

Linking Sensitive Data

Peter Christen • Thilina Ranbaduge • Rainer Schnell

Linking Sensitive Data

Methods and Techniques for Practical
Privacy-Preserving Information Sharing

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>

© Springer Nature Switzerland AG 2020

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

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 pro-

vides 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,
Canberra,
Lechtingen,
10 August 2020

Peter Christen
Thilina Ranbaduge
Rainer Schnell

Contents

Part I Introduction	1
1 Introduction	3
1.1 The Increase in Linking Data	3
1.2 Why Should Data be Linked at All?	5
1.3 Sources of Data and their Linkage	6
1.4 Direct and Indirect Identifiers	7
1.5 What are Sensitive Data?	9
1.6 Example Case Studies	10
1.6.1 Financial Fraud	10
1.6.2 Law Enforcement and Counter Terrorism	11
1.6.3 Health Service Research	12
1.6.4 Longitudinal Studies	14
1.6.5 Survey Methodology	16
1.6.6 Official Statistics	19
1.7 Ethical Challenges	23
1.8 What this Book Covers	24
1.9 Summary and Further Reading	25
2 Regulatory Frameworks	27
2.1 Privacy Norms: The Privacy Paradox and Contextual Integrity	27
2.2 Basic Ethical Principles of Research	29
2.3 Regulations in the European Union and the United Kingdom	30
2.3.1 Austria	32
2.3.2 Germany	33
2.3.3 United Kingdom	34
2.4 Regulations in the United States	36
2.5 Regulations in other Countries	37
2.5.1 Australia	37
2.5.2 Switzerland	38
2.6 Statistical Disclosure Control	38
2.7 Best Practice Approaches	39
2.7.1 Organisational Measures	40
2.7.2 Professional Guidelines	41
2.7.3 Social Embeddings of Research	42

2.8	Summary and Further Reading	44
3	Linking Sensitive Data Background	47
3.1	A Short History of Linking Data	47
3.2	The Process of Linking Records across Databases	50
3.3	Data Quality Aspects Relevant to Linking Databases	57
3.4	Evaluation Measures	60
3.4.1	Linkage Quality Measures	60
3.4.2	Group Linkage Quality Measures	64
3.4.3	Linkage Complexity Measures	66
3.5	Major Challenges to Linking Data	69
3.6	Introduction to Privacy-Preserving Record Linkage	72
3.7	Summary and Further Reading	75
	 Part II Methods and Techniques	 77
4	Private Information Sharing Protocols	81
4.1	Roles of Different Linkage Participants	81
4.2	Separation Principle	83
4.3	Linkage Protocols	87
4.4	Adversarial Models	89
4.5	Additional Aspects of Private Information Sharing Protocols	93
4.5.1	Secure Key Exchange Algorithms	94
4.5.2	Access Control Mechanisms	96
4.6	Summary and Further Reading	97
5	Assessing Privacy and Risks	99
5.1	Measuring Privacy and Risks when Linking Sensitive Data . .	99
5.2	Privacy Measures for Linking Sensitive Databases	102
5.2.1	Information Entropy based Privacy Measures	102
5.2.2	Disclosure Risk based Privacy Measures	104
5.3	Data Breaches and Mishaps when Dealing with Sensitive Data	106
5.4	Attacks on Sensitive Data	108
5.4.1	Insider Attacks and Social Engineering	109
5.4.2	Dictionary Attacks	110
5.4.3	Frequency Attacks	111
5.4.4	Composition Attacks	112
5.4.5	Collusion Attacks	113
5.4.6	Linkage Attacks	115
5.4.7	Motivation, Costs, and Gains of Attacks	117
5.5	Statistical Disclosure Control Methods	118
5.5.1	Statistical Disclosure Control Techniques	118
5.5.2	Evaluating Statistical Disclosure Control Techniques .	120
5.6	Summary and Further Reading	122

6	Building Blocks for Linking Sensitive Data	123
6.1	Random Number Generation	123
6.2	Hashing Techniques	125
6.2.1	One-way Hashing	127
6.2.2	Keyed Cryptographic Hashing and Message Authentication	129
6.2.3	Locality Sensitive Hashing	131
6.3	Anonymisation and Pseudonymisation Techniques	134
6.3.1	Randomisation	137
6.3.2	Generalisation	138
6.3.3	Differential Privacy	141
6.4	Encryption Techniques	142
6.4.1	Symmetric Key Encryption	144
6.4.2	Public Key Encryption	146
6.4.3	Homomorphic Encryption	148
6.5	Secure Multiparty Computation	149
6.5.1	Secure Summation	150
6.5.2	Secure Set Intersection	151
6.5.3	Oblivious Transfer Protocols	152
6.5.4	Secret Sharing	152
6.6	Phonetic Encoding	153
6.7	Statistical Linkage Keys	154
6.8	Similarity Measures	156
6.8.1	Set-based Similarities between Strings	157
6.8.2	Edit-based Similarities between Strings	159
6.8.3	Calculating Similarities between Numerical Values	162
6.8.4	Calculating Similarities between Date Values	164
6.9	Choosing Suitable Building Blocks	165
6.10	Summary and Further Reading	167
7	Encoding and Comparing Sensitive Values	169
7.1	A Taxonomy of Techniques for Linking Sensitive Values	169
7.2	Generations of Privacy-Preserving Linkage Techniques	171
7.3	Phonetic Encoding based Techniques	172
7.4	Hashing-based Techniques	174
7.5	Reference Values based Techniques	180
7.6	Embedding-based Techniques	182
7.7	Differential Privacy based Techniques	184
7.8	Secure Multiparty Computation based Techniques	186
7.9	Choosing Suitable Encoding Techniques	189
7.10	Summary and Further Reading	190
8	Bloom Filter based Encoding Methods	193
8.1	Bloom Filter Encoding	193
8.2	Hashing Techniques for Bloom Filters	196

8.2.1	Double Hashing	197
8.2.2	Triple Hashing	197
8.2.3	Enhanced Double Hashing	198
8.2.4	Random Hashing	199
8.3	Encoding Techniques for Textual Data	200
8.3.1	Attribute Level Bloom Filter Encoding	200
8.3.2	Cryptographic Long-term Key	201
8.3.3	Record Level Bloom Filters	203
8.3.4	CLK-RBF	205
8.4	Encoding Numerical Data	206
8.4.1	Absolute Difference Similarity Encoding	206
8.4.2	Distance Aware Numerical Encoding	210
8.5	Encoding Hierarchical Classification Codes	211
8.6	Choosing Suitable Settings for Bloom Filter Encoding	213
8.6.1	Encoding Parameters and Best Practice Suggestions	213
8.6.2	Optimal Parameters for Bloom Filter Encoding	214
8.7	Summary and Further Reading	218
9	Attacking and Hardening Bloom Filter Encoding	221
9.1	Overview of Attack Methods	221
9.2	Frequency-based Cryptanalysis Attacks	224
9.2.1	Constrain Satisfaction based Attack	224
9.2.2	Bloom Filter Atoms based Attack	225
9.2.3	Bloom Filter Construction Principle based Attack	228
9.3	Pattern Mining based Cryptanalysis Attacks	231
9.4	Graph based Cryptanalysis Attacks	235
9.4.1	Q-gram Graph based Attack	235
9.4.2	Similarity Graph based Attack	237
9.5	Hardening Techniques for Bloom Filter Encoding	238
9.5.1	Salting	239
9.5.2	Balancing	240
9.5.3	XOR-folding	242
9.5.4	Rule 90	243
9.5.5	Adding Random Noise	244
9.5.6	Bloom and Flip	244
9.5.7	Rehashing	246
9.5.8	Markov Chaining	247
9.6	Recommended Best Practice for Bloom Filter Hardening	249
9.7	Summary and Further Reading	250
10	Computational Efficiency	253
10.1	Blocking and Indexing Techniques	253
10.1.1	Requirements of Privacy-Preserving Blocking	255
10.1.2	Phonetic Blocking	257
10.1.3	Reference Values based Blocking	259

10.1.4	Hashing-based Blocking	260
10.1.5	Multibit Tree based Blocking	263
10.2	Meta-Blocking Techniques	264
10.3	Filtering Techniques	266
10.3.1	Length Filtering	266
10.3.2	Prefix and Position Filtering	267
10.3.3	Metric Space Filtering	269
10.4	Blocking and Indexing for Multiple Parties	270
10.5	Many-Party Linkage Methods	274
10.6	Parallel and Distributed Computing Techniques	278
10.6.1	Non-framework based Parallel Approaches	279
10.6.2	Framework based Parallel Approaches	281
10.6.3	Communication Protocols	282
10.7	Summary and Further Reading	284

Part III Practical Aspects, Evaluation, and Applications **287**

11	Practical Considerations	289
11.1	Introduction	289
11.2	Data Related Considerations	290
11.2.1	Dealing with Dirty Data	291
11.2.2	Dealing with Missing Values	293
11.2.3	Temporal and Dynamic Data	297
11.2.4	Dealing with Bias in Linked Data	299
11.2.5	Availability or Lack of Ground Truth Data	303
11.2.6	Costs of False Matches and False Non-Matches	306
11.3	Technical Considerations	307
11.3.1	Suitability of Linkage Protocols	308
11.3.2	Suitability of Linkage Techniques	310
11.3.3	Availability of Software	312
11.3.4	Customisation and Parameter Tuning	314
11.3.5	Computational Requirements	314
11.4	Institutional Considerations	315
11.4.1	Required Domain and Technical Expertise	316
11.4.2	Legal and Ethical Concerns	317
11.5	Guidelines for Practical Linkage Projects	319
11.6	Summary and Further Reading	320
12	Empirical Evaluation	323
12.1	Evaluation Framework and Setup	323
12.1.1	Databases used in Evaluation	324
12.1.2	Experimental Setup	325
12.2	Evaluating Linkage Quality	327

12.3	Evaluating Scalability	330
12.4	Evaluating Privacy	333
12.4.1	Frequency Distributions of 1-bits in Bloom Filters . .	335
12.4.2	Attack Evaluation	336
12.5	Summary and Further Reading	343
13	Real-world Applications	345
13.1	Australia	345
13.2	Brazil	347
13.3	Canada	348
13.4	Germany	349
13.5	Switzerland	350
13.6	United Kingdom	352
13.7	United States	355
13.8	Summary and Further Reading	357
Part IV	Outlook	359
14	Future Research Challenges and Directions	361
14.1	Conceptual Research Questions	361
14.2	Practical Challenges	365
14.3	Linking in the Era of Big Data	369
14.4	Linking Biometric and Genetic Data	371
14.5	Summary and Further Reading	374
Appendices		377
A	Software and Data Sets	379
A.1	Software Prototypes	379
A.2	Public Data Collections and Benchmark Data Sets	383
A.3	Synthetic Data Generation	385
A.4	Summary and Further Reading	390
B	Details of the Empirical Evaluation	391
B.1	Modules Overview	391
B.2	Installation Requirements	393
B.3	Examples for Running an Evaluation	394
B.4	Data Set Generation	396
Glossary		397
References		419
Index		463

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 **a**, **b**, **c**); while for matrices, lists, and sets of lists, vectors, or sets we use uppercase bold font (such as **A**, **B**, **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 **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 **l** of numbers from 100 to 106 is denoted by $\mathbf{l} = [100, 101, 102, 103, 104, 105, 106]$, where the first element in **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 **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 **s** given above, this would give $l = 9$, while for the string $s = \text{‘hello’}$ its length $l = |s|$ is $l = 5$. We use $||$ to symbolise the concatenation of strings, for example $\text{‘hello’} || \text{‘World’}$ 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 $\mathbf{b}_1 \oplus \mathbf{b}_2 = [0, 1, 0, 1]$.