; ii) If $k = 0$, P^0 always processes P^1's request; if $k > 1$, P^k decides whether to process P^1's request (which may involve accessing the system player on P^1's behalf) or reject it. The players' strategies are $S^0 = \{s^0_1\}$ (process P^1's request); $S^1 = \{s^1_0, s^1_2, \cdots, s^1_N\}$, where s^1_j means that P^1 selects P^j; for $i > 1$, $S^i = \{s^i_1, s^i_2\}$, where s^i_1 means processing P^1's request and s^i_2 rejecting it. Game*

Π is said to be coprivate *with respect to the set $U = (u_1, \cdots, u_N)$ of privacy utility functions if $s^{1*}_{u_1} = s^1_k$ for some $k > 1$ such that $s^{k*}_{u_k} = s^k_1$, that is, if a peer P^k exists such that (s^1_k, s^k_1) is a pure strategy Nash equilibrium [14,15] between P^1 and P^k.*

An intuition on the above definition is that there is coprivacy if the best strategy for player P^1 to preserve her privacy is to ask some player P^k for help, and the best strategy for player P^k to preserve his privacy is to provide the help requested by P^1. Note that the notions of privacy utility function and therefore of coprivacy are based on the aforementioned privacy graduality: one can have a varying degree of privacy preservation, hence it makes sense to trade it off. A quantification of coprivacy follows:

Definition 2 (δ-Coprivacy). *Given $\delta \in [0, 1]$, the game of Definition 1 is said to be δ-coprivate with respect to the set $U = (u_1, \cdots, u_N)$ of privacy utility functions if the probability of it being coprivate for U is at least δ.*

The following extensions of coprivacy are conceivable:

- **General coprivacy** can be defined by replacing the set U of privacy utility functions in Definition 1 with a set $\mathcal{U}$ of general utility functions for peer players P^k combining privacy preservation with security and/or functionality. In general coprivacy, the interests of peers include, in addition to privacy, functionality and/or security.
- **General δ-coprivacy** can be defined by replacing U with $\mathcal{U}$ in Definition 2.
- **Mixed coprivacy** results if one allows mixed strategies for players and replaces the requirement of pure strategy Nash equilibrium in Definition 1 by a mixed strategy Nash equilibrium. The good point of mixed coprivacy is that a theorem by Nash [14] guarantees that any game with a finite set of players and a finite set of strategies has a mixed strategy Nash equilibrium, and is therefore *mixedly coprivate*.
- **Correlated coprivacy** results if one replaces the requirement of pure Nash equilibrium in Definition 1 by a correlated equilibrium. Indeed, the outcome of independent rational behavior by users, provided by Nash equilibria, can be inferior to a centrally designed outcome. Correlated equilibria resulting from coordination of strategies may give a higher outcome. We will illustrate this in Section 4 below.
- The above extensions can be combined to yield **mixed general coprivacy** and **correlated general coprivacy**. Since mixed coprivacy is always achievable if any mixed strategy is valid for any player, **mixed δ-coprivacy** and **mixed general δ-coprivacy** only make sense when players have boundary conditions that define a subset of feasible mixed strategies. The same holds for correlated coprivacy, which is also always achievable.

If a privacy preservation problem can be solved by using a protocol based on a coprivate game, the advantage is that it is in a player's rational privacy interest to help other players to preserve their privacy.

3 Coprivacy in P2P User-Private Information Retrieval

Private information retrieval (PIR) is normally modeled as a game between two players: a user and a database. The user retrieves some item from the database without the latter learning which item was retrieved. Most PIR protocols are ill-suited to provide PIR from a search engine or large database, not only because their computational complexity is linear in the size of the database, but also because they (unrealistically) assume active cooperation by the database in the PIR protocol.

Pragmatic approaches to guarantee some query privacy have therefore been based so far on two relaxations of PIR: standalone and peer-to-peer (P2P). In the standalone approach, a program running locally in the user's computer either keeps submitting fake queries to cover the user's real queries (TrackMeNot, [12]) or masks the real query keywords with additional fake keywords (GooPIR, [7]). In the P2P approach, a user gets her queries submitted by other users in the P2P community; in this way, the database still learns which item is being retrieved, but it cannot obtain the real query histories of users, which become diffused among the peer users, thereby achieving user-private information retrieval (UPIR). We first proposed a P2P UPIR system in [8].

Consider a system with two peers P^1 and P^2, who are interested in querying a database DB playing the role of system player P^0. If P^1 originates a query for submission to DB, she can send the query directly to DB or ask P^2 to submit the query on P^1's behalf and return the query results. The roles of P^1 and P^2 are exchangeable.

More formally, for $i, j \in \{1, 2\}$ and $i \neq j$, the strategies available for a requesting P^i are:

Sii: P^i submits her query directly to DB;
Sij: P^i forwards her query to P^j and requests P^j to submit the query on P^i's behalf.

When receiving P^i's query, P^j has two possible strategies:

Tji: P^j submits P^i's query to DB and returns the answer to P^i;
Tjj: P^j ignores P^i's query and does nothing.

Let $X^i(t)$ be the set of queries originated by P^i, let $Y^i(t)$ be the set of queries submitted to DB and $Y^{ij}(t)$ be the set of queries forwarded by P^i to P^j with $j \neq i$ up to time t. The privacy utility function for P^i should reflect the following intuitions: (i) the more "distant" is $X^i(t)$ from $Y^i(t)$, the more private is P^i vs DB; (ii) the more "distant" is $X^i(t)$ from $Y^{ij}(t)$, the more private is P^i vs P^j. Given a distance $d(\cdot, \cdot)$ between sets of queries, plausible utilities for a requesting P^i under strategies Sii and Sij at time $t + 1$ are:

$$U_{Sii}(t+1) = (d(X^i(t+1), Y^i(t+1)))^{\alpha_{i,DB}} (d(X^i(t+1), Y^{ij}(t)))^{\alpha_{i,j}}$$

$$U_{Sij}(t+1) = (d(X^i(t+1), Y^i(t)))^{\alpha_{i,DB}} (d(X^i(t+1), Y^{ij}(t+1)))^{\alpha_{i,j}}$$

where $\alpha_{i,DB}$ and $\alpha_{i,j}$ are weights in $[0,1]$ denoting how critical is for P^i privacy in front of DB and j, respectively. The utilities for the requested player P^j follow.

$$U_{Tji}(t+1) = (d(X^j(t+1), Y^j(t+1)))^{\alpha_{j,DB}} (d(X^j(t), Y^{ji}(t)))^{\alpha_{j,i}}$$

Since P^j does nothing under Tjj, we have

$$U_{Tjj}(t+1) = U_{Tjj}(t) = (d(X^j(t), Y^j(t)))^{\alpha_{j,DB}} (d(X^j(t), Y^{ji}(t)))^{\alpha_{j,i}}$$

If the α-values are all identical, the above privacy utilities are maximized when the distance from the set of originated queries to the set of submitted queries is equal to the distance from the set of originated queries to the set of forwarded queries.

Assume all α values are identical. Assume also that $X^i(t)$ and $Y^i(t)$ are "closer" than $X^i(t)$ and $Y^{ij}(t)$. Since maximum privacy utility is obtained when the within-pair distances are equal to each other, the interest of P^i is to increase the distance between $X^i(t)$ and $Y^i(t)$, that is, to submit a new query via P^j (strategy Sij); formally, we have $U_{Sij}(t+1) > U_{Sii}(t+1)$. Assume also that $X^j(t)$ and $Y^j(t)$ are "closer" than $X^j(t)$ and $Y^{ji}(t)$. Hence, the interest of P^j is to increase the distance between $X^j(t)$ and $Y^j(t)$ and this can be done by accepting to submit P^i's query to DB (strategy Tji); formally, $U_{Tji}(t+1) > U_{Tjj}(t+1)$. Under both closeness assumptions above, (Sij, Tji) is a pure-strategy Nash equilibrium between P^i and P^j and *there is coprivacy* between P^1 and P^2.

We give a detailed formalization and empirical results for the N-player P2P user-private information retrieval game in the manuscript [9].

4 Correlated Coprivacy in Social Networks

Social networking web sites or, for short, social networks (SNs) have become an important web service with a broad range of applications: collaborative work, collaborative service rating, resource sharing, friend search, etc. Facebook, MySpace, Xing, etc., are well-known examples. In an SN, a user publishes and shares information and services.

There are two types of privacy in SNs:

– *Content privacy.* The information a user publishes clearly affects her privacy. Recently, a privacy risk score [13] has been proposed for the user to evaluate the privacy risk caused by the publication of a certain information. Let the information attributes published by the users in an SN be labeled from 1 to n. Then the privacy score risk of user j is

$$PR(j) = \sum_{i=1}^{n} \sum_{k=1}^{\ell} \beta_{ik} \times V(i,j,k)$$

where $V(i,j,k)$ is the visibility of user j's value for attribute i to users which are at most k links away from j and β_{ik} is the sensitivity of attribute i vs those users.

– *Relationship privacy.* In some SNs, the user can specify how much it trusts other users, by assigning them a trust level. It is also possible to establish several types of relationships among users (like "colleague of", "friend of", etc.). The trust level and the relationship type are used to decide whether access is granted to resources and services being offered (*access rule*). The availability of information on relationships (trust level, relationship type) has increased with the advent of the Semantic Web and raises privacy concerns: knowing who is trusted by whom and to what extent discloses a lot about the user's thoughts and feelings. For a list of related abuses see [2]. In [5], we described a new protocol offering private relationships in an SN while allowing resource access through indirect relationships without requiring a mediating trusted third party.

We focus here on content privacy in SNs. A possible privacy-functionality score for user j reflecting the utility the user derives from participating in an SN is the amount of information the user learns from the other SN users divided by the amount the user discloses to them. This rational view of disclosure suits better SNs for professional contact (where employers and professionals target their disclosures) than SNs for personal contact (where users often disclose a lot without requiring much in return). A formalization of this privacy-functionality score is

$$PRF_1(j) = \frac{\sum_{j'=1,j'\neq j}^{N} \sum_{i=1}^{n} \sum_{k=1}^{\ell} \beta_{ik} V(i,j',k) I(j,j',k)}{1 + PR(j)}$$

$$= \frac{\sum_{j'=1,j'\neq j}^{N} \sum_{i=1}^{n} \sum_{k=1}^{\ell} \beta_{ik} V(i,j',k) I(j,j',k)}{1 + \sum_{i=1}^{n} \sum_{k=1}^{\ell} \beta_{ik} V(i,j,k)}$$

where $I(j,j',k)$ is 1 if j and j' are k links away from each other, and it is 0 otherwise.

Note that:

– $PRF_1(j)$ decreases as the privacy score $PR(j)$ in its denominator increases, that is, as user j discloses more of her attributes;
– $PRF_1(j)$ increases as its numerator increases; this numerator adds up the components of privacy scores of users $j' \neq j$ due to those users disclosing attribute values to j.

The dichotomous version of the above privacy-functionality score, for the case where an attribute is simply either made public or kept secret, is:

$$PRF_2(j) = \frac{\sum_{j'=1,j'\neq j}^{N} \sum_{i=1}^{n} \beta_i V(i,j')}{1 + PR(j)}$$

$$= \frac{\sum_{j'=1,j'\neq j}^{N} \sum_{i=1}^{n} \beta_i V(i,j')}{1 + \sum_{i=1}^{n} \beta_i V(i,j)} \quad (1)$$

If we regard $PRF_1(j)$ (resp. $PRF_2(j)$) as a game-theoretic utility function [19], the higher $PRF_1(j)$ (resp. $PRF_2(j)$), the higher the utility for user j.

For instance, take a strategy vector $s = (s_1, \cdots, s_N)$ formed by the strategies *independently and selfishly* chosen by all users and consider the dichotomous case, that is, let the utility incurred by user j under strategy s be $u_j(s) = PRF_2(j)$. It is easy to see (and it is formally shown in [10]) that rational and independent choice of strategies leads to a Nash equilibrium where no user offers any information on the SN, which results in the SN being shut down. See Example 1 below.

A similar pessimistic result is known for the P2P file sharing game, in which the system goal is to leverage the upload bandwidth of the downloading peers: the dominant strategy is for all peers to attempt "free-riding", that is, to refuse to upload [1], which causes the system to shut down.

Example 1. The simplest version of the above game is one with two users having each one attribute, which they may decide to keep hidden (a strategy denoted by H, which implies visibility 0 for the attribute) or publish (a strategy denoted by P, which implies visibility 1). Assuming a sensitivity $\beta = 1$ for that attribute and using $u_j(s) = PRF_2(j)$, the user utilities for each possible strategy vector are as follows:

$$u_1(H,H) = 0; u_1(H,P) = 1; u_1(P,H) = 0; u_1(P,P) = 1/2$$

$$u_2(H,H) = 0; u_2(H,P) = 0; u_2(P,H) = 1; u_1(P,P) = 1/2$$

This simple game can be expressed in matrix form:

User 2 / User 1	H	P
H	0 0	0 1
P	1 0	1/2 1/2

The above matrix corresponds to the Prisoner's Dilemma [19], perhaps the best-known and best-studied game. Consistently with our argument for the general case, it turns out that (H, H) is a dominant strategy, because:

$$u_1(H,P) = 1 \geq u_1(P,P) = 1/2; u_1(H,H) = 0 \geq u_1(P,H) = 0$$

$$u_2(P,H) = 1 \geq u_1(P,P) = 1/2; u_2(H,H) = 0 \geq u_2(H,P) = 0$$

The second and fourth equations above guarantee that (H, H) is a Nash equilibrium (in fact, the only one). The Prisoner's Dilemma with $N > 2$ users is known as the Pollution Game [19] and corresponds to the dichotomous SN game considered above.

The outcome of independent rational behavior by users, provided by Nash equilibria and dominant strategies, can be inferior to a centrally designed outcome. This is clearly seen in Example 1: the strategy (P, P) would give more utility than (H, H) to *both* users. However, usually no trusted third-party accepted by all users is available to enforce correlated strategies; in that situation, the problem is how User 1 (resp. User 2) can guess whether User 2 (resp. User 1) will choose P.

Using a solution based on cryptographic protocols for bitwise fair exchange of secrets would be an option, but it seems impractical in current social networks, as it would require a cryptographic infrastructure, unavailable in most SNs.

A more practical solution to this problem may be based on direct reciprocity (*i.e.* tit-for-tat) or reputation, two approaches largely used in the context of P2P file-sharing systems. We describe in [10] two correlated equilibrium protocols based on tit-for-tat and reputation, respectively. They are intended as "assistants" to the human user of the SN in deciding whether to disclose an attribute to another user; however, the ultimate decision belongs to the human, who may quit and renounce to reach the equilibrium.

Those correlated equilibrium protocols offer *correlated general coprivacy*, referred to a utility combining privacy and functionality.

5 Conclusions and Research Directions

We have introduced in this paper the novel concept of coprivacy, as well as an incipient generalization theory of it. The main contribution of coprivacy is to make data privacy an attractive feature, especially in peer-to-peer applications:

- In many situations, players can better preserve their own privacy if they help other players in preserving theirs. We say that those situations can be handled by so-called coprivate protocols.
- In other situations, the utility of players consists of a combination of privacy plus security and/or functionality. If they can increase their own utility by helping others in increasing theirs, the situation can be handled by a generally coprivate protocol.

We have shown that P2P private information retrieval can be solved with a coprivate protocol. Furthermore, we have shown that content privacy in social networks can be solved with a generally coprivate protocol.

Future research directions include developing the theory of coprivacy in the following non-exhaustive directions:

- Develop a theory of coprivacy which, given a privacy preservation problem and a parameter $\delta \in [0, 1]$, can answer under which conditions a δ-coprivate game (*i.e.* protocol) that solves the problem exists.
- Elaborate a theory of general coprivacy which also takes security and functionality into account. In this generalization, the Nash or the correlated equilibrium that characterizes coprivacy is to be reached by considering utilities which combine the privacy with the security and/or the functionality obtained by the players.

– Elaborate a theory of mixed coprivacy to characterize when mixed strategies and therefore mixed coprivacy make sense for utilities about privacy, security and functionality.
– Create new cryptographic protocols to implement the privacy graduality needed in coprivacy. Specifically, *ad hoc* broadcast encryption and anonymous *ad hoc* broadcast encryption inspired in [20], (n, N)-anonymity signatures and some multiparty computation protocols for social networks are needed.

Acknowledgments and Disclaimer

This work was partly funded by the Spanish Government through projects TSI2007-65406-C03-01 "E-AEGIS" and CONSOLIDER INGENIO 2010 CSD2007-00004 "ARES", and by the Government of Catalonia through grant 2009 SGR 1135. The author is partly supported as an ICREA-Acadèmia researcher by the Government of Catalonia. He holds the UNESCO Chair in Data Privacy, but the views expressed in this paper are his own and do not commit UNESCO.

References

1. Babaioff, M., Chuang, J., Feldman, M.: Incentives in peer-to-peer systems. In: Nisan, N., Roughgarden, T., Tardos, É., Vazirani, V.V. (eds.) Algorithmic Game Theory, pp. 593–611. Cambridge University Press, Cambridge (2007)
2. Barnes, S.B.: A privacy paradox: social networking in the United States. First Monday 11(9) (2006)
3. Chaum, D., van Heyst, E.: Group signatures. In: Davies, D.W. (ed.) EUROCRYPT 1991. LNCS, vol. 547, pp. 257–265. Springer, Heidelberg (1991)
4. Domingo-Ferrer, J., Sebé, F., Solanas, A.: A polynomial-time approximation to optimal multivariate microaggregation. Computers & Mathematics with Applications 55(4), 717–732 (2008)
5. Domingo-Ferrer, J., Viejo, A., Sebé, F., González-Nicolás, Ú.: Privacy homomorphisms for social networks with private relationships. Computer Networks 52, 3007–3016 (2008)
6. Domingo-Ferrer, J.: The functionality-security-privacy game. In: Torra, V., Narukawa, Y., Inuiguchi, M. (eds.) MDAI 2009. LNCS, vol. 5861, pp. 92–101. Springer, Heidelberg (2009)
7. Domingo-Ferrer, J., Solanas, A., Castellà-Roca, J.: $h(k)$-Private information retrieval from privacy-uncooperative queryable databases. Online Information Review 33(4), 720–744 (2009)
8. Domingo-Ferrer, J., Bras-Amorós, M., Wu, Q., Manjón, J.: User-private information retrieval based on a peer-to-peer community. Data and Knowledge Engineering 68(11), 1237–1252 (2009)
9. Domingo-Ferrer, J., González-Nicolás, Ú.: Peer-to-peer user-private information retrieval: a game-theoretic analysis (2010) (manuscript)
10. Domingo-Ferrer, J.: Rational privacy disclosure in social networks. In: Proc. of MDAI 2010. LNCS (2010, to appear)

11. Groth, J.: Fully anonymous group signatures without random oracles. In: Kurosawa, K. (ed.) ASIACRYPT 2007. LNCS, vol. 4833, pp. 164–180. Springer, Heidelberg (2007)
12. Howe, D.C., Nissenbaum, H.: TrackMeNot: Resisting surveillance in web search. In: Lessons from the Identity Trail, pp. 409–428. Oxford University Press, Oxford (2009)
13. Liu, K., Terzi, E.: A framework for computing the privacy scores of users in online social networks. In: Proc. of ICDM 2009-The 9th IEEE International Conference on Data Mining, pp. 288–297 (2009)
14. Nash, J.: Non-cooperative games. Annals of Mathematics 54, 289–295 (1951)
15. Nisan, N., Roughgarden, T., Tardos, É., Vazirani, V.V. (eds.): Algorithmic Game Theory. Cambridge University Press, Cambridge (2007)
16. Rubin, D.B.: Discussion on statistical disclosure limitation. Journal of Official Statistics 9(2), 461–468 (1993)
17. Samarati, P.: Protecting respondents' identities in microdata release. IEEE Transactions on Knowledge and Data Engineering 13(6), 1010–1027 (2001)
18. Solove, D.J.: Understanding Privacy. Harvard University Press, Cambridge (2008)
19. Tardos, É., Vazirani, V.V.: Basic solution concepts and computational issues. In: Nisan, N., Roughgarden, T., Tardos, É., Vazirani, V.V. (eds.) Algorithmic Game Theory, pp. 3–28. Cambridge University Press, Cambridge (2007)
20. Wu, Q., Mu, Y., Susilo, W., Qin, B., Domingo-Ferrer, J.: Asymmetric group key agreement. In: Joux, A. (ed.) EUROCRYPT 2009. LNCS, vol. 5479, pp. 153–170. Springer, Heidelberg (2010)

Privacy-Preserving Record Linkage

Rob Hall and Stephen E. Fienberg

Department of Statistics, Machine Learning Department, and Cylab,
Carnegie Mellon University, Pittsburgh PA 15213, USA
rjhall@cs.cmu.edu, fienberg@stat.cmu.edu

Abstract. Record linkage has a long tradition in both the statistical and the computer science literature. We survey current approaches to the record linkage problem in a privacy-aware setting and contrast these with the more traditional literature. We also identify several important open questions that pertain to private record linkage from different perspectives.

1 Introduction

Record linkage is an historically important statistical problem arising when data about some population of individuals, is spread over several files. Most of the literature focuses on the two file setting. The record linkage goal is to determine whether a record from one file corresponds to a record of a second file, in the sense that the records describe the same individual. Winkler and others describe application areas, computational techniques and statistical underpinnings in detail in [19,2,38,39]. The typical purposes of record linkage are:

- data integration.
- as an intermediate step in performing a computation on the integrated data.
- to create a public use file that will allow others to analyze the integrated data.

The overarching goal of privacy-preserving datamining (PPDM) [37] is to perform "data mining" computations on a set of data, in a manner that prevents both the computation, and the output of the computation from revealing "too much" sensitive information about the units represented in the data. Our goal in this paper is to detail recent advances at the intersection of record linkage and PPDM, largely as a followup to an earlier survey by Winkler [39,40]. Whereas Winkler assumed that all the data files were accessible to the party running the computation, in our setting we remove this assumption. Instead, depending on the setting and the problem at hand, we are interested in access to files that may be somehow restricted, or not available at all.

Record linkage already plays a large role as a building block for privacy preserving statistical analysis. For example, numerous papers already tacitly assume that the files that are input to their procedures are a-priori matched, in the sense that the correspondence between the units is known [24,23,22,12,13]. We describe

J. Domingo-Ferrer and E. Magkos (Eds.): PSD 2010, LNCS 6344, pp. 269–283, 2010.

some key challenges at the interface of record linkage and PPDM and show the steps various authors have taken to address them. We overview the basic record linkage approach and the secure multiparty computation literature with the intent of demonstrating some common failure modes of so called privacy preserving schemes. Then we survey the recent literature on privacy preserving schemes for performing record linkage, and conclude by outlining what we see are the key unsolved challenges in this area.

2 Record Linkage Overview

We begin by providing an overview of the record linkage problem in a non-private setting. We see the traditional approaches as being composed of a series of different steps, which we explain in turn. We then give these steps the privacy-preserving treatment in section 5. Currently, the literature on record linkage involving two files is fairly mature [39,19], whereas the problem of linking many files has only begun to be studied (see for example the discussions in the context of merging files for the purposes of multiple capture-recapture [14,19,32]). Therefore in this paper we focus on the problem of record linkage between two files.

2.1 Problem Definition

Suppose there are two data files A and B, each of which contains possibly different numbers of records, say $a_i,\ (i = 1 \ldots n)$ are the records belonging to file A and likewise $b_j,\ (j = 1 \ldots m)$ are the records in B. The records are in essence vectors in which each component is a "field" or an attribute of the record, and we may regard the records as being the elements of the product space of the fields. For the purpose of this exposition we suppose that the fields in the two files are the same (or otherwise somehow the data has been cleaned ahead of time). When this is not the case then the problem is called "schema matching" —see [33] for a treatment of this topic. Suppose that there are p fields that are common to the files. We denote by a_i^k the k^{th} field of record a_i and likewise for B. In the database terminology, records correspond to rows of a file, whereas the fields correspond to columns. The goal of record linkage is to determine the pairs of records (a_i, b_j) corresponding to the same underlying individual.

Fellegi and Sunter formally studied this problem in their seminal paper [11]. They described an approach that partitioned the cartesian product of the files $A \times B$ into three disjoint sets: M the set of "matches", U the set of "non-matches", and C a set which requires human intervention in order to classify. The presence of this latter set is due to ambiguity in the data which is hard or impossible for an automated procedure to solve. For example, several people with a common first initial and last name may inhabit the same house, and so further data may be required to determine whether or not two records correspond to the same individual of such a household. The Felegi-Sunter approach [11] aims to minimize the cardinality of C, subject to a user-specified upper bound on

the error rates in M and U. There are several modifications of this approach, a number of which are described in [19].

2.2 Computing Similarity of Record Pairs

In essence, most modern statistical record linkage techniques build on the Fellegi-Sunter idea and follow a common pattern. In a first stage, the cartesian product $A \times B$ is preprocessed and cleaned. Then some "similarity function" is applied to each element in the resulting file. Historically, the functions were indicators of whether corresponding fields of the records matched or not, i.e., whether their values for a particular fields were identical. These binary flags are referred to as the "match variables." Let $m_{i,j}$ be the vector of match variable corresponding to the pair (a_i, b_j). We may have:

$$m_{i,j} \in \{0,1\}^p,\ m_{i,j}^k = \mathbf{1}\{a_i^k = b_i^k\}$$

where we use $\mathbf{1}\{\cdot\}$ to mean the function that takes value 1 when the predicate in the braces is true, and 0 otherwise. In principle, there could be more match variables than fields, as multiple different similarity functions could be applied to different pairs of fields. For simplicity we omit a discussion of this variation here. The alternative to exact match indicators is to compute a distance function for the individual fields [8]. When fields are numeric then perhaps absolute or euclidian difference is appropriate. When fields are strings such as names and addresses, then string edit distances [21,2] are useful. Such distance measures may be thresholded, i.e., reduced to binary match variables where the flag is "on" whenever the distance falls below some cutoff. In this case, we may have:

$$m_{i,j} \in \{0,1\}^p,\ m_{i,j}^k = \mathbf{1}\{d^k(a_i^k, b_i^k) < \tau^k\}$$

where $d^k(\cdot,\cdot)$ is the appropriate distance function for field k, and τ_k is some parameter that determines the thresholding. After this first step, there are $n \times m$ sets of match variables, corresponding to the pairs of elements in the product of the files. The match variables are either binary or real numbers depending on what kinds of similarity functions that were applied.

2.3 Parameter Estimation

In the second step, we estimate the parameters of two models, namely the conditional probabilities of the match variables, given that the records match: $p_\theta(m_{i,j}|(a_i, b_j) \in M)$ and the probability for the match variables given that the records don't match $p_\theta(m_{i,j}|(a_i, b_j) \in U)$. Here the notation $p_\theta(\cdot)$ is used to mean a probability density or mass function which is parameterized by some vector θ of parameters.

If there is plentiful labeled data (i.e., hand linked records of a similar nature) to use for estimation, then we may estimate these parameters analytically using a simple maximum likelihood approach [19]. In the absence of such data (the

usual situation for PPDM) estimation is more problematic. Nevertheless, we can often use the EM algorithm [19]. Generally, there is not enough data to estimate a completely general model for the match variables, so instead some we impose additional structure [38]. Historically, a useful method was to restrict the models to force conditional independence of the individual match variables. Winkler [40] provides a discussion of more structured approaches, and Ravikumar et al. [30] give a specific model with good performance.

2.4 Classification of Record Pairs

Since we are treating record linkage as a statistical problem, it is unlikely that every record pair will be labeled correctly as a link or a non-link. Nevertheless, we can tradeoff the amount of error in the final linkage against the amount of pairs sent for clerical review. As Felligi and sunter demonstrated, the classification of a particular pair (a_i, b_j) into M, U, C may be done by considering the likelihood ratio of $m_{i,j}$ under the two models:

$$r_{i,j} = \frac{p_\theta(m_{i,j}|(a_i, b_j) \in M)}{p_\theta(m_{i,j}|(a_i, b_j) \in U)}$$

As Fellegi and Sunter [11] show, the optimal decision rule is given by:

$$\psi(a_i, b_j) = \begin{cases} M & r_{i,j} > C_1 \\ C & C_0 \leq r_{i,j} \leq C_1 \\ U & r_{i,j} < C_0 \end{cases}$$

This rule is essentially a simple test of hypothesis. One chooses constants C_0, C_1 for user-specified error levels for false-links and false non-links [36]. The rule is optimal in the sense that among the classification rules with that achieve these error rates, this rule assigns the fewest records for clerical review.

2.5 Blocking

When the sizes of the data files to be linked are moderate (e.g., tens of thousands of records or more) then applying the above theory may be too inefficient, since we would have to consider hundreds of millions of pairs. A common way to deal with this problem is to perform a "blocking" phase in which we remove clear non-links, leaving blocks of potential links. The terminology goes back in some sense to the census uses where the population is divided into physical blocks, but also reflects the experimental design notion of "blocking" to remove heterogeneity.

The idea is that a "reliable" field such as zip code or gender may be used to quickly label some of the non-links. See [19] for discussion. The result is a tradeoff of computational efficiency versus accuracy in the final linkage, however the impact on the accuracy is usually fairly mild.

3 Overview of Privacy Preserving Data Mining

The field of "privacy preserving data mining" (PPDM) primarily focuses on performing useful data analysis in such a way as to mitigate the risk of releasing some private or secret information. On the surface, there are two distinct sets of problems in this field. The first set includes problems of how two or more separate parties each with private data, may compute some function of the union of their data without having to reveal it. The second set focuses on how to determine whether the result of a computation alone constitutes an invasion of privacy (a identifiable release), and if so how to mitigate the release. When two parties need to link their private data and then perform some computation on the resulting linked records, both facets of PPDM are important to respect. In this section, we give a brief overview of the salient features of the field, the goal being to build enough sophistication to understand the subtleties of record linkage in a private setting.

3.1 Secure Multiparty Computation

Suppose two parties each hold a separate piece of private data which they would benefit from jointly analyzing. For example, the parties may be administrators of hospitals or government agencies, who are bound by law to not disclose the information of individuals in their databases. Nevertheless they may wish to join their data to that of some medical research center or another agency in order to fit a statistical model to the union of their data. Performing such computations is the concern of a mature area in the PPDM literature called "Secure Multiparty Computation" (SMC) see e.g., [27,26] for an overview. The goal is to develop protocols consisting of local computations by individual parties, and the transmitting of messages between the parties. Depending on the demands of the parties involved, one of several models of security may be appropriate.

Perhaps the most well studied and rigorous formulation of a secure computation comes from cryptography [17,16]. The idea is that the protocol should reveal no more information than would a fanciful "idealized" method in which the private data are presented to a completely trusted third party, who performs the computation and returns the results to each of the original parties. That is, to any specific party, the computation itself should reveal no more than whatever may be revealed by examining his input and output. An example of a protocol that would fail to meet this criteria is if one party was sent all the private inputs, performed the computation locally and then broadcast the results to the other parties. The reason this fails is because, in general, the party who does the computation cannot infer the other's data just from looking at his data and the result, and so the messages passed in the proposed protocol has revealed too much to him.

If it is understood that the parties will follow the protocol, but will try to covertly infer whatever they may from the messages, then this is called the "semi-honest" or "honest but curious" model. Using techniques from cryptography it is theoretically possible [16] to take a protocol for the semi-honest model and

make it work under a malicious model, in which one of the parties tries to deviate from the protocol in order to reveal information. Generally though, when the task is inference on joint data, it seems likely that both parties would benefit from the collaboration, and hence the semi-honest model may be a reasonable assumption.

In order to build a protocol for a particular computation, we first make an assumption about the computational power available to the parties. Then we choose a "security parameter" (similar in idea to a key length) so that for a particular party, to determine the others' private inputs becomes a computationally intractable problem (e.g., similar to breaking public key encryption) [16].

An important theoretical result in this area is given by Yao [42] and similarly [18], which show that any function of the parties private inputs may be computed in this setting. The idea is that the parties arrange their computation into a large circuit consisting of wires and gates, then apply a generic protocol to evaluate it on their inputs. Details are given in [16] although for the time being, such a generic protocol is primarily of theoretical interest, since it is prohibitively expensive for all but very small computations. Nevertheless see [28] for an implementation of the generic protocol. An area of study is the construction of protocols for specific problems, which often result in faster and more practically applicable methods. A cornerstone of such techniques is homomorphic encryption [29] which allows parties to perform mathematical operations on each others' encrypted values.

3.2 Alternative Security Models

An alternative which results in fast and often simpler protocols is the "weak" security model given in [9] and studied in [37] section 5.1.3. The idea of this model is that any protocol is fine, so long as the output doesn't reveal exactly what any parties particular input was. Specifically, if there exists an infinitely large set which could be substituted for a parties input, and result in the same output, then the protocol is secure in this weak model. The authors acknowledge that this definition is weak since this infinite set may be e.g., a small ball centered around some point in space, and so may still reveal a great deal of information [37]. Furthermore this definition has no mention of information leakage due to the protocol itself, however it could be amended so that the definition must hold for the intermediate messages as well as the final output. An analysis of some weakly secure inner product protocols is given by [15], who conclude that the weaker model presents a far greater prospect of information leakage than does the cryptographic model.

A second recent alternative is the so-called "differential privacy" approach due to Dwork and colleagues, e.g., see [10]. A randomized algorithm achieves differential privacy if its distribution of outputs doesn't change greatly when the input database is changed by one record. This technique was developed to prevent datamining schemes from releasing information which would identify individuals in the data. Nevertheless it may be brought to bear on multiparty computation. For example, for the problem of record linkage it is conceivable that each party

could use a randomized sanitization scheme on their data in order to achieve differential privacy. Then, the data could be revealed to the other parties, and then each party having his own copy of the complete sanitized data could run whatever record linkage or datamining algorithm he wanted to. The question which remains is whether differential privacy is a sufficiently strong guarantee compared to the cryptographic model, and whether this randomized sanitization scheme would corrupt the data so much that the results would be meaningless.

Finally in some settings the existence of a trusted third party may be realistic. Several protocols make explicit use of such a party [41,33,5,6,34], in a more limited way.

4 Privacy Preserving Record Linkage

When the files to be matched are held by two different parties and are deemed to be sensitive or private, then we may elicit the use of secure protocols in order to perform the record linkage and whatever may be the final statistical computation. This intersection of record linkage and PPDM has been of great interest in the last decade. The purpose of this section is to first highlight some of the unique challenges posed in this setting, and then to survey the results of research which has sought to solve them.

When the goal is for two parties to integrate their private data, typically they will only care about the set of matching records. If it was the case that they also wanted to share the non-links then there would be no need for secrecy since in the end all the data would become visible to both parties. Protocols which compute the set of linked records and then output them to both parties are perhaps the most well studied part of record linkage in the PPDM literature. In this case, the goal is to perform record linkage without revealing anything about the non-linked records (besides of course, whatever may be inferred of them by means of the linked records). In the cryptographic model this means e.g., that the values of the match variables as well as the parameters of p_θ should not become known explicitly to either party. Even if the computation of the match variables is done securely, for any party to know the values constitutes a failure of security since in general these values are not implied exactly by the linkage itself. For example, while it may be the case that linked records have high similarity, the exact values must remain unknown to either party.

It is important to pay attention to these details, consider a simple model where we allow both parties to learn the similarity measures. Say the data are real vectors and the computed similarity scores are the square or absolute errors between the components. In this case for example the party who holds A may consider two of his distinct values $a_h^k,\ a_i^k$ along with the computed similarities $m_{h,j}^k,\ m_{i,j}^k$. Now he has two distinct points on the real line as well as the distance of b_j^k to each point. Therefore he may solve to recover exactly the value of b_j^k, this way he may reveal he entirety of B, and likewise the owner of B may reveal A. This is a simple example but it serves to illustrate the problems that might arise from revealing intermediate values.

Another important distinction between the private and the usual non-private setting is that resorting to human clerical workers for disambiguation seems tantamount to an invasion of privacy. Although recent methods have focused on performing pure statistical linkage with no need for human intervention, there is a price to pay in the form of increased error rates. When the overarching goal is to perform some statistical analysis on the linked data, then the error in the linkage must be accounted for in order to obtain a valid analysis. This is in contrast to the usual setting where in essence the human-curated data may be treated as completely correct. Maintaining uncertainty about the linkage is an area which has begun to draw attention in the statistical literature, see e.g., [25].

When the goal is to perform some datamining task on the integrated data (e.g., [24,23,22,12,13]) then the data themselves are not part of the output. Instead, the final output of the protocol is e.g., a set of estimated regression coefficients on the integrated data, or some other such set of quantities. In this case, we need to take care to protect not only the non-links but also the linked data themselves. For instance, running a secure record linkage algorithm that outputs the links, and then using these data to fit a regression model does not constitute a secure protocol in the cryptographic model. The reason for this is that in general the data themselves are not implied by the regression output.

We repeat that, while in principle all the problems of privacy preserving datamining are solved by the generic protocol of Yao [42], the computational and communication demands of this method are too great in practice [37]. For this reason it is necessary in to devise protocols for the specific problem of record linkage, a problem that we now examine.

5 Methods in Privacy Preserving Record Linkage

While many authors in the literature propose end-to-end secure protocols for record linkage, oftentimes the individual steps may be seen as sub-protocols that are strung together into a secure protocol. Here we describe proposed methods for the steps identified in section 2. We begin, however, with a discussion of private exact matching, which is of historical importance.

5.1 Database Joins and Set Intersection

One of the earliest mentions of record linkage in a private setting is given in [1]. Here the author considers various classical problems from databases, ported to the private setting. The most relevant problem is the computing of a so called "equijoin." This may be considered a variant of record linkage in which two records link whenever they agree exactly on some specific subset of their fields. This then obviates both the need for parameter estimation and statistical inference of the joins, since a deterministic decision is made based upon the single match variable for each pair of records. The goal is to output the entire set of linked records, therefore it is not a concern if the match variables are revealed, since they are implied by the output.

A potential way to compute such an equijoin might be for both parties to apply some one-way hash function [17] to the fields of their records, then share them with each other and see which hashed values match. One might think that if the hash function is computationally hard to invert then this protocol would be safe. As shown in [1], this naive method fails since the hash function is deterministic. First it may be possible for either party to mount a dictionary attack in which they hash every possible value a field may take on and then see which ones match up to the other party's data. Secondly, when this attack is infeasible the parties may still consider the frequencies with which the hashed values appear. Using this along with knowledge of the distribution of field values (say, estimated empirically from their data), they may be able to reveal some values with high confidence. The way [1] resolve the issue is through the use of a semantically secure [17] encryption scheme. Using such a scheme guarantees that both of these proposed attacks will fails, since it implies that the encryptions are random, and the distributions of them do not differ significantly when the plaintext values are changed. The original protocol must then be modified to accommodate randomness in the hashing. Agrawal's idea paved the way for interest in private record linkage. From a theoretical perspective it is good starting point, however two questions remained. The first is whether the overhead of using this encryption scheme is too great. For example, in order for encryption to be sufficiently hard to break, usually the keys must be chosen to be thousands of bits long. This means that there is a great deal of communication cost, as well as computation since basic mathematical operations on such large numbers may be costly. The second question which remains is whether this approach may be extended to support non-exact matching such as is usual in record linkage.

5.2 Record Pair Similarity

The question of non-exact matching is partially addressed in [5,6,34]. These works in essence compute similarity scores for pairs of records via a reduction to a secure set intersection protocol. The idea applies mainly to text data such as names and addresses. First such fields are broken up into a set of "n-grams" which are the substrings of length n. Then since each field is now represented by a set, the size of the intersection of such sets may be compared with the size of the union, to get a measure of the degree of overlap between the two sets. If the intersection is large then the two strings have a large number of common substrings and so are regarded as close to each other and a potential candidate for matching. In principle, the secure protocol of [1] could be used for computing the intersections, however the authors are concerned about the computational overhead. Therefore they resort to a variant of the naive insecure approach mentioned in [1], in which a deterministic one way hashing function is used. To overcome the security issues the authors here instead suggest that a trusted third party may be employed to look at the hashed values and report the cardinality of the intersections. While in principle this approach would be very efficient, it is perhaps conceptually unappealing since the assumption of a trusted third party may be too restrictive in a wide variety of real problems.

An alternative method to compute string similarity is given by [31]. They present a secure two party protocol which computes approximate inner products between real vectors. Their idea is that strings which consist of multiple words may be represented in a vector space model by the well known TF-IDF transformation which was shown to be useful in record linkage [7]. Their approximation scheme makes use of a cryptographic protocol for secure set intersection, and therefore may be computationally demanding. Whats more, the approach is approximate and to increase the accuracy of the approximation requires increasing the size of the sets which get passed to the sub-protocol.

Another secure vector space method to compute edit distances is described by [33]. Their idea involves a so called metric embedding approach (see e.g., [4]). First some random set of strings is agreed upon by the two parties. Then each party computes the edit distance [2] of his records to each random string. With this in hand, the records may be described by a vector of real numbers in which each component is a distance to a random string. Then it may be shown that the euclidian distance between these vectors corresponds approximately to the string edit distance between the records. In principle, distances between strings could now be approximated by means of a secure inner product protocol, since if we use $\phi(\cdot)$ to denote the embedding we have:

$$d(a_i^k, b_j^k)^2 \approx ||\phi(a_i^k) - \phi(b_j^k)||_2^2 = ||\phi(a_i^k)||_2^2 + ||\phi(b_j^k)||_2^2 - 2\phi(a_i^k)^T\phi(b_j^k)$$

The last term is the inner product, and the other two terms may be computed locally by either party. The authors instead propose to use a third party protocol in which the embedded strings are sent to the third party for computation of these distances. It appears that despite the elegance of this approach, the third party would still be able to mount a frequency based attack on these embeddings. Nevertheless the metric embedding idea is compelling since it results in low-dimensional vectors [33], and so in principle it allows reduction of string edit distance computation to secure inner products which are already well-studied in the literature (e.g., [15]).

We note that all of the string similarity protocols make use of either set-intersection or inner products as a subprotocol. In essence any such protocol could be supplanted in place of the authors' suggestions, and the privacy guarantees and complexity of the resulting protocol would depend on those same characteristics of the sub-protocol. Therefore developing fast protocols for these two problems is important for the future of private record linkage. Although current protocols are reasonable in principle, remember that they will be run on every element of the direct product of the files, which could easily be millions of pairs for even modest size data.

Because a third party may decide whether or not certain similarity scores constitute a link, those protocols which use such a party evidently may output the linkage decision rather than just the similarity. For two party protocols it is less trivial to get the linkage classifications without revealing the similarity. One way, if the similarity scores are computed using a cryptographic protocol, would

be to threshold it before it is allowed to be decrypted. For example reducing similarity to the inner product and using [15] results in an encrypted value held by one party, where it may only be decrypted by the other. In this case the holding party may apply a certain sequence of operations to the ciphertext in order to reduce it to a binary flag corresponding to thresholding against some constant value. One such approach is via a reduction to the so called "millionaires problem" proposed by Yao, which in essence is a protocol to compute an inequality. See [3] for a recent approach.

5.3 Blocking

In the non-private setting, blocking [38] greatly reduces the number of record pairs to be classified. Several authors have ported this idea to the private setting. The idea of blocking is to use simple heuristics based on the record similarities to quickly remove obvious non-links from consideration. In the private setting, however, evaluating such heuristics may itself be a costly process.

One approach is given in [20]. In order to make the blocking step efficient the proposal is to first k-anonymize [35] the database rows, then share them. While the authors choose k-anonymity for its conceptual simplicity there is the prospect that other sanitization schemes could be used such as permuting with noise to achieve differential privacy [10]. After obtaining the sanitized version of the other party's data, the hope is that each party may infer a great deal of non-matches. However they won't be able to infer matches perfectly due to the corruption of the private data due to the sanitization. Therefore a second phase begins in which cryptographic protocols are used to resolve ambiguous record pairs. This way, the proposed scheme achieves a three-way tradeoff of computational overhead vs possible leakage of values vs accuracy of the solution. For example if the sanitization scheme leaves many values unchanged, then privacy is certainly breached, however the resulting accuracy of the linkage will be high, and the cost due to cryptographic protocols will be small. We note that since publication, there have been several published vulnerabilities in the k-anonymization framework [10].

Another paper which employs a blocking approach is [41]. Here the idea is to first transform the records into numeric vectors as in [33], and then perform a secure record linkage technique on these vectors. The protocol is structured in two rounds, the first of which is a blocking phase. The values are permuted and then shared so that the parties may quickly reject obvious non-matches. After this initial step, the remaining candidate record pairs are evaluated through a reduction to a secure inner product protocol as described above. The particular protocol they use may be considered as weakly secure [37].

Note that no matter the settings of the sanitization scheme, these methods will fail to meet the criteria of security in the cryptographic model. To achieve that standard, the sanitization scheme would have to render the data indistinguishable from any arbitrary dataset, and hence would render the blocking phase impossible. Therefore these approaches to blocking may only be used in a weaker security model. In principle it may be possible to do blocking in the

cryptographic model, by using a cryptographic protocol for the blocking heuristic; however, this may not be significantly faster than not performing blocking at all, e.g., if such a protocol is costly relative to the full matching protocol for a record pair. Nevertheless it is possible in practice that the guarantee afforded through the use of differential privacy [10] may be sufficient, so that a blocking scheme based on sanitized data may be feasible.

6 Prominent Unsolved Challenges

The main component of record linkage currently missing from the privacy-aware treatment is that of parameter estimation. All the works above made use of a-priori agreed upon thresholds for the similarity scores, and classify a record as a match if some a-priori agreed upon subset of fields are similar. This technique may result in good linkage under some conditions, however by sidestepping the difficult parameter estimation step, the result is a record linkage with no guarantees regarding error rates.

Another challenge which deserves attention is the development of techniques for record linkage which may propagate uncertainty through to subsequent statistical analysis. One approach is mentioned by Lahiri and Larsen [25] where the goal is to identify additional bias introduced by record linkage and remove it in the final calculation. More general techniques are required, but they may end up being different depending on the type of statistical analysis which is required. Such techniques will be very important, especially when the end result involves confidence intervals or hypothesis testing. The reason is that these are meant to come with well understood statistical guarantees (e.g., the probability of incorrectly rejecting a hypothesis is below some level α). When there is uncertainty in the data itself, then this uncertainty must be modeled in order to have such guarantees in the end.

In order for record linkage to be successfully applied to large databases, it will be important to increase the speed of the cryptographic underpinnings. While using clever protocols may reduce the number of operations (e.g., inner products) performed, ultimately the speed of these operations determines the feasibility of the secure approach.

Privacy-aware record linkage is a crucial problem lying at the intersection of statistics, computer science, and cryptography. We have provided an overview of the recent literature on the topic which builds on earlier reviews and the fundamental approach of Fellegi and Sunter pairs of data files. Extensions of of all of the methods described here to the case of linkage across multiple files, in the presence of measurement error remains a major statistical challenge.

Acknowledgement

This research was partially supported by Army contract DAAD19-02-1-3-0389 to Cylab at Carnegie Mellon University.

References

1. Agrawal, R., Evfimievski, A., Srikant, R.: Information sharing across private databases. In: SIGMOD 2003: Proceedings of the 2003 ACM SIGMOD international conference on Management of data, pp. 86–97. ACM, New York (2003)
2. Bilenko, M., Mooney, R.J., Cohen, W.W., Ravikumar, P., Fienberg, S.E.: Adaptive name matching in information integration. IEEE Intelligent Systems 18(5), 16–23 (2003)
3. Blake, I., Kolesnikov, V.: Strong conditional oblivious transfer and computing on intervals. In: Lee, P.J. (ed.) ASIACRYPT 2004. LNCS, vol. 3329, pp. 515–529. Springer, Heidelberg (2004)
4. Bourgain, J.: On lipschitz embedding of finite metric spaces in hilbert space. Israel Journal of Mathematics 52(1), 46–52 (1985)
5. Churches, T., Christen, P.: Blind data linkage using n-gram similarity comparisons. In: Dai, H., Srikant, R., Zhang, C. (eds.) PAKDD 2004. LNCS (LNAI), vol. 3056, pp. 121–126. Springer, Heidelberg (2004)
6. Churches, T., Christen, P.: Some methods for blindfolded record linkage. BMC Medical Informatics and Decision Making 4(1), 9 (2004)
7. Cohen, L.W., Cohen, W.W.: Data integration using similarity joins and a word-based information representation. ACM Transactions on Information Systems 18, 2000 (1998)
8. Domingo-Ferrer, J., Torra, V.: Validating distance-based record linkage with probabilistic record linkage. In: Escrig, M.T., Toledo, F.J., Golobardes, E. (eds.) CCIA 2002. LNCS (LNAI), vol. 2504, pp. 207–215. Springer, Heidelberg (2002)
9. Du, W., Chen, S., Han, Y.S.: Privacy-preserving multivariate statistical analysis: Linear regression and classification. In: Proceedings of the 4th SIAM International Conference on Data Mining, pp. 222–233 (2004)
10. Dwork, C.: Differential privacy: A survey of results. In: Agrawal, M., Du, D.-Z., Duan, Z., Li, A. (eds.) TAMC 2008. LNCS, vol. 4978, pp. 1–19. Springer, Heidelberg (2008)
11. Fellegi, I.P., Sunter, A.B.: A theory for record linkage. Journal of the American Statistical Association 64(328), 1183–1210 (1969)
12. Fienberg, S.E., Fulp, W.J., Slavkovic, A.B., Wrobel, T.A.: "Secure" log-linear and logistic regression analysis of distributed databases. In: Domingo-Ferrer, J., Franconi, L. (eds.) PSD 2006. LNCS, vol. 4302, pp. 277–290. Springer, Heidelberg (2006)
13. Fienberg, S., Slavkovic, A., Nardi, Y.: Valid statistical analysis for logistic regression with multiple sources. In: Gal, C.S., Kantor, P.B., Lesk, M.E. (eds.) ISIPS 2008. LNCS, vol. 5661, pp. 82–94. Springer, Heidelberg (2009)
14. Fienberg, S.E., Manrique-Vallier, D.: Integrated methodology for multiple systems estimation and record linkage using a missing data formulation. AStA Adv. Stat. Anal. 93, 49–60 (2009)
15. Goethals, B., Laur, S., Lipmaa, H., Mielikainen, T.: On secure scalar product computation for privacy-preserving data mining. In: ISISC 2004 (2004)
16. Goldreich, O.: Modern Cryptography, Probabilistic Proofs, and Pseudorandomness. Springer, New York (1998)
17. Goldreich, O.: Foundations of Cryptography. Basic Applications, vol. 2. Cambridge University Press, Cambridge (2004)
18. Goldreich, O., Micali, S., Wigderson, A.: How to play any mental game or a completeness theorem for protocols with honest majority. In: STOC, pp. 218–229. ACM, New York (1987)

19. Herzog, T.N., Scheuren, F.J., Winkler, W.E.: Data Quality and Record Linkage Techniques, 1st edn. Springer, Heidelberg (May 2007)
20. Inan, A., Kantarcioglu, M., Bertino, E., Scannapieco, M.: A hybrid approach to private record linkage. In: ICDE, pp. 496–505. IEEE, Los Alamitos (2008)
21. Jaro, M.: 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)
22. Karr, A., Lin, X., Reiter, J., Sanil, A.: Secure regression on distributed databases. Journal of Computational and Graphical Statistics 14(2), 263–279 (2005)
23. Karr, A., Lin, X., Reiter, J., Sanil, A.: Secure analysis of distributed databases. In: Olwell, D., Wilson, A.G., Wilson, G. (eds.) Statistical Methods in Counterterrorism: Game Theory, Modeling, Syndromic Surveillance, and Biometric Authentication, pp. 237–261. Springer, New York (2006)
24. Karr, A., Lin, X., Sanil, A., Reiter, J.: Privacy-preserving analysis of vertically partitioned data using secure matrix products. Journal of Official Statistics 25(1), 125–138 (2009)
25. Lahiri, P., Larsen, M.: Regression analysis with linked data. Journal of the American Statistical Association 100(469), 222–230 (2002)
26. Lindell, Y., Pinkas, B.: Privacy preserving data mining. Journal of Cryptology 15(3), 177–206 (2002)
27. Lindell, Y., Pinkas, B.: Secure multiparty computation for privacy-preserving data mining. Journal of Privacy and Confidentiality 1(1), 59–98 (2009)
28. Malkhi, D., Nisan, N., Pinkas, B., Sella, Y.: Fairplay–a secure two-party computation system. In: SSYM 2004: Proceedings of the 13th conference on USENIX Security Symposium, Berkeley, CA, USA, p. 20. USENIX Association (2004)
29. Paillier, P.: Public-key cryptosystems based on composite degree residuosity classes. In: Stern, J. (ed.) EUROCRYPT 1999. LNCS, vol. 1592, pp. 223–238. Springer, Heidelberg (1999)
30. Ravikumar, P., Cohen, W.W.: A hierarchical graphical model for record linkage. In: Chickering, D.M., Halpern, J.Y. (eds.) UAI, pp. 454–461. AUAI Press (2004)
31. Ravikumar, P., Cohen, W.W., Fienberg, S.E.: A secure protocol for computing string distance metrics. In: PSDM held at ICDM, pp. 40–46 (2004)
32. Sadinle, M.: Multiple record linkage: Generalizing the fellegi-sunter theory. Conict Analysis Resource Center (CERAC), Bogota, Columbia, January 22 (2010)
33. Scannapieco, M., Figotin, I., Bertino, E., Elmagarmid, A.K.: Privacy preserving schema and data matching. In: Chan, C.Y., Ooi, B.C., Zhou, A. (eds.) SIGMOD Conference, pp. 653–664. ACM, New York (2007)
34. Schnell, R., Bachteler, T., Reiher, J.: Privacy-preserving record linkage using bloom filters. BMC Medical Informatics and Decision Making 9(1), 41 (2009)
35. Sweeney, L.: k-anonymity: A model for protecting privacy. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems 10(5), 557–570 (2002)
36. Torra, V., Domingo-Ferrer, J.: Record linkage methods for multidatabase data mining. In: Torra, V. (ed.) Information Fusion in Data Mining, pp. 99–130. Springer, Heidelberg (2003)
37. Vaidya, J., Zhu, Y., Clifton, C.: Privacy Preserving Data Mining (Advances in Information Security). Springer, New York (2005)
38. Winkler, W.E.: Matching and record linkage. In: Business Survey Methods, pp. 355–384. Wiley, Chichester (1995)
39. Winkler, W.E.: The state of record linkage and current research problems. Technical report, Statistical Research Division, U.S. Bureau of the Census (1999)

40. Winkler, W.E.: Methods for record linkage and bayesian networks. Technical report, Series RRS2002/05, U.S. Bureau of the Census (2002)
41. Yakout, M., Atallah, M.J., Elmagarmid, A.K.: Efficient private record linkage. In: ICDE, pp. 1283–1286. IEEE, Los Alamitos (2009)
42. Yao, A.: Protocols for secure computations. In: Proceedings of the 23rd Annual IEEE Symposium on Foundations of Computer Science, pp. 160–164 (1982)

Strategies to Achieve SDC Harmonisation at European Level: Multiple Countries, Multiple Files, Multiple Surveys

Daniela Ichim and Luisa Franconi

Istituto Nazionale di Statistica, DCMT,
Via C. Balbo, 16, 00184 Roma, Italy
{ichim,franconi}@istat.it

Abstract. Preliminary considerations and an initial proposal are made for the harmonisation of different statistical disclosure limitation procedures at European level. Here we present the case of microdata file but the same approach could be successfully applied to other types of releases as well. The proposal is based on two pillars: in the methodological part, contrary to the proposal of Pérez-Duarte (2009), the harmonisation concept is defined by means of a set of minimal requirements on both the input and the output of the anonymisation process. In the organisational part, the burden is shared among actors in the European Statistical System. A proposal for a possible implementation of both the methodological and procedural/organisational framework is sketched. Issues related to the release of multiple files from the same survey i.e. from the same original dataset, are sketched. The release of multiple files is a new feature at European level stemming from the introduction of the public use file (PUF) concept in the new regulation on European statistics. This implies that for the same survey both a public use file and a microdata file for scientific purposes might be available: care must be taken in designing such files in order to avoid incoherence. Finally, the problem of the impact on the coherence of an anonymisation procedure of the release of a system of surveys is briefly explored.

Keywords: comparability, privacy in official statistics, SDC governance, public use file, microdata file for research.

1 Introduction

Under the umbrella of Regulation EC 831/2002 Eurostat releases European microdata for research purposes for several surveys ranging from social surveys such as the Labour force survey to business microdata (Community Innovation Survey – CIS – and Structure of Earning Survey – SES). Such microdata stem from a harmonised process usually ruled by European regulations which are mandatory for the Member States (MSs) sharing common definitions and common structure. However, in many cases, MSs do not agree with the anonymisation methodology proposed by Eurostat leading to the release of datasets that do not cover the whole of Europe. If an increase

J. Domingo-Ferrer and E. Magkos (Eds.): PSD 2010, LNCS 6344, pp. 284–296, 2010.

in the number of MSs participating to the European dissemination is aimed, a change in strategy is needed. In this paper preliminary considerations and an initial proposal are made in order to allow for more flexibility in the implementation of statistical disclosure limitation strategies and harmonise the anonymisation of microdata files at European level. Harmonisation is needed also to deal with the release of multiple files from the same survey. In fact, besides the release of European microdata for scientific purposes (MFR) the new European legislation on statistics allows for the dissemination of public use file. This introduces a further dimension to the multiple countries problem due to the release of multiple types of microdata for external users (PUF and MFR) from the same survey. Finally, in the last few years, a new way to systematically investigate the complexity of modern societies has lead to the development of systems of surveys that, although focussing in different areas, still present common structures and characteristics. SDC methods applied to such systems ought to be coherent in order to avoid limitations for the users.

In section 2 we describe the current limitation to the release of microdata in Europe and show how the same type of problems may occour in other international settings. In section 3 we address the multiple countries dimension of the European release of microdata by proposing a general framework that allows for flexibility within known boundaries. In Section 4 we sketch the problem of multiple releases from the same survey. Finally, the need to address the coherence of SDC methods when applied to multiple related surveys is briefly explored in section 5. The conclusions are presented in section 6.

2 European Anonymisation Process: Structural Constraints and Different Situations

The core of any dissemination procedure is the anonymisation process. The *input* of this process has two main parts: the original microdata file and the statistical disclosure limitation methodology that limits the disclosure risk and still provides utility to users. The *output* of the masking process is the microdata file to be released.

In this paper the input microdata files contain the original survey data collected by twenty seven MSs of the European Union. Usually, but this is certainly the case for the surveys mentioned in EC Commission Regulation 831/2002 that deals with the release of European microdata for scientific purposes, data collection and processing are harmonised at European level. What makes the European anonymisation procedure different from an anonymisation procedure in a single MS is the complexity derived from several different approaches and situations. The anonymisation of European microdata files ought to take into account both organisational heterogeneity of MSs and their needs, rights and duties to respect their own national standards.

The organisational heterogeneity of MSs is visible in several dimensions. Without being exhaustive, some of these dimensions are listed below. It should be observed that the dissemination of European microdata files should deal with all these features.

a) *Law*: Legislation in MSs obliges the data owner, i.e. institution that collects the data, to guarantee the confidentiality of respondents. The responsible institution is the data owner although the possible harm is propagated throughout the whole European Statistical System (ESS).

b) *Organisation of the Statistical System*: According to each national statistical system organisation, the data might be collected by a National Statistical Institute or by some other type of entity, for example a minister or a research institute. This is an important issue as national statistical laws may oblige only some types of organisations to preserve the confidentiality of respondents and not others. Moreover, the data collection via administrative registers is another type of organisation of a statistical system. From now on, for simplicity, we will refer to the data owner as the institution who carries out the survey.
c) *Access to original confidential microdata*: Some MSs allow access to the original microdata, some others may not allow such access, or, at least, (international) access to the original microdata might be extremely difficult.
d) *Microdata transmission*: Some MSs have the legal possibility of transmitting the original microdata to other institutions, under bilateral agreements. In some countries the transmission of microdata (even to Eurostat) is possible only if this is accounted for in a specific regulation that obliges the MS to do so.
e) *Microdata dissemination*: Some MSs have the legal possibility of disseminating anonymised microdata files; some others may not have such possibility. Also it is possible that a MS may easily allow the dissemination of some kind of microdata (e.g. social), while strictly prohibiting the dissemination of other data types (e.g. enterprises, or indeed the other way around).

Each MS should decide on its participation to a given dissemination channel. For example, the MSs may agree or disagree on an anonymisation procedure and they may or may not allow the release of a certain data set to a particular project under EC Regulation 831/2002. The release of an MFR is not compulsory. Nonetheless, the NSIs have the mission to provide society the information needed. That's why, if the national legislation allows it, MS are generally willing to disseminate anonymised microdata files, in the provision that the anonymisation process is up to their national standards. However, there are large countries and very small countries with completely different economic structures, with different perceptions of the disclosure risk and different approaches to confidentiality. Also, as disclosure limitation is a recent field of work for many data owner, significant differences are visible amongst MSs. An anonymisation process needs to take into account all such different standards.

2.1 Analysis of the Current Anonymisation Flow and Its Critical Points

Figure 1 summarily presents the current flow of the anonymisation of a European MFR. Usually Eurostat proposes a masking procedure to be adopted; subject to MS agreement, carries out the microdata anonymisation, produces the CD-ROM and takes care of its subsequent delivery. Of course, the same strategy could be easily extended to other forms of microdata dissemination, if agreement is got on this workflow.

The central role of Eurostat in the current anonymisation flow may be easily observed. Starting from the methodological proposal and ending by the CD-ROM dissemination, Eurostat is the most important actor and therefore the one that holds most of the work and responsibilities. From the harmonisation side, this is surely a very attractive feature of the European dissemination procedure. If on the one hand the implementation of a single method is an important characteristics for users and is

a crucial simplifying step for Eurostat (who might find difficult to apply different methodologies for different MSs); on the other hand, being a rigid solution it might limit the possibility of anonymisation for a large number of MSs. We will see how such situation could be modified by creating formal forms of cooperation that allow to share the burden inside the ESS in order to develop more sustainable anonymisation procedures in a predefined methodological framework.

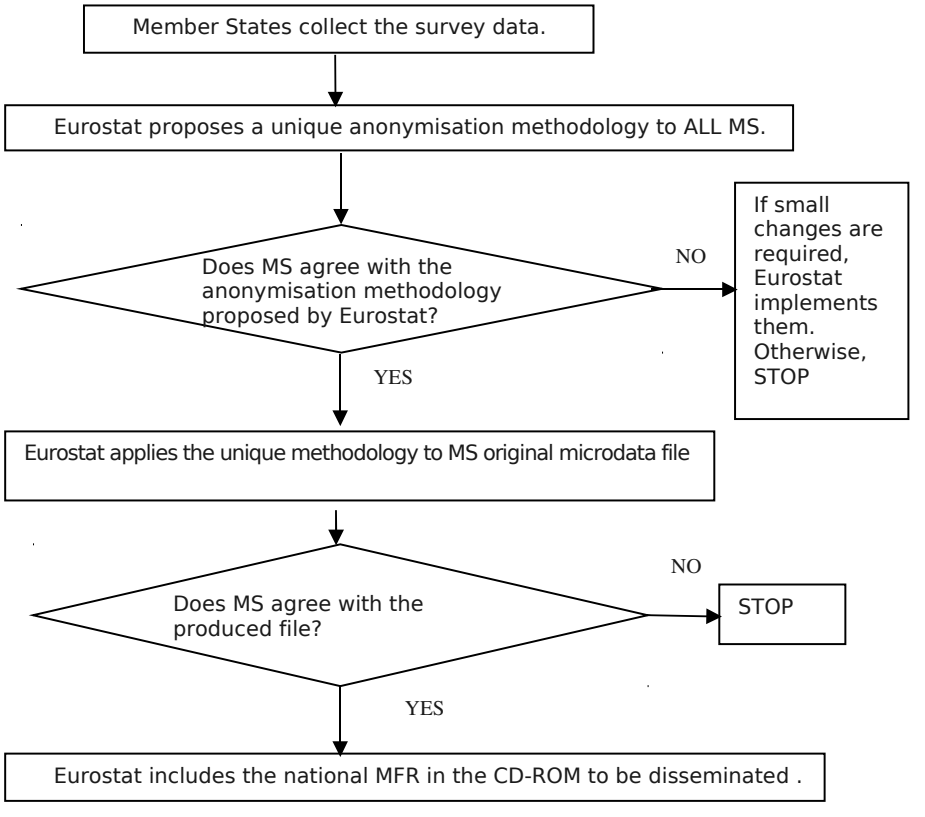


Fig. 1. Current flow of anonymisation of European microdata files for research purposes

Given a certain level of data utility, a European microdata file needs to satisfy a crucial requirement in order to meet the expectations of users: it has to be representative of all MSs in Europe. It might be possible that a single protection method may not be suitable for all MSs (and all real data sets and all waves of the survey). Indeed, the use of a single method does not take into consideration MSs organisational heterogeneity (discussed in section 2). If a MS wants to disseminate its own anonymised microdata file at European level, there is strong pressure to agree with the anonymisation methodology proposed by Eurostat.

In order to increase the number of MSs adhering to the release of a microdata file, some forms of flexibility need to be introduced in the anonymisation process to accommodate for organisational heterogeneity of MSs and their own standards. Alongside flexibility the other pillar on which to build harmonisation is the involvement of other MSs inside the ESS to share the burden of anonymisation.

2.2 Other Possible International Settings

In this paper we analyse only some problems related to the European system of information dissemination. However, it should be noted that the same approaches, analyses and solutions could be applied to other information dissemination systems. International organisations like UNECE or OECD make efforts to disseminate information at transnational level. To cite only a few, well-known examples, a) OECD is currently running a project to disseminate microdata files stemming from a variety of labour force surveys, see Brackfield and Ruiz (2010), b) the World Bank is coordinating the International Household Survey Network, c) Integrated Public Use Microdata Series (census microdata for social and economic research) and d) Demographic and Health Surveys maintained by the U.S. Agency for International Development (USAID). Sometimes, due to legal, logistic and cultural differences, it is not realistic to constrain national organizations to strictly follow some approved guidelines; an alternative could then be the proposed approach.

3 Proposal for a Harmonised European Anonymisation

A harmonised anonymisation of microdata files would be surely profitable for all actors in the release process: users, data owners, mainly National Statistical Institutes, and Eurostat. A harmonised anonymisation would increase the number of MSs releasing their microdata and therefore increase data utility. Moreover, the corresponding European data set would still share the same harmonisation properties of the original data files as the building phase would be harmonised as well. At the same time, data owners should be aware that a harmonised anonymisation would greatly benefit them, too. First, the recognition of structural differences and internal standards would allow more MSs to adhere to the anonymisation. Second, the exchange of experiences and competences surely generates improved results. Finally, Eurostat with the help of MSs with sound experience in the area of SDC would enforce its co-ordinating role at European level by promoting the definition and adoption of a set of common guidelines and by sustaining the implementation of software routines able to be applied to different microdata files.

At a first glance, one might believe that a harmonisation of the disseminated microdata files is very difficult. Since the organisational heterogeneity of MSs is a rigid constraint, we believe that a harmonised European anonymisation of microdata files could be achieved twofold: 1) modelling the input of the anonymisation methodology and 2) modelling the output of the anonymisation methodology. In other words, the harmonisation concept is defined by means of a set of minimal requirements on both input and output of the anonymisation process. The dissemination flow, as described in section 1, presents an *input* phase and an *output* phase. In principle, on the

input phase, a significant improvement might be reached by using flexible statistical methods. On the output phase, the definition of a battery of benchmarking statistics and corresponding quality criteria/thresholds could be used to put in practice the comparability concept. The changes to the European anonymisation flow are survey independent. Nonetheless, benchmarking statistics and quality criteria should be survey specific and should be applied to each survey wave. If appropriate, the same benchmarking statistics and quality criteria/thresholds could be applied to consecutive waves.

3.1 Working on the Input of the Process: Statistical Methodology

a) A single method

Currently the European anonymisation procedure foresees the application of a single statistical disclosure limitation methodology. This strategy surely has the lowest costs in terms of implementation, testing and application. It might be believed that this strategy also produces highly harmonised results. Nonetheless, the application of the same statistical disclosure limitation method to different data sets might produce different qualitative and quantitative results.

Given the organisational heterogeneity of the MSs, it is hard to believe that there exists a method that best suits the requirements and standards of twenty seven countries. The best practical option would be the choice from a list of candidate methods. Anyway, it should be observed that the choice of the statistical disclosure limitation methodology is not an easy task. Today many statistical disclosure control methods exist, each one with its own merits and drawbacks. To our knowledge, there is no final winner. The situation is much more complicated when both risk and data utility are considered as the scientific community didn't find a rigorous way to compare all the protection methods.

The choice (selection/definition) would not completely solve the acceptance problem of the MSs. Because the participation to this dissemination channel is not mandatory, even if a method is agreed, one MS could still refuse its application. This could mainly concern the MSs that today cannot legally disseminate anonymised microdata files. If, in future, their national law would change, those MSs could still not agree with *a priori* selected anonymisation methods.

b) More methods

A simple strategy that possibly could take into account the MSs organisational heterogeneity is the creation of a list of pre-defined candidate methodologies. This approach would surely require some more resources spent in implementation and testing.

An advantage could be the possibility to increase the number of MSs agreeing to disseminate anonymised microdata files at European level. For example, in the framework of enterprise microdata European dissemination, the MSs that could have accepted the individual ranking applied irrespective of the categorical structural key variables (i.e. irrespective of the stratification) have already agreed on. If an increase of the number of the MSs participating to the European dissemination is aimed, a change in strategy is needed.

The usage of a list of candidate statistical disclosure limitation methodologies could activate a sort of virtual competition among methods. Different strategies could be implemented and tested on real survey data. In medium-long term, empirical evidence would guide the selection of the most suitable strategy for the analysed survey.

c) Flexible methods – parameterisation

This proposal is just an extension of the previous one (point b), aiming at increasing the number of MSs disseminating anonymised microdata files at European level. Different variants of the same statistical disclosure limitation methodology could be easily implemented and tested. For example, the implementation of the individual ranking could depend on the microaggregation parameter p; then, each MS could select its own value for this parameter p, e.g. 3 or 5 or some other value. The implementation of a statistical disclosure limitation methodology with respect to different stratification domains is another form of flexibility. For example, the methodology could be applied to the entire microdata file or to the domains defined by the categorical key variables (generally the structural categorical variables). In other words, by simply changing the values of some parameters, the statistical disclosure methodology could be more easily adapted to many MSs.

Another option could be the usage of sound statistical methods allowing, by definition, the output control. That is, some output quality indicators could already be taken into account by the statistical disclosure limitation methodology. For example, in the framework of continuous variables, if the preservation of weighted totals is required, using a methodology that by definition satisfies this constraint (e.g. adding noise or regression models) could be very helpful. Moreover, the usage of such statistical methods would allow a sound study of the statistical properties of the anonymised microdata files.

3.2 Working on the Output of the Process: Comparable Dissemination

Data utility / data quality are one of the most important characteristics of the output of the European anonymisation flow. Timeliness, consistency, efficacy and comparability are only some dimensions of data quality which are of interest to the users. Data utility is neither easy to define nor easy to quantify. We propose to assess it through the definition of benchmarking statistics for the type of data under analysis. Then, thresholds / quality criteria on these benchmarking statistics should be set. Moreover, possible remedies should be indicated for the cases when the quality criteria are not met. For the definition of both benchmarking statistics and their corresponding thresholds / quality criteria, cooperation between survey experts and methodologists is strategic. The most relevant statistics (benchmarking statistics) could be identified from a review of previous analyses performed on the survey data and from information given by users groups.
The comparable dissemination procedure may be summarised by the following steps:

a. Given a single survey (CIS, SES, etc.);
b. Indicate a list of non-statistical quality indicators $Q_1, Q_2, \ldots, Q_n$;

c. Indicate a list of benchmarking statistics $S_1, S_2, \ldots, S_m$;
d. Indicate the thresholds / quality criteria $C_1, C_2, \ldots, C_M$, $M \geq m$ associated to the statistical indicators $S_1, S_2, \ldots, S_m$;
e. Suppose that a candidate statistical disclosure limitation methodology is applied to the original microdata file;
f. If the anonymised microdata file satisfies each of the non-statistical criteria $Q_1, Q_2, \ldots, Q_n$ and each of the quality criteria $C_1, C_2, \ldots, C_M$ corresponding to the statistical indicators $S_1, S_2, \ldots, S_m$, then the file should be accepted for dissemination at European level.

Using the above procedure, at least from the point of view of the considered statistical (and non statistical) indicators, the comparability among the MSs would be guaranteed.

Examples of non-statistical indicators are: fulfilment of a dissemination deadline, compatibility with a predefined electronic format, preservation of the original microdata file structure; examples of statistical indicators could be: preservation of an informative content of the most important variables, preservation of an informative content of the survey specific variables (generally the confidential variables), means of the most important variables, by stratification domain, variances of the most important variables, by stratification domain, distributions of the most important variables, by stratification domain, already published statistics (tables). Finally examples of quality criteria/thresholds could be: preservation of a minimum level of detail on categorical variables (for example NACE 2-digits or NUTS at regional level), bounds on variations (e.g. the anonymised total should not differ from the original total by more than given percentage), and coherence with the already published statistics. Some remarks on the process are outlined:

1. The procedure should be constructed and applied to each survey. This dependency on survey is due to the fact that the benchmarking statistics and their quality criteria/thresholds are strongly related to the survey type, to the kind of microdata and to the kind of analyses performed on such microdata.
2. The procedure should be constructed and applied to each survey wave (see item 3, too). The same motivations as above.
3. In order to ensure the comparability among distinct waves of the same survey, the same statistics and quality indicators should be chosen.
4. For each statistics S indicated in step c, different quality criteria/thresholds may be indicated, consequently, $M \geq m$. For example, one might bound the total variation, but at the same time, the total computed on anonymised data should be nonnegative.
5. The key point in the comparable dissemination procedure is the definition of the benchmarking statistics and their thresholds / quality criteria. Anyway, the importance of non-statistical criteria should also be stressed.

An example of such an approach can be found in Franconi and Ichim (2009). A proposal of governance structure is presented in Appendix 1.

4 Release of Multiple Types of Files

The release of different files from the same microdata is a new issue at European level. It derives from the entry into force of the new regulation on European statistics, Reg. (CE) 223/2009, introducing the definition of public use file (PUF) besides the already implemented file for scientific purposes (MFR). Although new at transnational level, the instances of the production of multiple files from the same dataset are however growing very fast as international institutes or EU or world based projects urge the need to develop "customised files" that could be compared at international level: recent examples are the "Generation and gender project" (http://www.unece.org/pau/ggp/Welcome.html) or the IPUMS project (https://international.ipums.org/international/). The problem encountered in such situation is a simple one: the file required by international institutions is generally not a problematic one *in itself*, but it might differ for some classifications from other files already released at national or EU level. For example, nowadays an international organisation could require for a certain survey a level of geography not extremely detailed but, at the same time, it would need of indications on the socio-demographic characteristics of the municipality. Such requirements could then be in contrast with previously released files with more detailed geography where information on the size of the town or its rural/urban nature were not present. This type of problem is the microdata counterpart of the linked tables problem and, as for the latter, an optimal solution can be found only when the different data to be released are anonymised at the same time. Therefore to be optimal at European level, the anonymisation of different types of microdata files should be planned at the same time.

At national level the multiple types of files (multiple releases) problem has already been encountered as the production of different files for different users is becoming a widespread practice (see for example Trottini *et al.* (2006) for a dissemination strategy proposal for the household expenditure survey in Italy). In Appendix 2, an overview of the problem of multiple releases is given.

However, despite of the need of data anonymisation procedures targeted to the different data users, the problem of releasing different files is still at an embryonic stage and indeed very rarely approached in practice (besides the previous citation an example of such implementations can be found in Abowd and Lane, 2003). This is due to the cost associated with a real differentiated data dissemination strategy and the complexity of its implementation. What is most commonly applied in most MSs adopting a dual dissemination (PUF and MFR) is the mere adoption of more aggregated classifications for the categorical variables and various forms of top and bottom coding as well as the introduction of bands for the continuous variables. This causes the needed drastic decrease of the risk of disclosure but presents, as a side effect, a severe drop in the information content of the microdata file. Also, till the present time, the dual release process at national level has been, in most cases, a *controlled release* also when "general use files" were involved. This means that in most countries the current procedure to release a microdata file implies the need for a formal request (therefore implying the clear identification of the user), specifications of the foreseen uses to be provided and some sort of confidentiality statement agreed. However, a new concept of PUF need to be developed where the dissemination mean will be the web and where a simple download could be the procedure to gain it. Possibly, in the future, for European PUF there will be no list of users, no control on reasons for access nor on

uses. This implies a completely new approach to the definition of a PUF with respect to the ones MSs are currently used to. The risk of disclosure will be surely higher as the risk is related also to the dissemination mean. However, new methods and a new attitude towards statistical disclosure control could supply strategies where the public nature, i.e. the free availability of the PUF, should not be the synonym of the production of files showing very limited interest and analytical validity for the final users. Targeted utility-based perturbation methods or, more recently, synthetic data generation methods can be used to release perturbed data that still present interesting level of information content.

Certainly PUF and MFR must be hierarchically designed in terms of information content (see Trottini et al. 2006). This means that all the information in the PUF should also be contained in the corresponding MFR. The hierarchical structure of the two data sets greatly simplifies assessment of the disclosure risk and information loss associated with the anonymisation procedure. Because of the hierarchy, in fact, there is no gain for a user having access to the MFR, to access the PUF. The hierarchy requires coherence in the choice of the variables to be included in both files and on the corresponding level of details. The inclusion of a variable in the PUF implies its inclusion in the MFR; non nested classifications for the same variable should not be allowed, and so on. The use of strategies outlined in section 3 (comparable dissemination) would allow the selection of the list of variables to be included and the agreement on the basic and broad classification for the PUF. Then, details on single respondents could be provided inside the broad band by means of perturbation or synthetic generation. So if a ten year classes for the variable age is agreed for the PUF, the age in single year of a the respondent could be generated inside the corresponding broad ten year band using also different methodologies. We foresee that the use of comparable dissemination coupled with perturbation/data generation procedures could allow both the definition of interesting PUF for the users and coherent multiple releases.

5 Release of Multiple Related Surveys

In many MSs the definition of a system of surveys structured in such a way that a basic questionnaire is present yearly but different modules are rotated year after year is becoming common; this is done in order to monitor cyclically a phenomenon of interest (in Italy the multipurpose system of surveys, at European level the future general social survey). Without reaching this level of definition, social surveys present always the same socio-demographic characteristics: gender, age, marital status, etc. It would be extremely appealing if a systematic recognition of such variables would be identified and harmonisation of the SDC practices applied in related surveys would be achieved.

6 Conclusions

The harmonisation of surveys and processes throughout Europe is recognised as a key feature for the future of European statistics. In this report we identify the dimensions in which the anonymisation process at European level should develop, highlight some of the corresponding critical points and cast possible ways to approach a solution.

The underlying idea is to develop a framework for harmonising the anonymisation process with the active cooperation of MSs by proposing possible sound alternative

methodologies and by setting benchmarking statistics and thresholds on such statistics in order to guarantee the users with a minimum standard of quality throughout the continent. The framework and such indicators could be simply part of the structure of the quality report that each survey under European regulation needs to comply with. The flexibility allowed by the process will increase the number of MSs adhering to the dissemination and therefore the number of data sets available to users and will foster the development of knowledge in the field of the statistical disclosure limitation methods within the ESS.

The comparable dissemination framework implies an initial investment in identifying the benchmarking statistics and relative thresholds / quality criteria but, then, the whole procedure is expected to become part of the production process. Also this initial stage can be performed with the help of MSs who have gained already experience in this field by creating an institutional form of collaboration on this particular area of expertise. It would be extremely beneficial if Eurostat would formally join together the experienced and willing MSs to develop and test the anonymisation process or even to take part to the production of the anonymised files. This systematic collaboration between the partners of the European Statistical System would allow sharing the burden that is currently on the shoulder of Eurostat and transferring the knowledge and expertise across the ESS, as suggested in Eurostat (2009).

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

References

Abowd, J.M., Lane, J.: Synthetic data and confidentiality protection. In: Workshop on Microdata, Stockholm, Sweden (August 2003)

Brackfield, D., Ruiz, N.: Harmonised Labour Force and Migration Statistics Based on Microdata. In: Joint UNECE/Eurostat work session on statistical data confidentiality (2009), http://www.unece.org/stats/documents/2009.12.confidentiality.htm

Eurostat, Communication on the production method of EU statistics: "a vision for the next decade" (2009)

Franconi, L., Ichim, D.: Community Innovation Survey: Comparable Dissemination. In: Joint UNECE/Eurostat work session on statistical data confidentiality 2007 (2009), ISBN 978-92-79-12055-8, Theme: General and regional statistics, Collection: Methodologies and working papers, http://www.unece.org/stats/documents/2007/12/confidentiality/wp.2.e.pdf, doi:10.2901/Eurostat.C2007.004

Ichim, D.: Community Innovation Survey: a Flexible Approach to the Dissemination of Microdata Files for Research. In: Proceedings of Q2008, European Conference on Quality in Official Statistics (2008), http://q2008.istat.it/sessions/24.html

Pérez-Duarte, S.: Harmonisation of anonymisation practices through partially synthetic files. In: Joint UNECE/Eurostat work session on statistical data confidentiality, Bilbao (December 2009), http://www.unece.org/stats/documents/2009.12.confidentiality.htm

Trottini, M., Franconi, L., Polettini, S.: Italian Household Expenditure Survey: A Proposal for Data Dissemination. In: Domingo-Ferrer, J., Franconi, L. (eds.) PSD 2006. LNCS, vol. 4302, pp. 318–333. Springer, Heidelberg (2006)

Appendix 1: Harmonised Anonymisation Flow

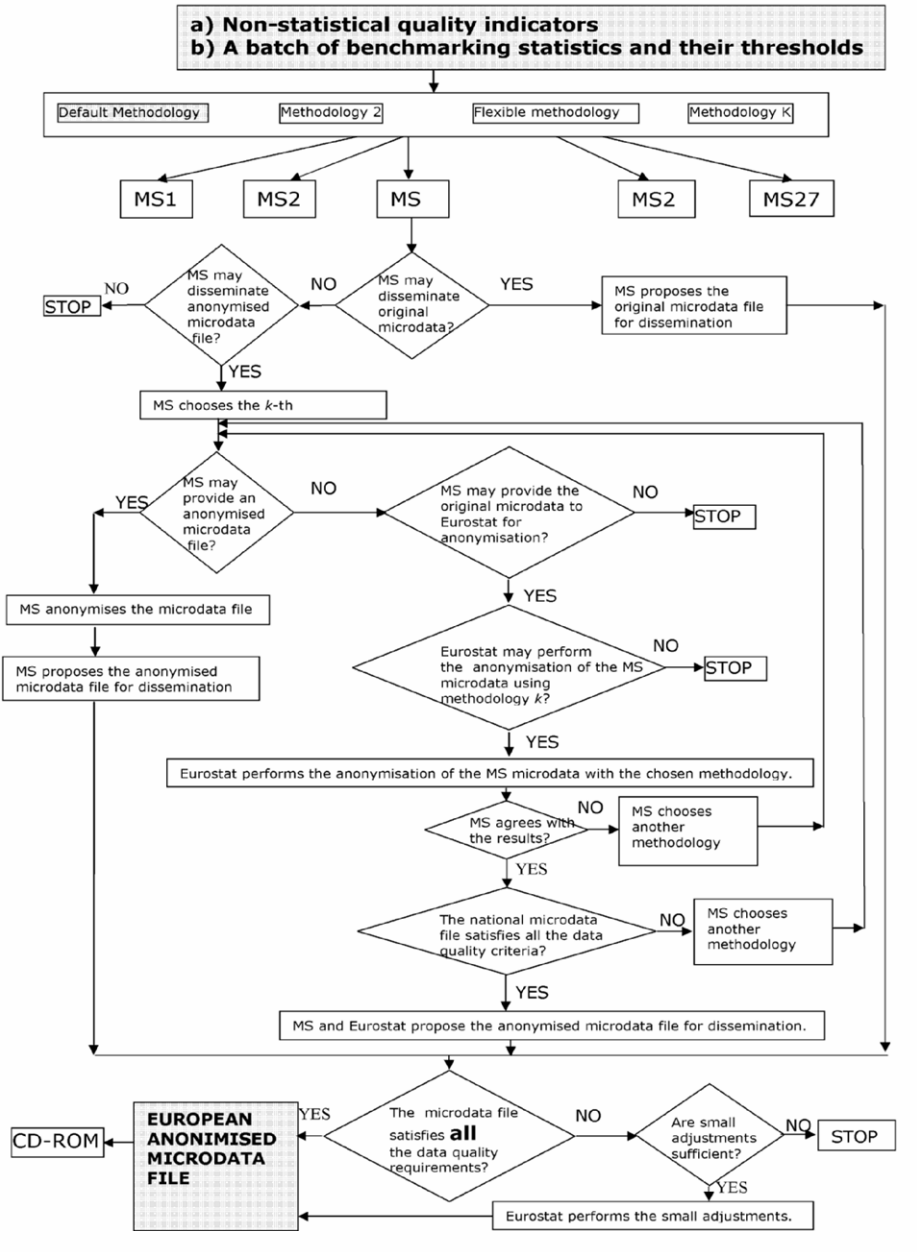


Fig. 2. Example of harmonised anonymisation flow

Appendix 2: Dimensions of the Harmonisation Problem

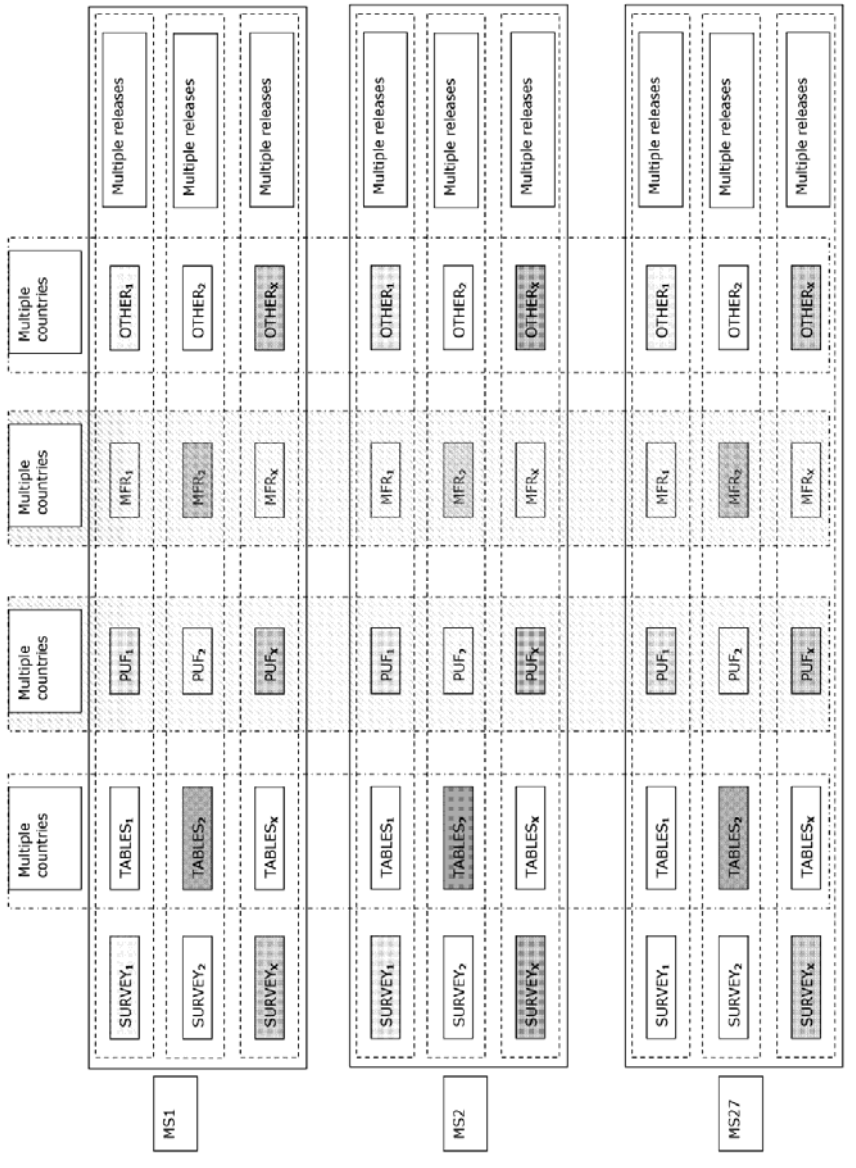


Fig. 3. Summary of the dimensions of the harmonisation problem: vertical boxes represent transnational releases, horizontal dashed (uniform colour) boxes represent release of multiple files and horizontal large boxes represent release of related surveys

Author Index