

Commenced Publication in 1973

Founding and Former Series Editors:

Gerhard Goos, Juris Hartmanis, and Jan van Leeuwen

Editorial Board

Takeo Kanade

Carnegie Mellon University, Pittsburgh, PA, USA

Josef Kittler

University of Surrey, Guildford, UK

Jon M. Kleinberg

Cornell University, Ithaca, NY, USA

Friedemann Mattern

ETH Zurich, Switzerland

John C. Mitchell

Stanford University, CA, USA

Moni Naor

Weizmann Institute of Science, Rehovot, Israel

Oscar Nierstrasz

University of Bern, Switzerland

C. Pandu Rangan

Indian Institute of Technology, Madras, India

Bernhard Steffen

University of Dortmund, Germany

Madhu Sudan

Massachusetts Institute of Technology, MA, USA

Demetri Terzopoulos

New York University, NY, USA

Doug Tygar

University of California, Berkeley, CA, USA

Moshe Y. Vardi

Rice University, Houston, TX, USA

Gerhard Weikum

Max-Planck Institute of Computer Science, Saarbruecken, Germany

Springer

Berlin

Heidelberg

New York

Hong Kong

London

Milan

Paris

Tokyo

Josep Domingo-Ferrer Vicenç Torra (Eds.)

Privacy in Statistical Databases

CASC Project Final Conference, PSD 2004
Barcelona, Catalonia, Spain, June 9-11, 2004
Proceedings



Springer

Volume Editors

Josep Domingo-Ferrer

Universitat Rovira i Virgili

Department of Computer Engineering and Mathematics

Av. Països Catalans 26, 43007 Tarragona, Catalonia, Spain

E-mail: jdomingo@etse.urv.es

Vicenç Torra

Institut d'Investigació en Intel·ligència Artificial

Campus de Bellaterra, 08193 Bellaterra, Catalonia, Spain

E-mail: vtorra@iiia.csic.es

Library of Congress Control Number: 2004106913

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

ISSN 0302-9743

ISBN 3-540-22118-2 Springer-Verlag Berlin Heidelberg New York

This work is subject to copyright. All rights are reserved, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, re-use of illustrations, recitation, broadcasting, reproduction on microfilms or in any other way, and storage in data banks. Duplication of this publication or parts thereof is permitted only under the provisions of the German Copyright Law of September 9, 1965, in its current version, and permission for use must always be obtained from Springer-Verlag. Violations are liable to prosecution under the German Copyright Law.

Springer-Verlag is a part of Springer Science+Business Media

springeronline.com

© Springer-Verlag Berlin Heidelberg 2004

Printed in Germany

Typesetting: Camera-ready by author, data conversion by Olgun Computergrafik

Printed on acid-free paper SPIN: 11008736 06/3142 5 4 3 2 1 0

Preface

Privacy in statistical databases is about finding tradeoffs to the tension between the increasing societal and economical demand for accurate information and the legal and ethical obligation to protect the privacy of individuals and enterprises, which are the source of the statistical data. Statistical agencies cannot expect to collect accurate information from individual or corporate respondents unless these feel the privacy of their responses is guaranteed; also, recent surveys of Web users show that a majority of these are unwilling to provide data to a Web site unless they know that privacy protection measures are in place.

“Privacy in Statistical Databases 2004” (PSD 2004) was the final conference of the CASC project (“Computational Aspects of Statistical Confidentiality”, IST-2000-25069). PSD 2004 is in the style of the following conferences: “Statistical Data Protection”, held in Lisbon in 1998 and with proceedings published by the Office of Official Publications of the EC, and also the AMRADS project SDC Workshop, held in Luxemburg in 2001 and with proceedings published by Springer-Verlag, as LNCS Vol. 2316.

The Program Committee accepted 29 papers out of 44 submissions from 15 different countries on four continents. Each submitted paper received at least two reviews. These proceedings contain the revised versions of the accepted papers. These papers cover the foundations and methods of tabular data protection, masking methods for the protection of individual data (microdata), synthetic data generation, disclosure risk analysis, and software/case studies.

Many people deserve our gratitude. The conference and these proceedings would not have existed without the Organization Chair, Enric Ripoll, and the Organizing Committee (Jordi Castellà, Antoni Martínez, Francesc Sebé and Julià Urrutia). In evaluating the papers submitted we received the help of the Program Committee and four external reviewers (Jörg Höhne, Silvia Polettini, Yosef Rinott and Giovanni Seri).

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

March 2004

Josep Domingo-Ferrer
Vicenç Torra

Privacy in Statistical Databases – PSD 2004

Program Committee

Rakesh Agrawal (IBM Almaden, USA)
David Brown (ONS, UK)
Jordi Castro (Polytechnical University of Catalonia)
Lawrence Cox (Nat. Center for Health Statistics, USA)
Ramesh Dandekar (EIA, USA)
Stephen Fienberg (Carnegie Mellon University, USA)
Luisa Franconi (ISTAT, Italy)
Sarah Giessing (Destatis, Germany)
Anco Hundepool (Statistics Netherlands, The Netherlands)
Julia Lane (Urban Institute, USA)
Josep M. Mateo-Sanz (Universitat Rovira i Virgili, Catalonia)
William E. Winkler (Census Bureau, USA)
Laura Zayatz (Census Bureau, USA)

Program Chairs

Josep Domingo-Ferrer (Universitat Rovira i Virgili, Catalonia)
Vicenç Torra (IIIA-CSIC, Catalonia)

Organization Committee

Jordi Castellà Roca (Universitat Rovira i Virgili, Catalonia)
Antoni Martínez-Ballesté (Universitat Rovira i Virgili, Catalonia)
Francesc Sebé (Universitat Rovira i Virgili, Catalonia)
Julià Urrutia (IDESCAT, Catalonia)

Organization Chair

Enric Ripoll (IDESCAT, Catalonia)

Table of Contents

Foundations of Tabular Protection

Survey on Methods for Tabular Data Protection in ARGUS	1
<i>Sarah Giessing</i>	

Data Swapping: Variations on a Theme by Dalenius and Reiss	14
<i>Stephen E. Fienberg and Julie McIntyre</i>	

Bounds for Cell Entries in Two-Way Tables Given Conditional Relative Frequencies	30
<i>Aleksandra B. Slavkovic and Stephen E. Fienberg</i>	

Methods for Tabular Protection

A New Tool for Applying Controlled Rounding to a Statistical Table in Microsoft Excel	44
<i>Juan-José Salazar-González and Markus Schoch</i>	

Getting the Best Results in Controlled Rounding with the Least Effort . . .	58
<i>Juan-José Salazar-González, Philip Lowthian, Caroline Young, Giovanni Merola, Stephen Bond, and David Brown</i>	

Computational Experiments with Minimum-Distance Controlled Perturbation Methods	73
<i>Jordi Castro</i>	

Balancing Quality and Confidentiality for Multivariate Tabular Data	87
<i>Lawrence H. Cox, James P. Kelly, and Rahul Patil</i>	

Reducing the Set of Tables τ -ARGUS Considers in a Hierarchical Setting .	99
<i>Peter-Paul de Wolf and Anneke Loeve</i>	

Approaches to Identify the Amount of Publishable Information in Business Surveys through Waivers	110
<i>Jean-Sébastien Provençal, Hélène Bérard, Jean-Marc Fillion, and Jean-Louis Tambay</i>	

Maximum Utility-Minimum Information Loss Table Server Design for Statistical Disclosure Control of Tabular Data	121
<i>Ramesh A. Dandekar</i>	

A Fast Network Flows Heuristic for Cell Suppression in Positive Tables . .	136
<i>Jordi Castro</i>	

Masking for Microdata Protection

On the Security of Noise Addition for Privacy in Statistical Databases . . . 149
Josep Domingo-Ferrer, Francesc Sebé, and Jordi Castellà-Roca

Microaggregation for Categorical Variables: A Median Based Approach . . . 162
Vicenç Torra

Evaluating Fuzzy Clustering Algorithms for Microdata Protection 175
Vicenç Torra and Sadaaki Miyamoto

To Blank or Not to Blank? A Comparison of the Effects
of Disclosure Limitation Methods on Nonlinear Regression Estimates 187
Sandra Lechner and Winfried Pohlmeier

Outlier Protection in Continuous Microdata Masking 201
Josep Maria Mateo-Sanz, Francesc Sebé, and Josep Domingo-Ferrer

Risk in Microdata Protection

Re-identification Methods for Masked Microdata 216
William E. Winkler

Masking and Re-identification Methods for Public-Use Microdata:
Overview and Research Problems 231
William E. Winkler

A Bayesian Hierarchical Model Approach to Risk Estimation
in Statistical Disclosure Limitation 247
Silvia Polettini and Julian Stander

Individual Risk Estimation in μ -Argus: A Review 262
Luisa Franconi and Silvia Polettini

Analysis of Re-identification Risk Based on Log-Linear Models 273
Elsayed A.H. Elamir

Synthetic Data

New Approaches to Confidentiality Protection:
Synthetic Data, Remote Access and Research Data Centers 282
John M. Abowd and Julia Lane

Multiply-Imputing Confidential Characteristics and File Links
in Longitudinal Linked Data 290
John M. Abowd and Simon D. Woodcock

Fast Generation of Accurate Synthetic Microdata 298
*Josep Maria Mateo-Sanz, Antoni Martínez-Ballesté,
and Josep Domingo-Ferrer*

Software and Case Studies

Trade-Off between Disclosure Risk and Information Loss

Using Multivariate Microaggregation: A Case Study on Business Data 307

Josep A. Sànchez, Julià Urrutia, and Enric Ripoll

The ARGUS Software in the CASC-Project 323

Anco Hundepool

Different Grades of Statistical Disclosure Control Correlated

with German Statistics Law 336

Thomas Wende

Developing Adoptable Disclosure Protection Techniques:

Lessons Learned from a U.S. Experience 343

Nicholas H. Greenia

Privacy Preserving and Data Mining in an On-Line Statistical Database

of Additive Type 353

Francesco M. Malvestuto and Mauro Mezzini

Author Index 367