Linking Sensitive Data

Peter Christen • Thilina Ranbaduge • Rainer Schnell

Linking Sensitive Data

Methods and Techniques for Practical Privacy-Preserving Information Sharing

Springer

Peter Christen
Research School of Computer Science
The Australian National University
Canberra, ACT, Australia

Thilina Ranbaduge
Research School of Computer Science
The Australian National University
Canberra, ACT, Australia

Rainer Schnell
Institut für Soziologie
Universität Duisburg-Essen
Duisburg, Germany

ISBN 978-3-030-59705-4 ISBN 978-3-030-59706-1 (eBook)
https://doi.org/10.1007/978-3-030-59706-1

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

To Gail, with all my love.
P. C.

To my loving family.
T. R.

To Katrin, my perfect match.
R. S.

Foreword

By now, the potential that data science has for benefitting society must be obvious to everyone. As more and more large data sets describing people and their behaviour accumulate, so the opportunities for improving public policy, for enhancing the efficiency of service industries, for increasing the efficiency of healthcare systems, and for a host of other ways of bettering the human condition are becoming apparent. Many of these possibilities arise as a consequence of linking data sets. Research programs in many countries have been established with the specific aim of combining data from disparate sources to enable opportunities that none of the data sets alone could do.

But all advanced technologies must be handled with care. And this is as true for data science, and in particular for data-linkage technology, as it is for nuclear or bio-technology. To achieve the gains which can be made by linking data sets, we need more than the physical and mathematical advances enabling us to do it. We must also have buy-in from those described by the data. We must handle their data with discretion, preserve their privacy when they want us to, treat their confidential data as sacrosanct, and only disclose what they want us to disclose. And, indeed, more than all this, we must often manage to do it in the face of malicious actors, keen to break into the databases to identify individuals and their characteristics.

Clearly this is a very challenging problem, so I am delighted that the authors of this book, leading experts in the domain of linking sensitive data, have provided us with the answers.

In an extraordinarily comprehensive discussion of linkage technology the book runs over regulatory frameworks, technical details, and practical application. It describes how matching methods work and how to evaluate their performance — something which is in my view under-rated and yet critically important. It covers all the major concepts and methods, including such things as Bloom filters and differential privacy, and also lesser known ideas likely to become more important in the future. But it is not simply an abstract technical manual — it also discusses practical matters such as computational efficiency, which are critical if the methods are to be used in practice. And it does all this in a highly accessible way, telling a fascinating story, ranging from the women who sorted through piles of London Underground tickets in the 1930s linking journeys so they could understand travel patterns, to modern cutting-edge technology involving possibly billions of data points.

This timely book will become a key text for a wide variety of data scientists, whether they are concerned with enhancing the human condition in the public domain, or with launching the latest start-up using data from a variety of sources.

London, UK

David J Hand
Imperial College, London

Preface

Sensitive personal data are created in many application domains, and there is now an increasing demand to share, integrate, and link such data within and across organisations in the public and private sectors. The ultimate aim of such linkage is to enable detailed data analysis that is not possible on individual data sets. The strong emphasis given to pseudo anonymisation (pseudonymisation) in recent privacy legislation, such as the Health Insurance Portability and Accountability Act (HIPAA) in the US and the EU's General Data Protection Regulation (GDPR), calls for novel solutions to allow secure sharing of sensitive information. Furthermore, the difficulty of obtaining individual consent for population covering databases requires the use of privacy-preserving record linkage methods.

Most scientists would consider as the aim of their profession the increase of knowledge by systematically testing theories to explain observed data. Since research also involves generating ideas, the amount of data needed for research cannot, in all cases, be minimised. Therefore, it makes sense to exempt scientific research from general data protection principles such as data minimisation. For example, the GDPR excludes scientific research and official statistics from many general data protection principles. This book is written from the perspective that linking data is a useful tool for scientific research. As other tools, linkage techniques can be used for malicious purposes as well. Therefore, a societal agreement for the use of such techniques is required. The techniques described in this book are designed to minimise the potential misuse of linking data.

A key message of this book is that any database that contains sensitive information about individuals in plaintext can be vulnerable to data breaches and attacks by adversaries, both external and internal to an organisation, as well as unintentional revealing or publication due to human or technical mishaps. Encoding personal sensitive information using the techniques and methods we discuss in this book can significantly reduce the risks of sensitive data being breached or revealed. This is because significant efforts would be required by an adversary to reidentify individuals in an encoded database.

This book covers modern technical answers to the legal requirements of pseudonymisation as recommended by privacy legislation. We describe advanced techniques and concepts for linking sensitive databases using privacy-preserving methods. Using such techniques there is no need to exchange or share private or confidential data that could be used to identify individuals. The book covers topics such as modern regulatory frameworks for sharing and linking sensitive information, concepts and algorithms for privacy-preserving record linkage and their computational aspects, practical considerations such as dealing with dirty and missing data, as well as privacy, risk, and performance assessment measures. Existing techniques for privacy-preserving record linkage are evaluated empirically and real-world application examples that scale to population sizes are described. The book also includes pointers to freely available software tools, benchmark data sets, and tools to generate synthetic data that can be used to test and evaluate linkage techniques.

Intended Audience

The intended audiences of this book include applied scientists, researchers, and practitioners in governments, industry, and universities who are concerned with developing, implementing, and deploying systems and tools to share sensitive information in administrative, commercial, or medical databases. Examples include researchers in public health, road injury research, demography, criminology, history, education, and urban planning, as well as IT managers in hospitals and in government agencies, lawyers in official statistics, data custodians in administration, and public health researchers.

Furthermore, we believe this book to be of high value to graduates from computer science and related fields coming out of university who are starting to work in an organisation that is tasked with linking sensitive data. The non-technical parts of the book will also be of value to decision makers in organisations that are linking sensitive databases as these corresponding chapters will provide high level descriptions of the main concepts of how modern computer based methods can be used to link sensitive data while at the same time the privacy of the individuals whose records are stored in these databases is being protected.

Organisation

This book consists of fourteen chapters grouped into four parts, and two appendices. The first part introduces the reader to the topic of linking sensitive data, the second part covers methods and techniques to link such data, the third part discusses aspects of practical importance, and the fourth part provides an outlook of future challenges and open (research) problems relevant to linking sensitive databases.

The first part consists of three chapters, where the first introduces the topic and motivates why linking databases is an important topic to consider in today's data driven society, and why linking sensitive data can lead to benefits in a variety of application areas as illustrated by several case studies. The second chapter then covers current regulatory frameworks and how they make novel techniques that allow anonymous linking of sensitive data necessary. This chapter also touches on statistical disclosure control (SDC) and how linking sensitive data relates to SDC. We end the first part of the book with Chapter 3 which covers the general aspects of how data can and have been linked, how data quality affects the linking of data, how to evaluate various aspects of the linkage process, and the general challenges of linking databases. We end this chapter with an introduction and formal definition of privacy-preserving record linkage.

We begin the second part of the book with Chapter 4 where we discuss the different conceptual protocols of how sensitive data can be shared and linked between organisations, as well as different models of privacy assumed in these protocols. This is followed by Chapter 5 where we discuss how risk, privacy, and utility can be measured and assessed, and how encoded sensitive data can be attacked by adversaries. We also provide an overview of the related important topic of statistical disclosure control methods. In Chapter 6 we then describe the various building blocks required to link sensitive data, ranging from encoding and encryption techniques to methods that allow names and addresses to be compared, as well as approaches to securely calculate functions across two or more parties. Based on these building blocks, in Chapter 7 we then cover the different techniques that have been proposed over the past two decades to allow the privacy-preserving linkage of sensitive data. In Chapter 8 we describe in detail Bloom filter encoding, the currently most widely used approach to linking sensitive data in a privacy-preserving way, and we discuss advantages and problems with this technique. Chapter 9 continues to cover Bloom filter encoding by describing several recently proposed cryptanalysis attack methods that have been developed with the aim to reidentify sensitive values encoded in Bloom filters, and hardening techniques that aim to overcome these attacks. We conclude the second part of the book with Chapter 10 discussing computational aspects that are becoming increasingly important as the databases to be linked are becoming ever larger. We describe blocking and indexing techniques, approaches that make use of modern parallel and distributed computing platforms, and how to link multiple (more than two) or even many (dozens to thousands) of sensitive databases.

The third part of the book in Chapter 11 discusses various practical aspects of linking sensitive databases, including how to deal with low quality data or incomplete or even missing data, and how to link heterogeneous, temporal, and dynamic data that are becoming more widespread in today's Big data applications, where data are collected in an ongoing basis and therefore often need to be processed, linked, and analysed in (near) real time. We also discuss practical implementation aspects, how to set and tune parameters for the algorithms and techniques described in the third part of the book, and what computational requirements to consider for practical use of these techniques. In Chapter 12 we then present a comparative evaluation of selected privacy-preserving record linkage techniques on example data sets, and how these techniques perform with regard to linkage quality, scalability, and the privacy protection they provide. Chapter 13 concludes the third part of the book with descriptions of selected real-world applications where sensitive databases are being linked in practice.

The fourth part of the book consists of Chapter 14 where we discuss future research challenges and directions, both practical problems as well as open conceptual challenges. We also describe new challenges posed by Big data applications, as well as the linking of other types of data such as biometric and genetic information about individuals, which opens up not only technical challenges but also new legal and ethical questions.

Finally, in Appendix A we provide pointers and describe currently existing software systems that allow the linkage of sensitive data. We limit ourselves to freely available, open-source software rather than commercial systems. In Appendix B we then provide further details about the evaluation presented in Chapter 12 to allow the interested reader install the software used for this evaluation and rerun the presented experiments.

We provide an extensive glossary, on page 397, covering many terms relevant to linking databases, sensitive data, and privacy aspects related to record linkage. Further notations used in this book are described on page xxi.

A companion Web site at https://dmm.anu.edu.au/lsdbook2020 provides additional material, such as the Python programs we used for the empirical evaluation described in Chapter 12 and Appendix B, any errata of the book, as well as electronic versions of the table of contents, glossary, and references.

Keywords: Data linkage, record linkage, data matching, entity resolution, administrative data, personal data, microdata, privacy, privacy-preserving, anonymisation, pseudonymisation, encoding, encryption, hashing, Bloom filter, GDPR, HIPAA.

Acknowledgements

The idea of this book started when the three of us were participating at the *Data Linkage and Anonymisation* programme held in 2016 at the *Isaac Newton Institute* (INI) for Mathematical Sciences at the University of Cambridge, UK. We therefore like to thank the INI for their fantastic support during this programme, which was funded by EPSRC grant EP/K032208/1.

We also like to thank David J. Hand, OBE, Imperial College London, for writing an inspiring foreword highlighting the importance of the topics covered in our book. A special thanks goes to our editor Ralf Gerstner from Springer, who supported this book project right from the start, and the anonymous reviewer who provided valuable detailed feedback and helpful suggestions. We like to thank Christian Borgs, Anushka Vidanage, and Sirintra Vaiwsri for co-authoring parts of certain chapters, Abel Kho and Brad Malin for advise and providing pointers to US resources on linking sensitive data, and Frauke Kreuter for commenting on the first part of the book. A big "thank you" goes also to Asara Senaratne, Anushka Vidanage, Charini Nanayakkara, Nishadi Kirielle, Sirintra Vaiwsri, Yanling Chen, and Youzhe Heng, for providing valuable feedback and proof-reading drafts of this book. All remaining errors are of course ours.

Peter Christen likes to acknowledge the Simons Foundation which supported his stay in Cambridge in 2016. He also likes to acknowledge the *Administrative Data Research Centre Scotland* (ADRC-S) and the *Digitising Scotland* project which funded his stays in Edinburgh, as well as Tash Vest in Greenwich, and Divers Lodge Lembeh and Liberty Dive Resort, both in Indonesia, where parts of this book were written. Peter furthermore likes to acknowledge the funding he received from the Australian Research Council (ARC) for conducting research on how to link sensitive databases under the two Discovery Projects DP130101801 and DP160101934.

Thilina Ranbaduge is sincerely thankful for the funding provided by the Australian Research Council (ARC Discovery Project DP160101934) for his research, without which it would not have been possible. He also thanks the

Research School of Computer Science and the Australian National University for offering him an opportunity to conduct his research studies. The school and university are well supportive of early career researchers.

Rainer Schnell thanks the University of London, City, to kindly relieve him from some of the duties in London to spend several months at the Isaac Newton Institute in Cambridge in 2016. He was supported by the German Research Foundation (DFG) by six different research grants on record linkage since 2005 (DFG-Grants 5369360, 200001560, 161924790, 407023611, 258933986, 87664861). Without these fundings, the development of many techniques described in this book would have been impossible. As part of these grants, DFG funded the setup of the German Record Linkage Center for its first years.

Canberra, *Peter Christen*
Canberra, *Thilina Ranbaduge*
Lechtingen, *Rainer Schnell*
10 August 2020

Contents

Notations

In the following we describe the style and mathematical notations used throughout this book. Additional notations will be introduced in specific chapters and sections as required. Furthermore, the glossary starting on page 397 describes many terms relevant to the topics covered in this book.

Throughout the book we show example textual values with single quotes, such as ‘John Smith’; while we show example attribute (or field) names in small caps font, for example FIRSTNAME or POSTCODE.

With regard to mathematical symbols and equations, we denote simple variables such as numbers or text strings using lowercase italics font (such as a, b, c); lists, sets, and vectors using lowercase bold font (for example $\mathbf{a}$, $\mathbf{b}$, $\mathbf{c}$); while for matrices, lists, and sets of lists, vectors, or sets we use uppercase bold font (such as $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$). Sets in the mathematical sense do not have an order, while lists and vectors (both one-dimensional) and matrices (two-dimensional) are ordered collections of elements. We denote sets with curly brackets, for example the set $\mathbf{s}$ of numbers from 1 to 9 (unordered) could be $\mathbf{s} = \{5, 9, 1, 3, 8, 2, 7, 6, 4\}$. Lists and vectors are shown with square brackets and their elements are indexed from 0 onwards. For example, the ordered list $\mathbf{l}$ of numbers from 100 to 106 is denoted by $\mathbf{l} = [100, 101, 102, 103, 104, 105, 106]$, where the first element in $\mathbf{l}$ is $\mathbf{l}[0] = 100$ and the fifth element is $\mathbf{l}[4] = 104$. Similarly, elements in a matrix are denoted by their row and column indices, both starting from 0. For example, $\mathbf{M}[1, 3]$ will be the element in the second row and fourth column in matrix $\mathbf{M}$.

We denote the number of elements in a set (its size) and the length of a text string (number of characters), list, or vector (number of elements) with two vertical bars: $l = |\mathbf{s}|$. For the example set $\mathbf{s}$ given above, this would give $l = 9$, while for the string $s =$ ‘hello’ its length $l = |s|$ is $l = 5$. We use $||$ to symbolise the concatenation of strings, for example ‘hello’ $||$ ‘World’ results in the concatenated string ‘helloWorld’.

For operations on bit vectors, such as Bloom filters as described in Chapters 8 and 9, we use $\wedge$ for the bitwise AND, $\vee$ for the bitwise OR, and $\oplus$

for the bitwise XOR (exclusive OR) operations, where the outcomes of these operations are shown in the following three tables.

Bitwise AND

x	y	$x \wedge y$
0	0	0
0	1	0
1	0	0
1	1	1

Bitwise OR

x	y	$x \vee y$
0	0	0
0	1	1
1	0	1
1	1	1

Bitwise XOR

x	y	$x \oplus y$
0	0	0
0	1	1
1	0	1
1	1	0

For example, for the two bit vectors $\mathbf{b}_1 = [1, 0, 0, 1]$ and $\mathbf{b}_2 = [1, 1, 0, 0]$, we obtain $\mathbf{b}_1 \wedge \mathbf{b}_2 = [1, 0, 0, 0]$, $\mathbf{b}_1 \vee \mathbf{b}_2 = [1, 1, 0, 1]$, and $\mathbf{b}_1 \oplus \mathbf{b}_2 = [0, 1, 0, 1]$.