

## Founding Editors

Gerhard Goos

*Karlsruhe Institute of Technology, Karlsruhe, Germany*

Juris Hartmanis

*Cornell University, Ithaca, NY, USA*

## Editorial Board Members

Elisa Bertino

*Purdue University, West Lafayette, IN, USA*

Wen Gao

*Peking University, Beijing, China*

Bernhard Steffen 

*TU Dortmund University, Dortmund, Germany*

Gerhard Woeginger 

*RWTH Aachen, Aachen, Germany*

Moti Yung

*Columbia University, New York, NY, USA*

More information about this series at <http://www.springer.com/series/7409>

Josep Domingo-Ferrer ·  
Krishnamurty Muralidhar (Eds.)

# Privacy in Statistical Databases

UNESCO Chair in Data Privacy, International Conference, PSD 2020  
Tarragona, Spain, September 23–25, 2020  
Proceedings

*Editors*

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

Krishnamurty Muralidhar  
University of Oklahoma  
Norman, OK, USA

ISSN 0302-9743

ISSN 1611-3349 (electronic)

Lecture Notes in Computer Science

ISBN 978-3-030-57520-5

ISBN 978-3-030-57521-2 (eBook)

<https://doi.org/10.1007/978-3-030-57521-2>

LNCS Sublibrary: SL3 – Information Systems and Applications, incl. Internet/Web, and HCI

© Springer Nature Switzerland AG 2020

Chapter “Explaining Recurrent Machine Learning Models: Integral Privacy Revisited” is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>). For further details see licence information in the chapter.

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Switzerland AG  
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

# Preface

Privacy in statistical databases is a discipline whose purpose is to provide solutions to the tension between the social, political, economic, and corporate demands of accurate information, and the legal and ethical obligation to protect the privacy of the various parties involved. In particular, the need to enforce the EU General Data Protection Regulation (GDPR) in our world of big data has made this tension all the more pressing. Stakeholders include the subjects, sometimes a.k.a. respondents (the individuals and enterprises to which the data refer), the data controllers (those organizations collecting, curating, and to some extent sharing or releasing the data) 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 controllers to invest in subject privacy: if individual subjects feel their privacy is guaranteed, they are likely to provide more accurate responses. Data controller privacy is primarily motivated by practical considerations: if an enterprise collects data at its own expense and responsibility, 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 data controller 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) or statistical disclosure limitation (SDL), and the second one originates from computer science and database technology. In official statistics, the basic concern is subject privacy. In computer science, the initial motivation was also subject privacy but, from 2000 onwards, growing attention has been devoted to controller 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. At the same time, the generalization of big data is challenging privacy technologies in many ways: this volume also contains recent research aimed at tackling some of these challenges.

Privacy in Statistical Databases 2020 (PSD 2020) was held in Tarragona, Catalonia, Spain, under the sponsorship of the UNESCO Chair in Data Privacy, which has provided a stable umbrella for the PSD biennial conference series since 2008. Previous PSD conferences were held in various locations around the Mediterranean, and had their proceedings published by Springer in the LNCS series: PSD 2018, Valencia, LNCS 11126; PSD 2016, Dubrovnik, LNCS 9867; PSD 2014, Eivissa, LNCS 8744; PSD 2012, Palermo, LNCS 7556; PSD 2010, Corfu, LNCS 6344; PSD 2008, Istanbul, LNCS 5262; PSD 2006 (the final conference of the Eurostat-funded CENEX-SDC project), Rome, LNCS 4302; and PSD 2004 (the final conference of the European FP5 CASC project) Barcelona, LNCS 3050. The nine PSD conferences held so far are a follow-up of a series of high-quality technical conferences on SDC which started

22 years ago with Statistical Data Protection-SDP 1998, held in Lisbon in 1998 with proceedings published by OPOCE, and continued with the AMRADS project SDC workshop, held in Luxemburg in 2001 with proceedings published by Springer in LNCS 2316.

The PSD 2020 Program Committee accepted for publication in this volume 25 papers out of 49 submissions. Furthermore, 10 of the above submissions were reviewed for short oral presentation at the conference. Papers came from 14 different countries and 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. Covered topics include privacy models, microdata protection, protection of statistical tables, protection of interactive and mobility databases, record linkage and alternative methods, synthetic data, data quality, and case studies.

We are indebted to many people. Firstly, 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 by Weiyi Xia, Zhiyu Wan, Chao Yan, and Jeremy Seeman as external reviewers. We also wish to thank all the authors of submitted papers and we apologize for possible omissions.

July 2020

Josep Domingo-Ferrer  
Krishnamurty Muralidhar

# Organization

## Program Committee

Jane Bambauer	University of Arizona, USA
Bettina Berendt	Technical University of Berlin, Germany
Elisa Bertino	CERIAS, Purdue University, USA
Aleksandra Bujnowska	EUROSTAT, EU
Jordi Castro	Polytechnical University of Catalonia, Spain
Anne-Sophie Charest	Université Laval, Canada
Chris Clifton	Purdue University, USA
Graham Cormode	University of Warwick, USA
Josep Domingo-Ferrer	Universitat Rovira i Virgili, Catalonia, Spain
Jörg Drechsler	IAB, Germany
Khaled El Emam	University of Ottawa, Canada
Mark Elliot	The University of Manchester, UK
Sébastien Gambs	Université du Québec à Montréal, Canada
Sarah Giessing	Destatis, Germany
Sara Hajian	Nets Group, Denmark
Hiroaki Kikuchi	Meiji University, Japan
Bradley Malin	Vanderbilt University, USA
Laura McKenna	Census Bureau, USA
Anna Monreale	Università di Pisa, Italy
Krishnamurty Muralidhar	University of Oklahoma, USA
Anna Oganyan	National Center for Health Statistics, USA
David Rebollo-Monedero	Universitat Rovira i Virgili, Catalonia, Spain
Jerome Reiter	Duke University, USA
Yosef Rinott	Hebrew University, Israel
Steven Ruggles	University of Minnesota, USA
Nicolas Ruiz	OECD, Universitat Rovira i Virgili, Catalonia, Spain
Pierangela Samarati	Università di Milano, Italy
David Sánchez	Universitat Rovira i Virgili, Catalonia, Spain
Eric Schulte-Nordholt	Statistics Netherlands, The Netherlands
Natalie Shlomo	The University of Manchester, UK
Aleksandra Slavković	Penn State University, USA
Jordi Soria-Comas	Catalan Data Protection Authority, Catalonia, Spain
Tamir Tassa	The Open University, Israel
Vicenç Torra	Umeå University, Sweden
Lars Vilhuber	Cornell University, USA
Peter-Paul de Wolf	Statistics Netherlands, The Netherlands

## **Program Chair**

Josep Domingo-Ferrer	UNESCO Chair in Data Privacy, Universitat Rovira i Virgili, Catalonia, Spain
----------------------	---

## **General Chair**

Krishnamurty Muralidhar	University of Oklahoma, USA
-------------------------	-----------------------------

## **Organization Committee**

Joaquín García-Alfaro	Télécom SudParis, France
Jesús Manjón	Universitat Rovira i Virgili, Catalonia, Spain
Romina Russo	Universitat Rovira i Virgili, Catalonia, Spain

# Contents

## Privacy Models

<i>P<sub><math>\alpha,\beta</math></sub></i> -Privacy: A Composable Formulation of Privacy Guarantees for Data Publishing Based on Permutation . . . . .	3
<i>Nicolas Ruiz</i>	
$\epsilon$ -Differential Privacy for Microdata Releases Does Not Guarantee Confidentiality (Let Alone Utility). . . . .	21
<i>Krishnamurty Muralidhar, Josep Domingo-Ferrer, and Sergio Martínez</i>	
A Bayesian Nonparametric Approach to Differentially Private Data. . . . .	32
<i>Fadhel Ayed, Marco Battiston, and Giuseppe Di Benedetto</i>	
A Partitioned Recoding Scheme for Privacy Preserving Data Publishing . . . .	49
<i>Chris Clifton, Eric J. Hanson, Keith Merrill, Shawn Merrill, and Amjad Zahraa</i>	
Explaining Recurrent Machine Learning Models: Integral Privacy Revisited . . . . .	62
<i>Vicenç Torra, Guillermo Navarro-Arribas, and Edgar Galván</i>	
Utility-Enhancing Flexible Mechanisms for Differential Privacy . . . . .	74
<i>Vaikkunth Mugunthan, Wanyi Xiao, and Lalana Kagal</i>	
Plausible Deniability . . . . .	91
<i>David Sidi and Jane Bambauer</i>	

## Microdata Protection

Analysis of Differentially-Private Microdata Using SIMEX . . . . .	109
<i>Anne-Sophie Charest and Leila Nombo</i>	
An Analysis of Different Notions of Effectiveness in k-Anonymity . . . . .	121
<i>Tanja Šarčević, David Molnar, and Rudolf Mayer</i>	
Multivariate Top-Coding for Statistical Disclosure Limitation . . . . .	136
<i>Anna Oganian, Ionut Iacob, and Goran Lesaja</i>	

## Protection of Statistical Tables

Calculation of Risk Probabilities for the Cell Key Method . . . . .	151
<i>Tobias Enderle, Sarah Giessing, and Reinhard Tent</i>	

On Different Formulations of a Continuous CTA Model . . . . . 166  
*Goran Lesaja, Ionut Iacob, and Anna Oganian*

**Protection of Interactive and Mobility Databases**

Privacy Analysis of Query-Set-Size Control . . . . . 183  
*Eyal Nussbaum and Michael Segal*

Statistical Disclosure Control When Publishing on Thematic Maps . . . . . 195  
*Douwe Hut, Jasper Goseling, Marie-Colette van Lieshout,  
Peter-Paul de Wolf, and Edwin de Jonge*

**Record Linkage and Alternative Methods**

Bayesian Modeling for Simultaneous Regression and Record Linkage . . . . . 209  
*Jiurui Tang, Jerome P. Reiter, and Rebecca C. Steorts*

Probabilistic Blocking and Distributed Bayesian Entity Resolution . . . . . 224  
*Ted Enamorado and Rebecca C. Steorts*

Secure Matrix Computation: A Viable Alternative to Record Linkage? . . . . . 240  
*Jörg Drechsler and Benjamin Klein*

**Synthetic Data**

A Synthetic Supplemental Public Use File of Low-Income Information  
Return Data: Methodology, Utility, and Privacy Implications . . . . . 257  
*Claire McKay Bowen, Victoria Bryant, Leonard Burman,  
Surachai Khitatrakun, Robert McClelland, Philip Stallworth,  
Kyle Ueyama, and Aaron R. Williams*

Integrating Differential Privacy in the Statistical Disclosure Control  
Tool-Kit for Synthetic Data Production . . . . . 271  
*Natalie Shlomo*

Advantages of Imputation vs. Data Swapping for Statistical  
Disclosure Control . . . . . 281  
*Satkartar K. Kinney, Charlotte B. Looby, and Feng Yu*

**Data Quality**

Evaluating Quality of Statistical Disclosure Control  
Methods – VIOLAS Framework . . . . . 299  
*Olga Dzięgielewska*

Detecting Bad Answers in Survey Data Through Unsupervised Machine Learning . . . . .	309
<i>Najeeb Moharram Jebreel, Rami Haffar, Ashneet Khandpur Singh, David Sánchez, Josep Domingo-Ferrer, and Alberto Blanco-Justicia</i>	
<b>Case Studies</b>	
Private Posterior Inference Consistent with Public Information: A Case Study in Small Area Estimation from Synthetic Census Data. . . . .	323
<i>Jeremy Seeman, Aleksandra Slavkovic, and Matthew Reimherr</i>	
Differential Privacy and Its Applicability for Official Statistics in Japan – A Comparative Study Using Small Area Data from the Japanese Population Census . . . . .	337
<i>Shinsuke Ito, Takayuki Miura, Hiroto Akatsuka, and Masayuki Terada</i>	
Disclosure Avoidance in the Census Bureau’s 2010 Demonstration Data Product . . . . .	353
<i>David Van Riper, Tracy Kugler, and Steven Ruggles</i>	
<b>Author Index . . . . .</b>	<b>369</b>