Josep Domingo-Ferrer
Luisa Franconi (Eds.)

LNCS 4302

Privacy in Statistical Databases

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

Springer

Josep Domingo-Ferrer   Luisa Franconi (Eds.)

Privacy in Statistical Databases

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

Springer

Volume Editors

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

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

Library of Congress Control Number: 2006936080

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

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

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

Springer is a part of Springer Science+Business Media

springer.com

Printed in Germany

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

Lecture Notes in Computer Science 4302

Lecture Notes in Computer Science

For information about Vols. 1–4211

please contact your bookseller or Springer

Vol. 4313: T. Margaria, B. Steffen (Eds.), Leveraging Applications of Formal Methods. IX, 197 pages. 2006.

Vol. 4312: S. Sugimoto, J. Hunter, A. Rauber, A. Morishima (Eds.), Digital Libraries: Achievements, Challenges and Opportunities. XVIII, 571 pages. 2006.

Vol. 4302: J. Domingo-Ferrer, L. Franconi (Eds.), Privacy in Statistical Databases. XI, 383 pages. 2006.

Vol. 4300: Y.Q. Shi (Ed.), Transactions on Data Hiding and Multimedia Security I. IX, 139 pages. 2006.

Vol. 4293: A. Gelbukh, C.A. Reyes-Garcia (Eds.), MICAI 2006: Advances in Artificial Intelligence. XXVIII, 1232 pages. 2006. (Sublibrary LNAI).

Vol. 4292: G. Bebis, R. Boyle, B. Parvin, D. Koracin, P. Remagnino, A. Nefian, G. Meenakshisundaram, V. Pascucci, J. Zara, J. Molineros, H. Theisel, T. Malzbender (Eds.), Advances in Visual Computing, Part II. XXXII, 906 pages. 2006.

Vol. 4291: G. Bebis, R. Boyle, B. Parvin, D. Koracin, P. Remagnino, A. Nefian, G. Meenakshisundaram, V. Pascucci, J. Zara, J. Molineros, H. Theisel, T. Malzbender (Eds.), Advances in Visual Computing, Part I. XXXI, 916 pages. 2006.

Vol. 4290: M. van Steen, M. Henning (Eds.), Middleware 2006. XIII, 425 pages. 2006.

Vol. 4283: Y.Q. Shi, B. Jeon (Eds.), Digital Watermarking. XII, 474 pages. 2006.

Vol. 4281: K. Barkaoui, A. Cavalcanti, A. Cerone (Eds.), Theoretical Aspects of Computing - ICTAC 2006. XV, 371 pages. 2006.

Vol. 4280: A.K. Datta, M. Gradinariu (Eds.), Stabilization, Safety, and Security of Distributed Systems. XVII, 590 pages. 2006.

Vol. 4279: N. Kobayashi (Ed.), Programming Languages and Systems. XI, 423 pages. 2006.

Vol. 4278: R. Meersman, Z. Tari, P. Herrero (Eds.), On the Move to Meaningful Internet Systems 2006: OTM 2006 Workshops, Part II. XLV, 1004 pages. 2006.

Vol. 4277: R. Meersman, Z. Tari, P. Herrero (Eds.), On the Move to Meaningful Internet Systems 2006: OTM 2006 Workshops, Part I. XLV, 1009 pages. 2006.

Vol. 4276: R. Meersman, Z. Tari (Eds.), On the Move to Meaningful Internet Systems 2006: CoopIS, DOA, GADA, and ODBASE, Part II. XXXII, 752 pages. 2006.

Vol. 4275: R. Meersman, Z. Tari (Eds.), On the Move to Meaningful Internet Systems 2006: CoopIS, DOA, GADA, and ODBASE, Part I. XXXI, 1115 pages. 2006.

Vol. 4273: I.F. Cruz, S. Decker, D. Allemang, C. Preist, D. Schwabe, P. Mika, M. Uschold, L. Aroyo (Eds.), The Semantic Web - ISWC 2006. XXIV, 1001 pages. 2006.

Vol. 4272: P. Havinga, M. Lijding, N. Meratnia, M. Wegdam (Eds.), Smart Sensing and Context. XI, 267 pages. 2006.

Vol. 4271: F.V. Fomin (Ed.), Graph-Theoretic Concepts in Computer Science. XIII, 358 pages. 2006.

Vol. 4270: H. Zha, Z. Pan, H. Thwaites, A.C. Addison, M. Forte (Eds.), Interactive Technologies and Sociotechnical Systems. XVI, 547 pages. 2006.

Vol. 4269: R. State, S. van der Meer, D. O'Sullivan, T. Pfeifer (Eds.), Large Scale Management of Distributed Systems. XIII, 282 pages. 2006.

Vol. 4268: G. Parr, D. Malone, M. Ó Foghlú (Eds.), Autonomic Principles of IP Operations and Management. XIII, 237 pages. 2006.

Vol. 4267: A. Helmy, B. Jennings, L. Murphy, T. Pfeifer (Eds.), Autonomic Management of Mobile Multimedia Services. XIII, 257 pages. 2006.

Vol. 4266: H. Yoshiura, K. Sakurai, K. Rannenberg, Y. Murayama, S. Kawamura (Eds.), Advances in Information and Computer Security. XIII, 438 pages. 2006.

Vol. 4265: L. Todorovski, N. Lavrač, K.P. Jantke (Eds.), Discovery Science. XIV, 384 pages. 2006. (Sublibrary LNAI).

Vol. 4264: J.L. Balcázar, P.M. Long, F. Stephan (Eds.), Algorithmic Learning Theory. XIII, 393 pages. 2006. (Sublibrary LNAI).

Vol. 4263: A. Levi, E. Savas, H. Yenigün, S. Balcisoy, Y. Saygin (Eds.), Computer and Information Sciences – ISCIS 2006. XXIII, 1084 pages. 2006.

Vol. 4261: Y. Zhuang, S. Yang, Y. Rui, Q. He (Eds.), Advances in Multimedia Information Processing - PCM 2006. XXII, 1040 pages. 2006.

Vol. 4260: Z. Liu, J. He (Eds.), Formal Methods and Software Engineering. XII, 778 pages. 2006.

Vol. 4259: S. Greco, Y. Hata, S. Hirano, M. Inuiguchi, S. Miyamoto, H.S. Nguyen, R. Słowiński (Eds.), Rough Sets and Current Trends in Computing. XXII, 951 pages. 2006. (Sublibrary LNAI).

Vol. 4257: I. Richardson, P. Runeson, R. Messnarz (Eds.), Software Process Improvement. XI, 219 pages. 2006.

Vol. 4256: L. Feng, G. Wang, C. Zeng, R. Huang (Eds.), Web Information Systems – WISE 2006 Workshops. XIV, 320 pages. 2006.

Vol. 4255: K. Aberer, Z. Peng, E.A. Rundensteiner, Y. Zhang, X. Li (Eds.), Web Information Systems – WISE 2006. XIV, 563 pages. 2006.

Vol. 4254: T. Grust, H. Höpfner, A. Illarramendi, S. Jablonski, M. Mesiti, S. Müller, P.-L. Patranjan, K.-U. Sattler, M. Spiliopoulou (Eds.), Current Trends in Database Technology – EDBT 2006. XXXI, 932 pages. 2006.

Vol. 4253: B. Gabrys, R.J. Howlett, L.C. Jain (Eds.), Knowledge-Based Intelligent Information and Engineering Systems, Part III. XXXII, 1301 pages. 2006. (Sublibrary LNAI).

Vol. 4252: B. Gabrys, R.J. Howlett, L.C. Jain (Eds.), Knowledge-Based Intelligent Information and Engineering Systems, Part II. XXXIII, 1335 pages. 2006. (Sublibrary LNAI).

Vol. 4251: B. Gabrys, R.J. Howlett, L.C. Jain (Eds.), Knowledge-Based Intelligent Information and Engineering Systems, Part I. LXVI, 1297 pages. 2006. (Sublibrary LNAI).

Vol. 4249: L. Goubin, M. Matsui (Eds.), Cryptographic Hardware and Embedded Systems - CHES 2006. XII, 462 pages. 2006.

Vol. 4248: S. Staab, V. Svátek (Eds.), Managing Knowledge in a World of Networks. XIV, 400 pages. 2006. (Sublibrary LNAI).

Vol. 4247: T.-D. Wang, X. Li, S.-H. Chen, X. Wang, H. Abbass, H. Iba, G. Chen, X. Yao (Eds.), Simulated Evolution and Learning. XXI, 940 pages. 2006.

Vol. 4246: M. Hermann, A. Voronkov (Eds.), Logic for Programming, Artificial Intelligence, and Reasoning. XIII, 588 pages. 2006. (Sublibrary LNAI).

Vol. 4245: A. Kuba, L.G. Nyúl, K. Palágyi (Eds.), Discrete Geometry for Computer Imagery. XIII, 688 pages. 2006.

Vol. 4244: S. Spaccapietra (Ed.), Journal on Data Semantics VII. XI, 267 pages. 2006.

Vol. 4243: T. Yakhno, E.J. Neuhold (Eds.), Advances in Information Systems. XIII, 420 pages. 2006.

Vol. 4242: A. Rashid, M. Aksit (Eds.), Transactions on Aspect-Oriented Software Development II. IX, 289 pages. 2006.

Vol. 4241: R.R. Beichel, M. Sonka (Eds.), Computer Vision Approaches to Medical Image Analysis. XI, 262 pages. 2006.

Vol. 4239: H.Y. Youn, M. Kim, H. Morikawa (Eds.), Ubiquitous Computing Systems. XVI, 548 pages. 2006.

Vol. 4238: Y.-T. Kim, M. Takano (Eds.), Management of Convergence Networks and Services. XVIII, 605 pages. 2006.

Vol. 4237: H. Leitold, E. Markatos (Eds.), Communications and Multimedia Security. XII, 253 pages. 2006.

Vol. 4236: L. Breveglieri, I. Koren, D. Naccache, J.-P. Seifert (Eds.), Fault Diagnosis and Tolerance in Cryptography. XIII, 253 pages. 2006.

Vol. 4234: I. King, J. Wang, L. Chan, D. Wang (Eds.), Neural Information Processing, Part III. XXII, 1227 pages. 2006.

Vol. 4233: I. King, J. Wang, L. Chan, D. Wang (Eds.), Neural Information Processing, Part II. XXII, 1203 pages. 2006.

Vol. 4232: I. King, J. Wang, L. Chan, D. Wang (Eds.), Neural Information Processing, Part I. XLVI, 1153 pages. 2006.

Vol. 4231: J. F. Roddick, R. Benjamins, S. Si-Saïd Cherfi, R. Chiang, C. Claramunt, R. Elmasri, F. Grandi, H. Han, M. Hepp, M. Hepp, M. Lytras, V.B. Mišić, G. Poels, I.-Y. Song, J. Trujillo, C. Vangenot (Eds.), Advances in Conceptual Modeling - Theory and Practice. XXII, 456 pages. 2006.

Vol. 4230: C. Priami, A. Ingólfsdóttir, B. Mishra, H.R. Nielson (Eds.), Transactions on Computational Systems Biology VII. VII, 185 pages. 2006. (Sublibrary LNBI).

Vol. 4229: E. Najm, J.F. Pradat-Peyre, V.V. Donzeau-Gouge (Eds.), Formal Techniques for Networked and Distributed Systems - FORTE 2006. X, 486 pages. 2006.

Vol. 4228: D.E. Lightfoot, C.A. Szyperski (Eds.), Modular Programming Languages. X, 415 pages. 2006.

Vol. 4227: W. Nejdl, K. Tochtermann (Eds.), Innovative Approaches for Learning and Knowledge Sharing. XVII, 721 pages. 2006.

Vol. 4226: R.T. Mittermeir (Ed.), Informatics Education – The Bridge between Using and Understanding Computers. XVII, 319 pages. 2006.

Vol. 4225: J.F. Martínez-Trinidad, J.A. Carrasco Ochoa, J. Kittler (Eds.), Progress in Pattern Recognition, Image Analysis and Applications. XIX, 995 pages. 2006.

Vol. 4224: E. Corchado, H. Yin, V. Botti, C. Fyfe (Eds.), Intelligent Data Engineering and Automated Learning – IDEAL 2006. XXVII, 1447 pages. 2006.

Vol. 4223: L. Wang, L. Jiao, G. Shi, X. Li, J. Liu (Eds.), Fuzzy Systems and Knowledge Discovery. XXVIII, 1335 pages. 2006. (Sublibrary LNAI).

Vol. 4222: L. Jiao, L. Wang, X. Gao, J. Liu, F. Wu (Eds.), Advances in Natural Computation, Part II. XLII, 998 pages. 2006.

Vol. 4221: L. Jiao, L. Wang, X. Gao, J. Liu, F. Wu (Eds.), Advances in Natural Computation, Part I. XLI, 992 pages. 2006.

Vol. 4220: C. Priami, G. Plotkin (Eds.), Transactions on Computational Systems Biology VI. VII, 247 pages. 2006. (Sublibrary LNBI).

Vol. 4219: D. Zamboni, C. Kruegel (Eds.), Recent Advances in Intrusion Detection. XII, 331 pages. 2006.

Vol. 4218: S. Graf, W. Zhang (Eds.), Automated Technology for Verification and Analysis. XIV, 540 pages. 2006.

Vol. 4217: P. Cuenca, L. Orozco-Barbosa (Eds.), Personal Wireless Communications. XV, 532 pages. 2006.

Vol. 4216: M.R. Berthold, R. Glen, I. Fischer (Eds.), Computational Life Sciences II. XIII, 269 pages. 2006. (Sublibrary LNBI).

Vol. 4215: D.W. Embley, A. Olivé, S. Ram (Eds.), Conceptual Modeling - ER 2006. XVI, 590 pages. 2006.

Vol. 4213: J. Fürnkranz, T. Scheffer, M. Spiliopoulou (Eds.), Knowledge Discovery in Databases: PKDD 2006. XXII, 660 pages. 2006. (Sublibrary LNAI).

Vol. 4212: J. Fürnkranz, T. Scheffer, M. Spiliopoulou (Eds.), Machine Learning: ECML 2006. XXIII, 851 pages. 2006. (Sublibrary LNAI).

Preface

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

There are at least two traditions in statistical database privacy: one stems from official statistics, where the discipline is also known as statistical disclosure control (SDC), and the other originates from computer science and database technology. Both started in the 1970s, but the 1980s and the early 1990s saw little privacy activity on the computer science side. The Internet era has strengthened the interest of both statisticians and computer scientists in this area. Along with the traditional topics of tabular and microdata protection, some research lines have revived and/or appeared, such as privacy in queryable databases and protocols for private data computation.

Privacy in Statistical Databases 2006 (PSD 2006) was the main conference of the CENEX-SDC project (Center of Excellence in SDC), funded by EUROSTAT (European Commission) and held in Rome, December 13–15, 2006. PSD 2006 is a successor of PSD 2004, the final conference of the CASC project (IST-2000-25069), held in Barcelona in 2004 and with proceedings published by Springer as LNCS vol. 3050. Those two PSD conferences follow a tradition of high-quality technical conferences on SDC which started with "Statistical Data Protection–SDP 1998", held in Lisbon in 1998 and with proceedings published by OPOCE, and continued with the AMRADS project SDC Workshop, held in Luxemburg in 2001 and with proceedings published in Springer LNCS vol. 2316.

The Program Committee accepted 31 papers out of 45 submissions from 17 different countries in four different continents. Each submitted paper received at least two reviews. These proceedings contain the revised versions of the accepted papers, which are a fine blend of contributions from official statistics and computer science. Covered topics include methods for tabular data protection, methods for individual data (microdata) protection, assessments of analytical utility and disclosure risk, protocols for private computation, case studies and SDC software.

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

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

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

September 2006

Josep Domingo-Ferrer
Luisa Franconi

Privacy in Statistical Databases - PSD 2006

Program Committee

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

Program Chair

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

General Chair

Luisa Franconi (ISTAT, Italy)

Organization Committee

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

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

Table of Contents

Methods for Tabular Protection

Utility and Risk in Tabular Protection

Methods for Microdata Protection

Utility and Risk in Microdata Protection

Protocols for Private Computation

Case Studies

Software

A Method for Preserving Statistical Distributions Subject to Controlled Tabular Adjustment

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

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

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

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

1 Introduction

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

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

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

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

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

2 The SDL Problem for Tabular Data

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

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

3 MDI-CTA

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

3.1 Definition of MDI

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

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

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

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

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

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

Some properties of the MDI are:

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

3.2 Applying MDI to CTA

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

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

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