
Preface

Data privacy is now a hot topic. Big data have increased its importance. Nevertheless, computational methods for data privacy have been studied and developed since the 70s, at least.

I would say that there are three different communities that work on data privacy from a technological perspective. One is the statistical disclosure control (people with a statistical background), another is the privacy-preserving data mining (people that proceed from databases and data mining), and finally the privacy-enhancing technologies (people that proceed from communications and security).

This book tries to give a general perspective of the field of data privacy in a unified way, and cover at least partially some of the problems and solutions studied by these three communities. The goal is not to present all methods and algorithms for all type of problems, nor the latest algorithms and results. I think that this is an almost impossible task. The goal is to give a broad view of the field, present the different approaches (some of them in their basic form), so that the reader can then deepen in their field of interest.

In this way, the book differs from others that focus on only one of the areas of data privacy. For example, we find the following reference books on statistical disclosure control [1–3], privacy-preserving data mining [4] (edited book), [5], and [6] (focusing on computation-driven/cryptographic approaches). We also edited [7], a book that presents the research on privacy in the ARES project. In addition, there are other books that focus on a specific type of data protection approach or type of problem (e.g., association rule hiding [8], data outsourcing [9], synthetic data generators [10], differential privacy [11], databases and microaggregation [12]). A book that gives a broad picture of privacy for big data is [13] but, in contrast to the others, it does not include details on data privacy technologies.

I have been working on data privacy for more than 15 years. My research has mainly focused on topics related to privacy for databases, and disclosure risk measurement. Because of that, the book is biased into data protection mechanisms for databases.

The book is partially based on the lectures I gave at the Universitat Autònoma de Barcelona and at the University of Linköping; and material from this book was used in these courses.

The book is written for last courses of undergraduate studies in computer engineering, statistics, data science and related fields. The book is expected to be self-contained.

Organization

The structure of the book is as follows. Chapter 1 gives an introduction to the field, reviewing the main terminology. Chapter 2 is a brief summary on techniques in machine and statistical learning. We review those tools that are needed later on. Chapter 3 gives a road-map of data protection procedures. It contains a classification of procedures from different perspectives. The chapter includes two sections on specific data protection methods. One related to result-driven approaches for association rules and the other related to methods for tabular data. Chapter 4 focuses on methods for user's privacy. We discuss methods in communication and for information retrieval.

Chapter 5 discusses privacy models and disclosure risk measures. Naturally, we discuss attribute and identity disclosure. The chapter includes a description of methods for data matching and record linkage as they are used for assessment of disclosure risk.

Chapter 6 is about masking methods (data protection methods for respondent and holder privacy). Literature on masking methods is very large. In this chapter we try to describe the major families of methods, and we include some algorithms of these families. We also refer to some alternative methods, but the chapter does not intend to be exhaustive (an impossible task, indeed!). Selection has been based on simplicity of the algorithm, well-knownness, and personal bias.

Chapter 7 is about information loss and data utility. We review the main approaches for evaluating information loss. This chapter tries to give a broad coverage of the alternative ways used in the literature to evaluate masking methods.

The book finishes with two final chapters, one (Chap. 8) on the selection of a masking method based on disclosure risk and information loss, and another (Chap. 9) that concludes the book.

How to Use This book

The book can be used to give a general introduction on data privacy, describing the different approaches in SDC, PPDM, and PETs. For this, the course would use the first and third chapters of the book. If the course is expected to be technical, then we would use most of the material in the book.

Alternatively, it can be used focusing on masking methods. In this case, emphasis would be given to Chap. 6, on the methods, with an overview of Chaps. 5 and 7.

The book can be used focusing on data privacy for big data. It contains a description of data privacy methods for big data. In Chap. 6 (on masking methods) there is for each family of methods a section focusing on big data, and Sect. 6.6, at the end of the chapter, wraps up all these partial discussions. The sections on each family of methods and their use for big data are as follows. Section 6.1.1 is on rank swapping, Sect. 6.1.2 on microaggregation, Sect. 6.1.3 on additive and multiplicative noise, Sect. 6.1.4 on PRAM, Sect. 6.1.5 on lossy compression, and Sect. 6.4.2 on algorithms for k -anonymity for big data. Then, we also discuss disclosure risk and information loss for big data in the corresponding chapters. That is, Sect. 5.9.3 is on disclosure risk (including a subsection on guidelines and research issues) and Sect. 7.5 is on information loss.

I used parts of this book in courses on data privacy with 6 and 12 h of lectures. The course with 12 h described the main concepts of data privacy, the classification of methods, and a high-level description of disclosure risk and information loss, and a summary of masking methods. A course syllabus for 8 lectures of 2 h follows. The structure maps the chapters of this book.

- L1. Introduction to data privacy.
- L2. Elements of statistical and machine learning.
- L3. Classification of data protection mechanisms.
- L4. Privacy models.
- L5. Masking methods I.
- L6. Masking methods II.
- L7. Information loss. Masking method selection.
- L8. Other privacy protection mechanisms. Result-driven approaches. Methods for tabular data. Methods for user privacy.

The book does not contain exercises. For experimentation, open software as the `sdcMicro` package [14] in R can be used. I used this in my courses. Another open-source software for data anonymization is ARX [15].

Acknowledgements

I was introduced to this field by Josep Domingo-Ferrer when we met in Tarragona, at the Universitat Rovira i Virgili, at the end of the 1990s. So, my acknowledgement first goes to him.

Second, special thanks go to former (Ph.D.) students and postdocs of my research group with whom we have researched in different areas of data privacy: Aïda Valls, Jordi Nin, Jordi Marés, Jordi Casas, Daniel Abril, David Nettleton, Sergi Martínez, Marc Juarez, Susana Ladra, Javier Jimenez, Julián Salas, Cristina Martínez, Javier Herranz, and Guillermo Navarro-Arribas.

Third, to J. Lane, J. Castro, N. Shahmehri, and the people of the ARES and CASC projects. During the period 2008–2014 most of our research on data privacy was funded by the Spanish CONSOLIDER research project titled ARES. The

project gathered most of the research groups in Spain working in the field of privacy and security. CASC was an EU project (2001–2004) on statistical confidentiality. Part of my research described here was funded by these projects. The research of my own described in this book was performed while working at IIIA-CSIC, and lately at SAIL group at the U. of Skövde.

Parts of this manuscript were read and commented by Eva Armengol, Yacine Atif, and Guillermo Navarro. Special thanks go to them.

Last but not least, thanks to my family for their help, and particularly to Klara for the long discussions on privacy-related issues.

Naturally, all errors in the book (except the ones to avoid disclosure) are mine.

Cap de creus and l’Escala
August 2016

Vicenç Torra

References

1. Willenborg, L., de Waal, T.: Elements of Statistical Disclosure Control. Lecture Notes in Statistics. Springer, New York (2001)
2. Duncan, G.T., Elliot, M., Salazar, J.J.: Statistical Confidentiality. Springer, New York (2011)
3. Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Nordholt, E.S., Spicer, K., de Wolf, P.-P.: Statistical Disclosure Control. Wiley, New York (2012)
4. Aggarwal, C.C., Yu, P.S. (eds.): Privacy-Preserving Data Mining: Models and Algorithms. Springer, New York (2008)
5. Fung, B.C.M., Wang, K., Fu, A.W.-C., Yu, P.S.: Introduction to Privacy-Preserving Data Publishing: Concepts and Techniques. CRC Press (2011)
6. Vaidya, J., Clifton, C.W., Zhu, Y.M.: Privacy Preserving Data Mining. Springer, New York (2006)
7. Navarro-Arribas, G., Torra, V. (eds.): Advanced Research in Data Privacy. Springer (2015)
8. Dasseni, E., Verykios, V.S., Elmagarmid, A. K., Bertino, E. Hiding association rules by using confidence and support (2001)
9. Foresti, S.: Preserving Privacy in Data Outsourcing. Springer (2011)
10. Drechsler, J.: Synthetic Datasets for Statistical Disclosure Control: Theory and Implementation. Springer, New York (2011)
11. Li, N., Lyu, M., Su, D., Yang, W.: Differential Privacy: From Theory to Practice, Morgan and Claypool publishers (2016)
12. Domingo-Ferrer, J., Sánchez, D., Soria-Comas, J.: Database Anonymization: Privacy Models, Data Utility, and Microaggregation. Morgan and Claypool publishers (2016)
13. Lane, J., Stodden, V., Bender, S., Nissenbaum, H. Privacy, big data, and the public good. Cambridge University Press (2014)
14. Templ, M.: Statistical disclosure control for microdata using the R-Package sdcMicro. Trans. Data Priv. 1, 67–85 (2008)
15. <http://arx.deidentifier.org/>. Accessed Jan 2017

Data Privacy: Foundations, New Developments and the
Big Data Challenge

Torra, V.

2017, XIV, 269 p. 22 illus., Hardcover

ISBN: 978-3-319-57356-4