Peter Christen, Thilina Ranbaduge, and
Rainer Schnell

Linking Sensitive Data

Methods and Techniques for Practical Privacy-Preserving Information Sharing

Springer

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

Contents

References

[1] Aberer, K., Datta, A., Hauswirth, M.: A decentralized public key infrastructure for customer-to-customer e-commerce. International Journal of Business Process Integration and Management **1**, 26–33 (2005)

[2] Abowd, J.: How will statistical agencies operate when all data are private? Journal of Privacy and Confidentiality **7**(3) (2017)

[3] Acar, A., Aksu, H., Uluagac, A.S., Conti, M.: A survey on homomorphic encryption schemes: Theory and implementation. ACM Computing Surveys **51**(4), 79 (2018)

[4] Acheson, E.: Oxford record linkage study: a central file of morbidity and mortality records for a pilot population. British Journal of Preventive and Social Medicine **18**(1), 8 (1964)

[5] Adams, C., Lloyd, S.: Understanding Public-key Infrastructure: Concepts, Standards, and Deployment Considerations. Sams Publishing (1999)

[6] Aggarwal, C.C., Yu, P.S.: Privacy-preserving Data Mining: Models and Algorithms, *Advances in Database Systems*, vol. 34. Springer (2008)

[7] Agrawal, R., Evfimievski, A., Srikant, R.: Information sharing across private databases. In: ACM Conference on Management of Data, pp. 86–97. San Diego (2003)

[8] Agrawal, R., Srikant, R.: Fast algorithms for mining association rules. In: Conference on Very Large Data Bases, pp. 487–499. Santiago de Chile (1994)

[9] Akgün, Ö., Dearle, A., Kirby, G., Christen, P.: Using metric space indexing for complete and efficient record linkage. In: Pacific-Asia Conference on Knowledge Discovery and Data Mining, pp. 89–101. Melbourne (2018)

[10] Akgün, Ö., Dearle, A., Kirby, G., Garrett, E., Dalton, T., Christen, P., Dibben, C., Williamson, L.: Linking Scottish vital event records using family groups. Historical Methods: A Journal of Quantitative and Interdisciplinary History pp. 1–17 (2019)

[11] Al-Lawati, A., Lee, D., McDaniel, P.: Blocking-aware private record linkage. In: International Workshop on Information Quality in Information Systems, pp. 59–68. Baltimore (2005)
[12] Alaggan, M., Cunche, M., Gambs, S.: Privacy-preserving Wi-Fi analytics. Privacy Enhancing Technologies **2018**(2), 4–26 (2018)
[13] Alaggan, M., Gambs, S., Kermarrec, A.M.: BLIP: non-interactive differentially-private similarity computation on Bloom filters. In: Symposium on Self-Stabilizing Systems, pp. 202–216. Toronto (2012)
[14] Alhaqbani, B., Fidge, C.: Access control requirements for processing electronic health records. In: BPM, pp. 371–382. Springer, Vienna (2007)
[15] Ali, M.S., Ichihara, M.Y., Lopes, L.C., Barbosa, G.C., Pita, R., Carreiro, R.P., dos Santos, D.B., Ramos, D., Bispo, N., Raynal, F., et al.: Administrative data linkage in Brazil: potentials for health technology assessment. Frontiers in Pharmacology **10** (2019)
[16] Allahbakhsh, M., Ignjatovic, A., Benatallah, B., Bertino, E., Foo, N., et al.: Collusion detection in online rating systems. In: ApWeb, pp. 196–207. Springer, Sydney (2013)
[17] Almquist, Y.B., Grotta, A., Vågerö, D., Stenberg, S.Å., Modin, B.: Cohort profile update: The Stockholm birth cohort study. International Journal of Epidemiology **49**(2), 367–367e (2020)
[18] Alvarez, R., Caballero-Gil, C., Santonja, J., Zamora, A.: Algorithms for lightweight key exchange. Sensors **17**(7), 1517 (2017)
[19] American Academy of Arts & Sciences: Perceptions of science in America. `https://www.amacad.org/sites/default/files/publication/downloads/PFoS-Perceptions-Science-America.pdf` (2018)
[20] Anderson, K., Durbin, E., Salinger, M.: Identity theft. The Journal of Economic Perspectives **22**(2), 171–192 (2008)
[21] Andreou, A., Goga, O., Loiseau, P.: Identity vs. attribute disclosure risks for users with multiple social profiles. In: IEEE/ACM Conference on Advances in Social Networks Analysis and Mining, pp. 163–170. Calgary (2017)
[22] Angrist, J.D., Krueger, A.B.: Empirical strategies in labor economics. In: O.C. Ashenfelter, D. Card (eds.) Handbook of Labor Economics, vol. 3, pp. 1277–1366. Elsevier, Amsterdam (1999)
[23] Antoni, M.: Record linkage of GDR's 'Data Fund of Societal Work Power' with administrative labour market biography data of the German Federal Employment Agency. German Record-Linkage Center Working paper series, 2018-02, Nürnberg (2018)
[24] Antoni, M., Schnell, R.: The past, present and future of the German Record Linkage Center. Journal of Economics and Statistics (2017)
[25] Arasu, A., Götz, M., Kaushik, R.: On active learning of record matching packages. In: ACM Conference on Management of Data, pp. 783–794. Indianapolis (2010)

[26] de Araujo Almeida, B., Barreto, M.L., Ichihara, M.Y., Barreto, M.E., Cabral, L., Fiaccone, R., Carreiro, R.P., Teles, C., Pita, R., Penna, G., et al.: The center for data and knowledge integration for health (CIDACS). International Journal of Population Data Science **4**(2) (2019)
[27] Arp, D., Quiring, E., Krueger, T., Dragiev, S., Rieck, K.: Privacy-enhanced fraud detection with Bloom filters. In: International Conference on Security and Privacy in Communication Systems, pp. 396–415. Singapore (2018)
[28] Asadova, S.: Privacy-preserving DNA sequence alignment. Ph.D. thesis, Master's thesis, Eindhoven University of Technology (2017)
[29] Ashton, K.: That 'Internet of Things' thing. RFID Journal **22**(7) (2009)
[30] Atallah, M., Kerschbaum, F., Du, W.: Secure and private sequence comparisons. In: Workshop on Privacy in the Electronic Society, pp. 39–44. Washington DC (2003)
[31] Audette, L.M., Hammond, M.S., Rochester, N.K.: Methodological issues with coding participants in anonymous psychological longitudinal studies. Educational and Psychological Measurement (2019)
[32] Aumann, Y., Lindell, Y.: Security against covert adversaries: Efficient protocols for realistic adversaries. In: Theory of Cryptography Conference, pp. 137–156. Amsterdam (2007)
[33] Australian Bureau of Statistics: Housing mobility and conditions, 2007–08 (2009)
[34] Australian National University: Incident Report on the Breach of the Australian National University's Administrative Systems (2019). Canberra, Australia
[35] Averdijk, M., Elffers, H.: The discrepancy between survey-based victim accounts and police reports revisited. International Review of Victimology **18**(2), 91–107 (2012)
[36] Avigad, J., Donnelly, K.: Formalizing O notation in Isabelle/HOL. In: Joint Conference on Automated Reasoning, pp. 357–371. Cork (2004)
[37] Bacher, J., Brand, R., Bender, S.: Re-identifying register data by survey data using cluster analysis: an empirical study. International Journal of Uncertainty, Fuzziness and Knowledge Based Systems **10**(5), 589–607 (2002)
[38] Bachteler, T., Reiher, J., Schnell, R.: Similarity filtering with multibit trees for record linkage. German Record Linkage Center, Nüremberg, Working Paper WP-GRLC-2013-02 (2013)
[39] Bakker, C.: Valuing the census. Technical Report, Statistics New Zealand (2013). URL `http://archive.stats.govt.nz/methods/research-papers/topss/valuing-census.aspx`
[40] Ballantyne, A.: Adjusting the focus: A public health ethics approach to data research. Bioethics **33**(3), 357–366 (2019)

[41] Ballantyne, A., Schaefer, G.O.: Consent and the ethical duty to participate in health data research. Journal of Medical Ethics **44**(6), 392–396 (2018)
[42] Bar-On, A., Dunkelman, O., Keller, N., Ronen, E., Shamir, A.: Improved key recovery attacks on reduced-round AES with practical data and memory complexities. In: International Cryptology Conference, pp. 185–212. Santa Barbara (2018)
[43] Barker, E., Barker, W., Burr, W., Polk, W., Smid, M.: Recommendation for key management part 1: General (revision 3). NIST Special Publication **800**(57), 1–147 (2012)
[44] Barros, J.E., French, J.C., Martin, W.N., Kelly, P.M., Cannon, T.M.: Using the triangle inequality to reduce the number of comparisons required for similarity-based retrieval. In: Storage and Retrieval for Still Image and Video Databases IV, vol. 2670, pp. 392–403. La Jolla, San Diego (1996)
[45] Batini, C., Scannapieco, M.: Data and Information Quality. Data-Centric Systems and Applications. Springer (2016)
[46] Bayardo, R.: Efficiently mining long patterns from databases. ACM SIGMOD Record **27**(2), 85–93 (1998)
[47] Bellahsene, Z., Bonifati, A., Rahm, E.: Schema Matching and Mapping. Data-Centric Systems and Applications. Springer (2011)
[48] Bellare, K., Iyengar, S., Parameswaran, A.G., Rastogi, V.: Active sampling for entity matching. In: ACM Conference on Knowledge Discovery and Data Mining, pp. 1131–1139. Beijing (2012)
[49] Benaloh, J.C.: Secret sharing homomorphisms: Keeping shares of a secret secret. In: Theory and Application of Cryptographic Techniques, pp. 251–260. Santa Barbara (1986)
[50] Benchimol, E.I., Smeeth, L., Guttmann, A., Harron, K., Moher, D., Petersen, I., Sørensen, H.T., von Elm, E., Langan, S.M., Committee, R.W., et al.: The REporting of studies Conducted using Observational Routinely-collected health Data (RECORD) statement. PLOS Med **12**(10), e1001,885 (2015)
[51] Benitez, K., Malin, B.: Evaluating re-identification risks with respect to the HIPAA privacy rule. Journal of the American Medical Informatics Association **17**(2), 169–177 (2010)
[52] Benjelloun, O., Garcia-Molina, H., Gong, H., Kawai, H., Larson, T., Menestrina, D., Thavisomboon, S.: D-Swoosh: A family of algorithms for generic, distributed entity resolution. In: IEEE ICDCS, pp. 37–37. Toronto (2007)
[53] Benjelloun, O., Garcia-Molina, H., Menestrina, D., Su, Q., Whang, S., Widom, J.: Swoosh: a generic approach to entity resolution. VLDB Journal **18**(1), 255–276 (2009)
[54] Benschop, T., Machingauta, C., Welch, M.: Statistical disclosure control: A practice guide. `https://sdcpractice.readthedocs.io/en/latest/` (2019)

[55] Berent, M.K., Krosnick, J.A., Lupia, A.: Measuring voter registration and turnout in surveys: Do official government records yield more accurate assessments? Public Opinion Quarterly **80**(3), 597–621 (2016)
[56] Berger, J.M.: A note on error detection codes for asymmetric channels. Information and Control **4**(1), 68–73 (1961)
[57] Berlin, C., Techel, F., Moor, B.K., Zwahlen, M., Hasler, R.M., et al.: Snow avalanche deaths in Switzerland from 1995 to 2014 – Results of a nation-wide linkage study. PLOS One **14**(12) (2019)
[58] Bhattacharya, I., Getoor, L.: Collective entity resolution in relational data. ACM Transactions on Knowledge Discovery from Data **1**(1) (2007)
[59] Bian, J., Loiacono, A., Sura, A., Mendoza Viramontes, T., Lipori, G., Guo, Y., Shenkman, E., Hogan, W.: Implementing a hash-based privacy-preserving record linkage tool in the OneFlorida clinical research network. Journal of the American Medical Informatics Association Open **2**(4), 562–569 (2019)
[60] Bianchi, G., Bracciale, L., Loreti, P.: "Better than nothing" privacy with Bloom filters: To what extent? In: Privacy in Statistical Databases, pp. 348–363. Palermo (2012)
[61] Biemer, P.: Errors and inference. In: I. Foster, R. Ghani, R.S. Jarmin, F. Kreuter, J. Lane (eds.) Big Data and Social Science, chap. 10, pp. 265–297. CRC Press, Boca Raton (2017)
[62] Bilenko, M., Mooney, R.J.: Adaptive duplicate detection using learnable string similarity measures. In: ACM Conference on Knowledge Discovery and Data Mining, pp. 39–48. Washington DC (2003)
[63] Bizer, C., Heath, T., Berners-Lee, T.: Linked data: The story so far. In: Semantic Services, Interoperability and Web Applications: Emerging Concepts, pp. 205–227. IGI Global (2011)
[64] Blair, J., Czaja, R.F., Blair, E.A.: Designing Surveys: A Guide to Decisions and Procedures, 3 edn. Sage, Thousand Oaks (2014)
[65] Blakely, T., Salmond, C.: Probabilistic record linkage and a method to calculate the positive predictive value. International Journal of Epidemiology **31:6**, 1246–1252 (2002)
[66] Blakley, G.R., et al.: Safeguarding cryptographic keys. In: National Computer Conference. New York (1979)
[67] Blondel, B., Cuttini, M., Hindori-Mohangoo, A., Gissler, M., Loghi, M., Prunet, C., Heino, A., Smith, L., van der Pal-de Bruin, K., Macfarlane, A., Zeitlin, J.: How do late terminations of pregnancy affect comparisons of stillbirth rates in Europe? analyses of aggregated routine data from the Euro-Peristat Project. BJOG: An International Journal of Obstetrics and Gynaecology **125**(2), 226–234 (2018)
[68] Bloom, B.: Space/time trade-offs in hash coding with allowable errors. Communications of the ACM **13**(7), 422–426 (1970)
[69] Bloor, M., Leyland, A., Barnard, M., McKeganey, N.: Estimating hidden populations: A new method of calculating the prevalence of drug-

injecting and non-injecting female street prostitution. British Journal of Addiction **86**(11), 1477–1483 (1991)
[70] Blustein, J., El-Maazawi, A.: Bloom filters – a tutorial, analysis, and survey. Tech. rep., Dalhousie University, Halifax (2002)
[71] Bohensky, M.A., Jolley, D., Sundararajan, V., Evans, S., Pilcher, D.V., Scott, I., Brand, C.A.: Data linkage: a powerful research tool with potential problems. BMC Health Services Research **10**(346), 1–7 (2010)
[72] Boneh, D., et al.: Twenty years of attacks on the RSA cryptosystem. Notices of the AMS **46**(2), 203–213 (1999)
[73] Bonomi, L., Fan, L., Xiong, L.: A review of privacy preserving mechanisms for record linkage. In: A. Gkoulalas-Divanis, G. Loukides (eds.) Medical Data Privacy Handbook, pp. 233–265. Springer (2015)
[74] Bonomi, L., Xiong, L., Chen, R., Fung, B.: Frequent grams based embedding for privacy preserving record linkage. In: ACM Conference on Information and Knowledge Management, pp. 1597–1601. Maui, Hawaii (2012)
[75] Bonomi, L., Xiong, L., Lu, J.J.: LinkIT: privacy preserving record linkage and integration via transformations. In: ACM Conference on Management of Data, pp. 1029–1032. New York (2013)
[76] Bopp, M., Spoerri, A., Zwahlen, M., Gutzwiller, F., Paccaud, F., Braun-Fahrlander, C., Rougemont, A., Egger, M.: Cohort profile: The Swiss national cohort – a longitudinal study of 6.8 Million people. International Journal of Epidemiology **38**(2), 379–384 (2009)
[77] Borgs, C.: Optimal parameter choice for Bloom filter-based privacy-preserving record linkage. Ph.D. thesis, University of Duisburg-Essen, Germany (2019)
[78] Borthakur, D., et al.: HDFS architecture guide. Hadoop Apache Project **53**(1-13), 2 (2008)
[79] Bose, P., Guo, H., Kranakis, E., Maheshwari, A., Morin, P., Morrison, J., Smid, M., Tang, Y.: On the false-positive rate of Bloom filters. Information Processing Letters **108**(4), 210–213 (2008)
[80] Bosu, A., Liu, F., Yao, D.D., Wang, G.: Collusive data leak and more: Large-scale threat analysis of inter-app communications. In: ACM Asia Conference on Computer and Communications Security, pp. 71–85. Abu Dhabi (2017)
[81] Boyd, J.H., Ferrante, A.M., Irvine, K., Smith, M., Moore, E., Brown, A., Randall, S.M.: Understanding the origins of record linkage errors and how they affect research outcomes. Australian and New Zealand Journal of Public Health **41**(2), 215–215 (2017)
[82] Boyd, J.H., Ferrante, A.M., O'Keefe, C.M., Bass, A.J., Randall, S.M., Semmens, J.B.: Data linkage infrastructure for cross-jurisdictional health-related research in Australia. BMC Health Services Research **12**, 480 (2012)

[83] Boyd, J.H., Randall, S.M., Brown, A., Max, M., Botes, D., Gillies, M., Ferrante, A.M.: Population data centre profiles: Centre for data linkage. International Journal of Population Data Science **4**(2) (2019)
[84] Boyd, J.H., Randall, S.M., Ferrante, A.M.: Application of privacy-preserving techniques in operational record linkage centres. In: A. Gkoulalas-Divanis, G. Loukides (eds.) Medical Data Privacy Handbook, pp. 267–287. Springer (2015)
[85] Brand, R.: Microdata protection through noise addition. In: Inference Control in Statistical Databases: From Theory to Practice, pp. 97–116. Springer, Berlin, Heidelberg (2002)
[86] Brick, J.M., Williams, D.: Explaining rising nonresponse rates in cross-sectional surveys. The Annals of the American Academy of Political and Social Science **645**(1), 36–59 (2013)
[87] Broder, A.: On the resemblance and containment of documents. In: IEEE Compression and Complexity of Sequences, pp. 21–29. Salerno, Italy (1997)
[88] Broder, A., Mitzenmacher, M., Mitzenmacher, A.: Network applications of Bloom filters: A survey. In: Internet Mathematics (2002)
[89] Brown, A., Borgs, C., Randall, S., Schnell, R.: Evaluating privacy-preserving record linkage using cryptographic long-term keys and multi-bit trees on large medical datasets. BMC Medical Informatics and Decision Making **17**(83), 1–7 (2017)
[90] Brown, A., Borgs, C., Randall, S., Schnell, R.: High quality linkage using multi-bit trees for privacy-preserving blocking. International Journal of Population Data Science **1**(1) (2017)
[91] Bu-Pasha, S.: Cross-border issues under EU data protection law with regards to personal data protection. Information and Communications Technology Law **26**(3), 213–228 (2017)
[92] Bundesamt für Statistik: Das neue Volkszählungssystem. Evaluationsbericht des Bundesrates. `https://www.bfs.admin.ch/bfsstatic/dam/assets/3922064/master`, Bern (2017)
[93] Bundesverfassungsgericht: Decision of the first senate, 4. April— 2006. `https://www.bundesverfassungsgericht.de/SharedDocs/Entscheidungen/DE/2006/04/rs20060404_1bvr051802.html` (2006)
[94] Caldicott, F.: The Information Governance Review. Information: To share or not to share? Department of Health (2013)
[95] Caldicott Committee: Report on the review of patient-identifiable information. Department of Health, 11934 CA Q 1000 1P Dec 97 (1997)
[96] Campbell, K., Deck, D., Krupski, A.: Record linkage software in the public domain: a comparison of Link Plus, The Link King, and a basic deterministic algorithm. Health Informatics Journal **14**(1), 5 (2008)
[97] Carey, P. (ed.): Data Protection: A Practical Guide to UK and EU Law, 5 edn. Oxford University Press, Oxford (2018)

[98] Ceglar, A., Roddick, J.F.: Association mining. ACM Computing Surveys (CSUR) **3**(2), 5 (2006)
[99] Chen, J., Swamidass, S.J., Dou, Y., Bruand, J., Baldi, P.: ChemDB: a public database of small molecules and related chemoinformatics resources. Bioinformatics **21**(22), 4133–4139 (2005)
[100] Chen, M., Mao, S., Liu, Y.: Big data: A survey. Mobile Networks and Applications **19**(2), 171–209 (2014)
[101] Chetty, R.: Time trends in the use of administrative data for empirical research (2012). URL `http://www.rajchetty.com/chettyfiles/admin_data_trends.pdf`. NBER Summer Institute
[102] Chi, L., Zhu, X.: Hashing techniques: A survey and taxonomy. ACM Computing Surveys **50**(1), 11 (2017)
[103] Chi, Y., Hong, J., Jurek, A., Liu, W., O'Reilly, D.: Privacy preserving record linkage in the presence of missing values. Information Systems **71**, 199–210 (2017)
[104] Chiang, Y.H., Doan, A., Naughton, J.F.: Tracking entities in the dynamic world: A fast algorithm for matching temporal records. VLDB Endowment **7**(6), 469–480 (2014)
[105] Chor, B., Kushilevitz, E.: Secret sharing over infinite domains. Journal of Cryptology **6**(2), 87–95 (1993)
[106] Christen, P.: Privacy-preserving data linkage and geocoding: Current approaches and research directions. In: Privacy Aspects of Data Mining, held at IEEE ICDM. Hong Kong (2006)
[107] Christen, P.: Automatic record linkage using seeded nearest neighbour and support vector machine classification. In: ACM Conference on Knowledge Discovery and Data Mining, pp. 151–159. Las Vegas (2008)
[108] Christen, P.: Development and user experiences of an open source data cleaning, deduplication and record linkage system. ACM SIGKDD Explorations **11**(1), 39–48 (2009)
[109] Christen, P.: Data Matching – Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection. Data-Centric Systems and Applications. Springer (2012)
[110] Christen, P.: A survey of indexing techniques for scalable record linkage and deduplication. Transactions on Knowledge and Data Engineering **24**(9), 1537–1555 (2012)
[111] Christen, P.: Preparation of a real temporal voter data set for record linkage and duplicate detection research. Australian National University (2014)
[112] Christen, P.: Data linkage: The big picture. Harvard Data Science Review **1**(2) (2019)
[113] Christen, P., Churches, T., Hegland, M.: Febrl – A parallel open source data linkage system. In: Pacific-Asia Conference on Knowledge Discovery and Data Mining, pp. 638–647. Sydney (2004)

[114] Christen, P., Churches, T., Willmore, A.: A probabilistic geocoding system based on a national address file. In: Australasian Data Mining Conference. Cairns (2004)
[115] Christen, P., Gayler, R., Hawking, D.: Similarity-aware indexing for real-time entity resolution. In: ACM Conference on Information and Knowledge Management, pp. 1565–1568. Hong Kong (2009)
[116] Christen, P., Gayler, R.W.: Adaptive temporal entity resolution on dynamic databases. In: Pacific-Asia Conference on Knowledge Discovery and Data Mining, pp. 558–569. Springer, Gold Coast, Australia (2013)
[117] Christen, P., Gayler, R.W., Tran, K.N., Fisher, J., Vatsalan, D.: Automatic discovery of abnormal values in large textual databases. ACM Journal of Data and Information Quality **7**(1-2), 1–31 (2016)
[118] Christen, P., Pudjijono, A.: Accurate synthetic generation of realistic personal information. In: Pacific-Asia Conference on Knowledge Discovery and Data Mining, pp. 507–514. Bangkok (2009)
[119] Christen, P., Ranbaduge, T., Vatsalan, D., Schnell, R.: Precise and fast cryptanalysis for Bloom filter based privacy-preserving record linkage. Transactions on Knowledge and Data Engineering (2018)
[120] Christen, P., Schnell, R., Vatsalan, D., Ranbaduge, T.: Efficient cryptanalysis of Bloom filters for privacy-preserving record linkage. In: Pacific-Asia Conference on Knowledge Discovery and Data Mining, pp. 628–640. Springer, Jeju, Korea (2017)
[121] Christen, P., Vatsalan, D.: Flexible and extensible generation and corruption of personal data. In: ACM Conference on Information and Knowledge Management, pp. 1165–1168. San Francisco (2013)
[122] Christen, P., Vatsalan, D., Wang, Q.: Efficient entity resolution with adaptive and interactive training data selection. In: IEEE International Conference on Data Mining, pp. 727–732. Atlantic City (2015)
[123] Christen, P., Vidanage, A., Ranbaduge, T., Schnell, R.: Pattern-mining based cryptanalysis of Bloom filters for privacy-preserving record linkage. In: Pacific-Asia Conference on Knowledge Discovery and Data Mining, pp. 628–640. Springer, Melbourne (2018)
[124] Christen, V., Christen, P., Rahm, E.: Informativeness-based active learning for entity resolution. In: Workshop on Data Integration and Applications, held at PKDD/ECML. Würzburg (2019)
[125] Churches, T., Christen, P.: Some methods for blindfolded record linkage. BMC Medical Informatics and Decision Making **4**(9) (2004)
[126] Churches, T., Christen, P., Lim, K., Zhu, J.X.: Preparation of name and address data for record linkage using hidden Markov models. BMC Medical Informatics and Decision Making **2**(9) (2002)
[127] Clark, D.E.: Practical introduction to record linkage for injury research. Injury Prevention **10**, 186–191 (2004)
[128] Clarke, N., Vale, G., Reeves, E.P., Kirwan, M., Smith, D., Farrell, M., Hurl, G., McElvaney, N.G.: GDPR: an impediment to research? Irish Journal of Medical Science **188**(4), 1129–1135 (2019)

[129] Clifton, C., Kantarcioglu, M., Doan, A., Schadow, G., Vaidya, J., Elmagarmid, A., Suciu, D.: Privacy-preserving data integration and sharing. In: Workshop on Research issues in Data Mining and Knowledge Discovery, held at ACM SIGKDD, pp. 19–26. Paris (2004)

[130] Cochinwala, M., Kurien, V., Lalk, G., Shasha, D.: Efficient data reconciliation. Information Sciences **137**(1–4), 1–15 (2001)

[131] Cohen, W.W., Ravikumar, P., Fienberg, S.: A comparison of string distance metrics for name-matching tasks. In: Workshop on Information Integration on the Web, held at IJCAI, pp. 73–78. Acapulco (2003)

[132] Cohen, W.W., Richman, J.: Learning to match and cluster large high-dimensional data sets for data integration. In: ACM Conference on Knowledge Discovery and Data Mining, pp. 475–480. Edmonton (2002)

[133] Colquitt, J.A., Rodell, J.B.: Measuring justice and fairness. In: R. Cropanzano, M. Ambrose (eds.) Oxford Handbook of Justice in the Workplace, pp. 187–202. Oxford University Press (2015)

[134] Committee on Professional Ethics of the American Statistical Association: Ethical guidelines for statistical practice. `https://www.amstat.org/asa/files/pdfs/EthicalGuidelines.pdf` (2018)

[135] Connelly, R., Playford, C.J., Gayle, V., Dibben, C.: The role of administrative data in the big data revolution in social science research. Social Science Research **59**(Supplement C), 1–12 (2016)

[136] Connor, R., Dearle, A.: Querying metric spaces with bit operations. In: International Conference on Similarity Search and Applications, pp. 33–46. Lima, Peru (2018)

[137] Conrad, J.G., Guo, X.S., Schriber, C.P.: Online duplicate detection: Signature reliability in a dynamic retrieval environment. In: ACM Conference on Information and Knowledge Management, pp. 443–452. New Orleans (2003)

[138] Coppersmith, D.: The data encryption standard (DES) and its strength against attacks. IBM Journal of Research and Development **38**(3), 243–250 (1994)

[139] Cormode, G., Muthukrishnan, S.: An improved data stream summary: the count-min sketch and its applications. Journal of Algorithms **55**(1), 58–75 (2005)

[140] Culnane, C., Rubinstein, B.I., Teague, V.: Health data in an open world. arXiv Preprint (2017)

[141] Culnane, C., Rubinstein, B.I., Teague, V.: Vulnerabilities in the use of similarity tables in combination with pseudonymisation to preserve data privacy in the UK Office for National Statistics' privacy-preserving record linkage. arXiv Preprint (2017)

[142] Dalenius, T., Reiss, S.P.: Data-swapping: A technique for disclosure control. Journal of Statistical Planning and Inference **6**(1), 73–85 (1982)

[143] Damerau, F.J.: A technique for computer detection and correction of spelling errors. Communications of the ACM **7**(3), 171–176 (1964)

[144] Das, L.: Role of data in improving care within a health system: A case study of the australian health system. Master's thesis, RAND Corporation (2017)
[145] Datar, M., Immorlica, N., Indyk, P., Mirrokni, V.S.: Locality-sensitive hashing scheme based on p-stable distributions. In: ACM SOCG, pp. 253–262. Brooklyn (2004)
[146] Datta, A., Tschantz, M.C., Datta, A.: Automated experiments on ad privacy settings. In: Privacy Enhancing Technologies, pp. 92–112. Philadelphia (2015)
[147] Day, C.: Record linkage I: Evaluation of commercially available record linkage software for use in NASS. Tech. Rep. STB Research Report STB-95-02, National Agricultural Statistics Service, Washington DC (1995)
[148] Dean, J., Ghemawat, S.: Mapreduce: simplified data processing on large clusters. Communications of the ACM **51**(1), 107–113 (2008)
[149] DeBell, M., Krosnick, J.A., Gera, K., Yeager, D.S., McDonald, M.P.: The turnout gap in surveys: Explanations and solutions. Sociological Methods and Research (2018)
[150] Delnord, M., Szamotulska, K., Hindori-Mohangoo, A., Blondel, B., Macfarlane, A., Dattani, N., Barona, C., Berrut, S., Zile, I., Wood, R., Sakkeus, L., Gissler, M., Zeitlin, J., the Euro-Peristat Scientific Committee: Linking databases on perinatal health: A review of the literature and current practices in Europe. The European Journal of Public Health **26**(3), 422–430 (2016)
[151] Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society: Series B (Methodological) **39**(1), 1–22 (1977)
[152] Desai, T., Ritchie, F., Welpton, R.: Five safes: Designing data access for research. Tech. rep., Department of Accounting, Economics and Finance, Bristol Business School, University of the West of England (2016)
[153] Deutscher Bundestag: Entwurf eines Gesetzes zur Durchführung des Zensus im Jahr 2021. Tech. Rep. Drucksache 19/8693, Bundestag, Berlin (2019)
[154] DH Informatics: Supplementary guidance: Public interest disclosures. UK Department of Health and Social Care (2010)
[155] Diffie, W., Hellman, M.: New directions in cryptography. Transactions on Information Theory **22**(6), 644–654 (1976)
[156] Dillinger, P.C., Manolios, P.: Bloom filters in probabilistic verification. In: International Conference on Formal Methods in Computer-Aided Design, pp. 367–381. Springer (2004)
[157] Dillinger, P.C., Manolios, P.: Fast and accurate bitstate verification for spin. In: International SPIN Workshop on Model Checking of Software, pp. 57–75. Springer (2004)

[158] Doan, A., Halevy, A., Ives, Z.: Principles of Data Integration. Elsevier (2012)
[159] Doidge, J.C., Harron, K.L.: Reflections on modern methods: linkage error bias. International Journal of Epidemiology **48**(6), 2050–2060 (2019)
[160] Domingo-Ferrer, J., Mateo-Sanz, J.: Practical data-oriented microaggregation for statistical disclosure control. Transactions on Knowledge and Data Engineering **14**(1), 189–201 (2002)
[161] Domingo-Ferrer, J., Ricci, S., Soria-Comas, J.: Disclosure risk assessment via record linkage by a maximum-knowledge attacker. In: IEEE Privacy, Security and Trust, pp. 28–35. Izmir (2015)
[162] Domingo-Ferrer, J., Sánchez, D., Soria-Comas, J.: Database Anonymization: Privacy Models, Data Utility, and Microaggregation-based Inter-model Connections. Synthesis Lectures on Information Security, Privacy, and Trust. Morgan and Claypool Publishers (2016)
[163] Domingo-Ferrer, J., Sebé, F., Castella-Roca, J.: On the security of noise addition for privacy in statistical databases. In: Privacy in Statistical Databases, pp. 149–161. Barcelona (2004)
[164] Domingo-Ferrer, J., Torra, V.: Disclosure risk assessment in statistical microdata protection via advanced record linkage. Statistics and Computing **13**(4), 343–354 (2003)
[165] Dong, B., Liu, R., Wang, W.H.: Prada: Privacy-preserving data-deduplication-as-a-service. In: ACM Conference on Information and Knowledge Management, pp. 1559–1568. Shanghai (2014)
[166] Dong, C., Chen, L., Wen, Z.: When private set intersection meets big data: an efficient and scalable protocol. In: ACM Conference on Computer and Communications Security, pp. 789–800. Berlin (2013)
[167] Dong, X.L., Halevy, A., Madhavan, J.: Reference reconciliation in complex information spaces. In: ACM Conference on Management of Data, pp. 85–96. Baltimore (2005)
[168] Dong, X.L., Srivastava, D.: Big Data Integration. Synthesis Lectures on Data Management. Morgan and Claypool Publishers (2015)
[169] Dove, E.S., Phillips, M.: Privacy law, data sharing policies, and medical data: A comparative perspective. In: A. Gkoulalas-Divanis, G. Loukides (eds.) Medical Data Privacy Handbook, pp. 639–678. Springer (2015)
[170] Draisbach, U., Christen, P., Naumann, F.: Transforming pairwise duplicates to entity clusters for high-quality duplicate detection. Journal of Data and Information Quality **12**(1), 1–30 (2019)
[171] Draisbach, U., Naumann, F., Szott, S., Wonneberg, O.: Adaptive windows for duplicate detection. In: IEEE International Conference on Data Engineering, pp. 1073–1083. Washington, DC (2012)
[172] Du, W., Atallah, M., Kerschbaum, F.: Protocols for secure remote database access with approximate matching. In: ACM Workshop on Security and Privacy in E-Commerce. Athens (2000)

[173] Duncan, G., Elliot, M., Salazar-González, J.J.: Statistical Confidentiality: Principles and Practice. Springer, New York (2011)
[174] Duncan, G., Lambert, D.: The risk of disclosure for microdata. Journal of Business & Economic Statistics **7**(2), 207–217 (1989)
[175] Dunn, H.: Record linkage. American Journal of Public Health **36**(12), 1412 (1946)
[176] Durham, E., Kantarcioglu, M., Xue, Y., Toth, C., Kuzu, M., Malin, B.: Composite Bloom filters for secure record linkage. Transactions on Knowledge and Data Engineering **26**(12), 2956–2968 (2014)
[177] Durham, E., Xue, Y., Kantarcioglu, M., Malin, B.: Quantifying the correctness, computational complexity, and security of privacy-preserving string comparators for record linkage. Information Fusion **13**(4), 245–259 (2012)
[178] Durham, E.A.: A framework for accurate, efficient private record linkage. Ph.D. thesis, Faculty of the Graduate School of Vanderbilt University, Nashville, TN (2012)
[179] Durstenfeld, R.: Algorithm 235: Random permutation. Communications of the ACM **7**(7), 420 (1964)
[180] Dusserre, L., Quantin, C., Bouzelat, H.: A one way public key cryptosystem for the linkage of nominal files in epidemiological studies. Medinfo **8**, 644–647 (1995)
[181] Dwork, C.: Differential privacy. In: International Colloquium on Automata, Languages and Programming, pp. 1–12. Venice (2006)
[182] Dwork, C.: Differential privacy: A survey of results. In: Theory and Applications of Models of Computation, pp. 1–19. Xi'an, China (2008)
[183] Dwork, C., Roth, A., et al.: The algorithmic foundations of differential privacy. Foundations and Trends in Theoretical Computer Science **9**(3–4), 211–407 (2014)
[184] Efthymiou, V., Papadakis, G., Papastefanatos, G., Stefanidis, K., Palpanas, T.: Parallel meta-blocking for scaling entity resolution over big heterogeneous data. Information Systems **65**, 137–157 (2017)
[185] El Emam, K.: Guide to the De-Identification of Personal Health Information. CRC Press (2013)
[186] Elfeky, M.G., Verykios, V.S., Elmagarmid, A.K.: TAILOR: A record linkage toolbox. In: IEEE International Conference on Data Engineering, pp. 17–28. San Jose (2002)
[187] Elliot, M., Mackey, E., O'Hara, K., Tudor, C.: The Anonymisation Decision-making Framework. UKAN Manchester (2016)
[188] Elliot, M., O'Hara, K., Raab, C., O'Keefe, C.M., Mackey, E., Dibben, C., Gowans, H., Purdam, K., McCullagh, K.: Functional anonymisation: Personal data and the data environment. Computer Law and Security Review **34**(2), 204–221 (2018)
[189] Emmert-Streib, F., Dehmer, M., Shi, Y.: Fifty years of graph matching, network alignment and network comparison. Information Sciences **346**, 180–197 (2016)

[190] Erlingsson, Ú., Pihur, V., Korolova, A.: Rappor: Randomized aggregatable privacy-preserving ordinal response. In: ACM SIGSAC, pp. 1054–1067. Scottsdale, Arizona (2014)
[191] Esayas, S.: The role of anonymisation and pseudonymisation under the EU data privacy rules: beyond the 'all or nothing' approach. European Journal of Law and Technology **6**(2) (2015)
[192] Etheridge, Y.: PKI (public key infrastructure) – how and why it works. Health Management Technology **22**(1), 20 (2001)
[193] Etienne, B., Cheatham, M., Grzebala, P.: An analysis of blocking methods for private record linkage. In: AAAI Fall Symposium Series. Arlington, Virginia (2016)
[194] European Commission: Flash Eurobarometer 443: e-Privacy. doi:10.2759/249540 (2016)
[195] European Parliament: Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the Protection of Natural Persons with Regard to the Processing of Personal Data and on the Free Movement of such Data, and Repealing Directive 95/46/EC (General Data Protection Regulation). https://eur-lex.europa.eu/eli/reg/2016/679/oj (2016)
[196] European Statistical System Committee: European Statistics Code of Practice. Luxembourg (2018)
[197] European Union Agency for Fundamental Rights: Handbook on European Data Protection Law: 2018 Edition. Publications Office of the European Union, Luxembourg (2018)
[198] Fair, M.: Generalized record linkage system–Statistics Canada's record linkage software. Austrian Journal of Statistics **33**(1&2), 37–53 (2004)
[199] Fan, L., Cao, P., Almeida, J., Broder, A.Z.: Summary cache: a scalable wide-area web cache sharing protocol. IEEE/ACM Transactions on Networking **8**(3), 281–293 (2000)
[200] Farrow, J.M.: Comparing geospatial distance without revealing location. Presentation given at the International Health Data Linkage Conference, Vancouver (2014)
[201] Farrow, J.M.: Using graph databases to manage linked data. In: K. Harron, H. Goldstein, C. Dibben (eds.) Methodological Developments in Data Linkage, pp. 125–169. John Wiley & Sons (2015)
[202] Farrow, J.M.: Method and system for comparative data analysis (2017). US Patent App. 15/305,335
[203] Fellegi, I.P., Holt, D.: A systematic approach to automatic edit and imputation. Journal of the American Statistical Association **71**(353), 17–35 (1976)
[204] Fellegi, I.P., Sunter, A.B.: A theory for record linkage. Journal of the American Statistical Association **64**(328), 1183–1210 (1969)
[205] Ferguson, J., Hannigan, A., Stack, A.: A new computationally efficient algorithm for record linkage with field dependency and missing data

imputation. International Journal of Medical Informatics **109**, 70–75 (2018)
[206] Ferrante, A.M., Boyd, J.H.: A transparent and transportable methodology for evaluating data linkage software. Journal of Biomedical Informatics **45**(1), 165–172 (2012)
[207] Fickermann, D., Doll, J.: Potential und Technik der Verknüpfung von Befragungsdaten mit schulstatistischen Individualdaten und Leistungsdaten im Projekt EIBISCH. DDS - Die Deutsche Schule **107**(4), 365–374 (2015)
[208] Fienberg, S.E.: Confidentiality and disclosure limitation. Encyclopedia of Social Measurement **1**, 463–69 (2005)
[209] Fienberg, S.E.: Homeland insecurity: Data mining, privacy, disclosure limitation, and the hunt for terrorists. In: H. Chen, E. Reid, J. Sinai, A. Silke, B. Ganor (eds.) Terrorism Informatics: Knowledge Management and Data Mining for Homeland Security, pp. 197–218. Springer (2008)
[210] Fienberg, S.E., Steele, R.J.: Disclosure limitation using perturbation and related methods for categorical data. Journal of Official Statistics **14**(4), 485 (1998)
[211] Finance, B., Medjdoub, S., Pucheral, P.: Privacy of medical records: From law principles to practice. In: IEEE International Symposium on Computer Based Medical Systems, pp. 220–225. Dublin (2005)
[212] Fisher, J., Christen, P., Wang, Q., Rahm, E.: A clustering-based framework to control block sizes for entity resolution. In: ACM Conference on Knowledge Discovery and Data Mining, pp. 279–288. Sydney (2015)
[213] Flack, F., Adams, C., Allen, J.: Authorising the release of data without consent for health research: The role of data custodians and HRECs in Australia. Journal of Law and Medicine **26**(3), 655–680 (2019)
[214] Flack, F., Kemp-Casey, A., Wray, N.: Using linked administrative data in clinical trials: A guide for clinical trialists and researchers. Australian Clinical Trials Alliance (2019)
[215] Ford, D.V., Jones, K.H., Verplancke, J.P., Lyons, R.A., John, G., Brown, G., Brooks, C.J., Thompson, S., Bodger, O., Couch, T., et al.: The SAIL databank: building a national architecture for e-health research and evaluation. BMC Health Services Research **9**(1), 157 (2009)
[216] Fortini, M., Liseo, B., Nuccitelli, A., Scanu, M.: On Bayesian record linkage. Research in Official Statistics **4**(1), 185–198 (2001)
[217] Fox, W.R., Lasker, G.W.: The distribution of surname frequencies. International Statistical Review pp. 81–87 (1983)
[218] Franke, M., Gladbach, M., Sehili, Z., Rohde, F., Rahm, E.: ScaDS research on scalable privacy-preserving record linkage. Datenbank-Spektrum **19**(1), 31–40 (2019)
[219] Franke, M., Sehili, Z., Rahm, E.: PRIMAT: a toolbox for fast privacy-preserving matching. VLDB Endowment **12**(12), 1826–1829 (2019)

[220] Fu, Z., Zhou, J., Christen, P., Boot, M.: Multiple instance learning for group record linkage. In: Pacific-Asia Conference on Knowledge Discovery and Data Mining, pp. 171–182. Kuala Lumpur (2012)
[221] Fuller, W.: Masking procedures for microdata disclosure. Journal of Official Statistics **9**(2), 383–406 (1993)
[222] Galbraith, S.D.: Mathematics of Public Key Cryptography. Cambridge University Press (2012)
[223] Ganta, S.R., Kasiviswanathan, S.P., Smith, A.: Composition attacks and auxiliary information in data privacy. In: ACM Conference on Knowledge Discovery and Data Mining, pp. 265–273. Las Vegas (2008)
[224] Garfinkel, S.L.: De-identification of personal information. Tech. Rep. NIST IR 8053, National Institute of Standards and Technology (2015)
[225] Gilbert, R., Lafferty, R., Hagger-Johnson, G., Harron, K., Zhang, L.C., Smith, P., Dibben, C., Goldstein, H.: GUILD: Guidance for information about linking data sets. Journal of Public Health **40**(1), 191–198 (2017)
[226] Gill, L.: Methods for automatic record matching and linking and their use in national statistics. Tech. Rep. Methodology Series, no. 25, National Statistics, London (2001)
[227] Gionis, A., Indyk, P., Motwani, R., et al.: Similarity search in high dimensions via hashing. In: Conference on Very Large Data Bases, pp. 518–529. Edinburgh (1999)
[228] Goldberg, A., Borthwick, A.: The Choicemaker 2 record matching system. ChoiceMaker Technologies, Inc. (2004)
[229] Goldreich, O.: Secure multi-party computation. Tech. rep., Department of Computer Science and Applied Mathematics, Weizmann Institute of Science, Israel (2002)
[230] Goldreich, O.: Foundations of Cryptography: Volume 2, Basic Applications. Cambridge University Press (2009)
[231] Goldreich, O., Micali, S., Wigderson, A.: How to play any mental game. In: ACM Symposium on Theory of computing, pp. 218–229 (1987)
[232] Goldstein, H., Harron, K.: Record linkage: a missing data problem. In: K. Harron, H. Goldstein, C. Dibben (eds.) Methodological Developments in Data Linkage, pp. 110–124. John Wiley & Sons (2015)
[233] Goldstein, H., Harron, K., Wade, A.: The analysis of record-linked data using multiple imputation with data value priors. Statistics in Medicine **31**(28), 3481–3493 (2012)
[234] Goldwasser, S., Micali, S.: Probabilistic encryption. Journal of computer and system sciences **28**(2), 270–299 (1984)
[235] Gomatam, S., Larsen, M.D.: Record linkage and counterterrorism. Chance **17**(1), 25–29 (2004)
[236] Gouweleeuw, J.M., Kooiman, P., Willenborg, L.C.R.J., de Wolf, P.P.: Post randomisation for statistical disclosure control: Theory and implementation. Journal of Official Statistics **14**(4), 463–478 (1998)
[237] Goyal, V., Mohassel, P., Smith, A.: Efficient two party and multi party computation against covert adversaries. In: Annual International Con-

ference on the Theory and Applications of Cryptographic Techniques, pp. 289–306. Istanbul (2008)
[238] Grama, J.L.: Legal Issues in Information Security, 2 edn. Jones and Batlett Learning, Burlinton (2015)
[239] Grannis, S.J., Overhage, J.M., McDonald, C.J.: Analysis of identifier performance using a deterministic linkage algorithm. In: AMIA Annual Symposium Proceedings, p. 305. American Medical Informatics Association (2002)
[240] Gropp, W., Thakur, R., Lusk, E.: Using MPI-2: Advanced Features of the Message Passing Interface. MIT press (1999)
[241] Groves, R.M.: Survey Errors and Survey Costs. Wiley Series in Survey Methodology. John Wiley and Sons (2004)
[242] Groves, R.M., Fowler, F.J., Couper, M.P., Lebkowski, J.M., Singer, E., Tourangeau, R.: Survey Methodology, 2 edn. Wiley, Hoboken (2009)
[243] Gu, L., Baxter, R.: Decision models for record linkage. In: Selected Papers from AusDM, pp. 146–160. Springer (2006)
[244] Gu, L., Baxter, R., Vickers, D., Rainsford, C.: Record linkage: Current practice and future directions. CSIRO Mathematical and Information Sciences Technical Report **3**, 83 (2003)
[245] Guesdon, M., Benzenine, E., Gadouche, K., Quantin, C.: Securizing data linkage in French public statistics. BMC Medical Informatics and Decision Making **16**(1), 129 (2016)
[246] Guisado-Gámez, J., Prat-Pérez, A., Nin, J., Muntés-Mulero, V., Larriba-Pey, J.L.: Parallelizing record linkage for disclosure risk assessment. In: Privacy in Statistical Databases, pp. 190–202. Istanbul (2008)
[247] Gymrek, M., McGuire, A.L., Golan, D., Halperin, E., Erlich, Y.: Identifying personal genomes by surname inference. Science **339**(6117), 321–324 (2013)
[248] Hafner, H.P., Lenz, R., Ritchie, F.: User-focused threat identification for anonymised microdata. Statistical Journal of the IAOS **35**(4), 703–713 (2019)
[249] Hagger-Johnson, G., Harron, K., Goldstein, H., Aldridge, R., Gilbert, R.: Probabilistic linkage to enhance deterministic algorithms and reduce data linkage errors in hospital administrative data. Journal of Innovation in Health Informatics **24**(2), 891 (2017)
[250] Hajian, S., Bonchi, F., Castillo, C.: Algorithmic bias: From discrimination discovery to fairness-aware data mining. In: ACM Conference on Knowledge Discovery and Data Mining, pp. 2125–2126. San Francisco (2016)
[251] Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.: The WEKA data mining software: an update. ACM SIGKDD Explorations **11**(1), 10–18 (2009)
[252] Hall, P.A., Dowling, G.R.: Approximate string matching. ACM Computing Surveys **12**(4), 381–402 (1980)

[253] Hall, R., Fienberg, S.: Privacy-preserving record linkage. In: Privacy in Statistical Databases, pp. 269–283. Corfu, Greece (2010)
[254] Han, J., Kamber, M., Pei, J.: Data Mining: Concepts and Techniques, 3 edn. Morgan Kaufmann (2011)
[255] Han, S., Shen, D., Nie, T., Kou, Y., Yu, G.: Private blocking technique for multi-party privacy-preserving record linkage. Data Science and Engineering **2**(2), 187–196 (2017)
[256] Hand, D.J.: Classifier technology and the illusion of progress. Statistical Science **21**(1), 1–14 (2006)
[257] Hand, D.J.: Assessing the performance of classification methods. International Statistical Review **80**(3), 400–414 (2012)
[258] Hand, D.J.: Aspects of data ethics in a changing world: where are we now? Big data **6**(3), 176–190 (2018)
[259] Hand, D.J.: Dark Data: Why What You Don't Know Matters. Princeton University Press (2020)
[260] Hand, D.J., Christen, P.: A note on using the f-measure for evaluating record linkage algorithms. Statistics and Computing **28**(3), 539–547 (2018)
[261] Harada, M., Sato, S., Kazama, K.: Finding authoritative people from the web. In: ACM/IEEE Joint Conference on Digital Libraries, pp. 306–313. Tucson (2004)
[262] Harper, G.: A study of the use of linked routinely collected administrative data at the local level to count and profile populations. Ph.D. thesis, City, University of London (2017)
[263] Harper, G., Mayhew, L.: Using administrative data to count local populations. Applied Spatial Analysis and Policy **5**(2), 97–122 (2012)
[264] Harron, K., Dibben, C., Boyd, J., Hjern, A., Azimaee, M., Barreto, M.L., Goldstein, H.: Challenges in administrative data linkage for research. Big Data and Society **4**(2), 1–12 (2017)
[265] Harron, K., Gilbert, R., Cromwell, D., van der Meulen, J.: Linking data for mothers and babies in de-identified electronic health data. PLOS One **11**(10), 1–18 (2016)
[266] Harron, K., Goldstein, H., Dibben, C.: Methodological Developments in Data Linkage. John Wiley and Sons (2015)
[267] Harron, K., Wade, A., Gilbert, R., Muller-Pebody, B., Goldstein, H.: Evaluating bias due to data linkage error in electronic healthcare records. BMC Medical Research Methodology **14**(1), 36 (2014)
[268] Harron, K.L., Doidge, J.C., Knight, H.E., Gilbert, R.E., Goldstein, H., Cromwell, D.A., van der Meulen, J.H.: A guide to evaluating linkage quality for the analysis of linked data. International Journal of Epidemiology **46**(5), 1699–1710 (2017)
[269] Hassanzadeh, O., Chiang, F., Lee, H.C., Miller, R.J.: Framework for evaluating clustering algorithms in duplicate detection. VLDB Endowment **2**(1), 1282–1293 (2009)

[270] He, X., Machanavajjhala, A., Flynn, C., Srivastava, D.: Composing differential privacy and secure computation: A case study on scaling private record linkage. In: ACM Conference on Computer and Communications Security, pp. 1389–1406. Dallas (2017)
[271] van Herk-Sukel, M.P., Lemmens, V.E., van de Poll-Franse, L.V., Herings, R.M., Coebergh, J.W.W.: Record linkage for pharmacoepidemiological studies in cancer patients. Pharmacoepidemiology and Drug Safety **21**(1), 94–103 (2012)
[272] Hernandez, M.A., Stolfo, S.J.: The merge/purge problem for large databases. In: ACM Conference on Management of Data, pp. 127–138. San Jose (1995)
[273] Hernandez, M.A., Stolfo, S.J.: Real-world data is dirty: Data cleansing and the merge/purge problem. Data Mining and Knowledge Discovery **2**(1), 9–37 (1998)
[274] Herzog, T., Scheuren, F., Winkler, W.: Data Quality and Record Linkage Techniques. Springer Verlag (2007)
[275] Hill, K.: The secretive company that might end privacy as we know it. New York Times (2020)
[276] Hillestad, R., Bigelow, J.H., Chaudhry, B., Dreyer, P., Greenberg, M.D., Meili, R., Ridgely, M.S., Rothenberg, J., Taylor, R.: Identity crisis? An examination of the costs and benefits of a unique patient identifier for the US health care system. RAND Corporation (2008)
[277] Hintze, M., El Emam, K.: Comparing the benefits of pseudonymisation and anonymisation under the GDPR. Journal of Data Protection & Privacy **2**(2), 145–158 (2018)
[278] Hippisley-Cox, J.: Validity and completeness of the NHS Number in primary and secondary care: electronic data in England 1991-2013. Tech. rep., University of Nottingham (2013). URL `http://eprints.nottingham.ac.uk/3153`
[279] Hochbaum, D.S., Shmoys, D.B.: A best possible heuristic for the k-center problem. Mathematics of Operations Research **10**(2), 180–184 (1985)
[280] Hodges, S., Eitelhuber, T., Merchant, A., Alan, J.: Population data centre profile – the Western Australian Data Linkage Branch. International Journal of Population Data Science **4**(2) (2019)
[281] Hodgins, S., Janson, C.G.: Criminality and Violence among the Mentally Disordered: the Stockholm Metropolitan Project. Cambridge University Press, Cambridge, UK (2002)
[282] Holman, C.D.J., Bass, J.A., Rosman, D.L., Smith, M.B., Semmens, J.B., Glasson, E.J., Brook, E.L., Trutwein, B., Rouse, I.L., Watson, C.R., et al.: A decade of data linkage in Western Australia: strategic design, applications and benefits of the wa data linkage system. Australian Health Review **32**(4), 766–777 (2008)
[283] Holmes, D., McCabe, C.M.: Improving precision and recall for Soundex retrieval. In: IEEE ITCC. Las Vegas (2002)

[284] Honer, M.: BVerfG zu Recht auf Vergessen I und II Teil 2. Legal Tribune Online, https://www.lto.de/persistent/a_id/39109 (2019)
[285] Hoptroff, R., Mosquera, L., El Emam, K.: Practical Synthetic Data Generation. O'Reilly Media (2020)
[286] van den Hout, A., Elamir, E.A.H.: Statistical disclosure control using post randomisation: Variants and measures for disclosure risk. Journal of Official Statistics **22**(4), 711–731 (2006)
[287] Hu, Y., Wang, Q., Vatsalan, D., Christen, P.: Improving temporal record linkage using regression classification. In: Pacific-Asia Conference on Knowledge Discovery and Data Mining, pp. 561–573 (2017)
[288] Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Schulte Nordholt, E., Spicer, K., de Wolf, P.: Statistical Disclosure Control. Wiley, Chichester (2012)
[289] Inan, A., Kantarcioglu, M., Bertino, E., Scannapieco, M.: A hybrid approach to private record linkage. In: IEEE International Conference on Data Engineering, pp. 496–505. Cancun (2008)
[290] Inan, A., Kantarcioglu, M., Ghinita, G., Bertino, E.: Private record matching using differential privacy. In: Conference on Extending Database Technology, pp. 123–134. Lausanne (2010)
[291] Indyk, P., Motwani, R.: Approximate nearest neighbors: towards removing the curse of dimensionality. In: Annual ACM Symposium on Theory of Computing, pp. 604–613. ACM (1998)
[292] Information Commissioner's Office: Data sharing code of practice (2011)
[293] Information Commissioner's Office: Anonymisation: Managing data protection risk code of practice (2012)
[294] Institute for Social and Economic Research: Linked understanding society – national pupil database wave 1 linkage user manual. UK Data Archive Study Number 7642, University of Essex, Colchester (2015)
[295] Institute of Medicine: Health Services Research: Opportunities for an Expanding Field of Inquiry. The National Academies Press, Washington, DC (1994)
[296] International Organization for Standardization: ISO 8601:2004 – Data elements and interchange formats – Information interchange – Representation of dates and times (2019)
[297] Izakian, H.: Privacy preserving record linkage meets record linkage using unencrypted data. International Journal of Population Data Science **3**(4), 61 (2018)
[298] Jacobs, J.A., Boulis, A., Messikomer, C.: The movement of physicians between specialties. Research in Social Stratification and Mobility **18**(1), 63–95 (2001)
[299] Jakobsen, T.: A fast method for cryptanalysis of substitution ciphers. Cryptologia **19**(3), 265–274 (1995)

[300] Japec, L., Kreuter, F., Berg, M., Biemer, P., Decker, P., Lampe, C., Lane, J., O'Neil, C., Usher, A.: Big data in survey research: AAPOR task force report. Public Opinion Quarterly **79**(4), 839–880 (2015)
[301] Jaro, M.A.: Advances in record-linkage methodology a applied to matching the 1985 Census of Tampa, Florida. Journal of the American Statistical Association **84**, 414–420 (1989)
[302] Jentzsch, N.: The Economics and Regulation of Financial Privacy: An International Comparison of Credit Reporting Systems. Physica-Verlag, Heidelberg (2006)
[303] Jiang, W., Clifton, C.: AC-framework for privacy-preserving collaboration. In: SIAM International Conference on Data Mining, pp. 47–56. Minneapolis (2007)
[304] Jiang, W., Clifton, C., Kantarcıoğlu, M.: Transforming semi-honest protocols to ensure accountability. Data and Knowledge Engineering **65**(1), 57–74 (2008)
[305] Jin, L., Li, C., Mehrotra, S.: Efficient record linkage in large data sets. In: Conference on Database Systems for Advanced Applications, pp. 137–146. Tokyo (2003)
[306] Johnson, W.: Understanding the genetics of intelligence: Can height help? Can corn oil? Current Directions in Psychological Science **19**(3), 177–182 (2010)
[307] Johnston, D.: Random Number Generators—Principles and Practices: A Guide for Engineers and Programmers. Walter de Gruyter GmbH & Co KG (2018)
[308] Jokinen, P., Tarhio, J., Ukkonen, E.: A comparison of approximate string matching algorithms. Software – Practice and Experience **26**(12), 1439–1458 (1996)
[309] Jonas, J., Harper, J.: Effective counterterrorism and the limited role of predictive data mining. Policy Analysis (584) (2006)
[310] Jones, K.H., Ford, D.V.: Population data science: advancing the safe use of population data for public benefit. Epidemiology and Health **40** (2018)
[311] Jones, K.H., Ford, D.V.: Privacy, confidentiality and practicalities in data linkage. National Statistician's Quality Review into Privacy and Data Confidentiality Methods, Government Statistical Service (2018)
[312] Jones, K.H., Ford, D.V., Jones, C., Dsilva, R., Thompson, S., Brooks, C.J., Heaven, M.L., Thayer, D.S., McNerney, C.L., Lyons, R.A.: A case study of the secure anonymous information linkage (SAIL) Gateway: a privacy-protecting remote access system for health-related research and evaluation. Journal of Biomedical Informatics **50**, 196–204 (2014)
[313] Jones, K.H., Heys, S., Tingay, K.S., Jackson, P., Dibben, C.: The Good, the Bad, the Clunky. International Journal of Population Data Science **4**(1) (2019)

[314] Jurczyk, P., Lu, J., Xiong, L., Cragan, J., Correa, A.: FRIL: A tool for comparative record linkage. In: AMIA Annual Symposium Proceedings, p. 440. American Medical Informatics Association (2008)
[315] Kalashnikov, D., Mehrotra, S.: Domain-independent data cleaning via analysis of entity-relationship graph. ACM Transactions on Database Systems **31**(2), 716–767 (2006)
[316] Kalkman, S., Mostert, M., Gerlinger, C., van Delden, J.J.M., van Thiel, G.J.M.W.: Responsible data sharing in international health research: a systematic review of principles and norms. BMC Medical Ethics **20** (2019)
[317] Kalton, G.: Designs for surveys over time. In: D. Pfeffermann, C.R. Rao (eds.) Handbook of Statistics: Sample Surveys, vol. 29A, pp. 89–108. Elsevier, Amsterdam (2009)
[318] Kantarcioglu, M., Inan, A., Jiang, W., Malin, B.: Formal anonymity models for efficient privacy-preserving joins. Data and Knowledge Engineering **68**(11), 1206–1223 (2009)
[319] Kantarcioglu, M., Jiang, W., Malin, B.: A privacy-preserving framework for integrating person-specific databases. In: Privacy in Statistical Databases, pp. 298–314. Istanbul (2008)
[320] Karakasidis, A., Koloniari, G., Verykios, V.S.: Scalable blocking for privacy preserving record linkage. In: ACM Conference on Knowledge Discovery and Data Mining, pp. 527–536. Sydney (2015)
[321] Karakasidis, A., Verykios, V.S.: Privacy preserving record linkage using phonetic codes. In: Fourth Balkan Conference in Informatics, pp. 101–106. Thessaloniki (2009)
[322] Karakasidis, A., Verykios, V.S.: Secure blocking + secure matching = secure record linkage. Journal of Computing Science and Engineering **5**(3) (2011)
[323] Karakasidis, A., Verykios, V.S.: Reference table based k-anonymous private blocking. In: ACM Symposium on Applied Computing, pp. 859–864. Trento, Italy (2012)
[324] Karakasidis, A., Verykios, V.S.: A sorted neighborhood approach to multidimensional privacy preserving blocking. In: IEEE ICDM Workshops, pp. 937–944. Brussels (2012)
[325] Karakasidis, A., Verykios, V.S., Christen, P.: Fake injection strategies for private phonetic matching. In: International Workshop on Data Privacy Management. Leuven, Belgium (2011)
[326] Karapiperis, D., Gkoulalas-Divanis, A., Verykios, V.S.: LSHDB: a parallel and distributed engine for record linkage and similarity search. In: IEEE ICDM Workshops, pp. 1–4. Barcelona (2016)
[327] Karapiperis, D., Gkoulalas-Divanis, A., Verykios, V.S.: Distance-aware encoding of numerical values for privacy-preserving record linkage. In: IEEE International Conference on Data Engineering, pp. 135–138. San Diego (2017)

[328] Karapiperis, D., Gkoulalas-Divanis, A., Verykios, V.S.: FEDERAL: A framework for distance-aware privacy-preserving record linkage. Transactions on Knowledge and Data Engineering **30**(2), 292–304 (2017)
[329] Karapiperis, D., Gkoulalas-Divanis, A., Verykios, V.S.: FEMRL: A framework for large-scale privacy-preserving linkage of patients' electronic health records. In: IEEE International Smart Cities Conference, pp. 1–8. Kansas City (2018)
[330] Karapiperis, D., Gkoulalas-Divanis, A., Verykios, V.S.: Summarizing and linking electronic health records. Distributed and Parallel Databases pp. 1–40 (2019)
[331] Karapiperis, D., Vatsalan, D., Verykios, V.S., Christen, P.: Large-scale multi-party counting set intersection using a space efficient global synopsis. In: Conference on Database Systems for Advanced Applications, pp. 329–345. Hanoi (2015)
[332] Karapiperis, D., Verykios, V.S.: A distributed near-optimal LSH-based framework for privacy-preserving record linkage. Computer Science and Information Systems **11**(2), 745–763 (2014)
[333] Karapiperis, D., Verykios, V.S.: An LSH-based blocking approach with a homomorphic matching technique for privacy-preserving record linkage. Transactions on Knowledge and Data Engineering **27**(4), 909 – 921 (2015)
[334] Karapiperis, D., Verykios, V.S.: A fast and efficient Hamming LSH-based scheme for accurate linkage. Knowledge and Information Systems **49**(3), 861–884 (2016)
[335] Karmel, R.: Data linkage protocols using a statistical linkage key. Australian Institute of Health and Welfare (CS1) (2005)
[336] Karnin, E., Greene, J., Hellman, M.: On secret sharing systems. Transactions on Information Theory **29**(1), 35–41 (1983)
[337] Katikireddi, S.V., Whitley, E., Lewsey, J., Gray, L., Leyland, A.H.: Socioeconomic status as an effect modifier of alcohol consumption and harm: analysis of linked cohort data. The Lancet Public Health **2**(6), e267–e276 (2017)
[338] Katz, A., Enns, J., Smith, M., Burchill, C., Turner, K.: Population data centre profiles: Centre for data linkage. Population Data Centre Profile: The Manitoba Centre for Health Policy **4**(2) (2019)
[339] Kawai, H., Garcia-Molina, H., Benjelloun, O., Menestrina, D., Whang, E., Gong, H.: P-Swoosh: Parallel algorithm for generic entity resolution. Tech. Rep. 2006-19, Department of Computer Science, Stanford University (2006)
[340] Kelman, C.W., Bass, J., Holman, D.: Research use of linked health data – A best practice protocol. Aust NZ Journal of Public Health **26**, 251–255 (2002)
[341] Kendrick, S.: The development of record linkage in Scotland: The responsive application of probabilistic matching. In: Record Linkage Techniques, pp. 319–332. Arlington (1997)

[342] Kerckhoffs, A.: Military cryptography. French Journal of Military Science (1883)
[343] Keskustalo, H., Pirkola, A., Visala, K., Leppanen, E., Jarvelin, K.: Non-adjacent digrams improve matching of cross-lingual spelling variants. In: String Processing and Information Retrieval, pp. 252–265. Manaus, Brazil (2003)
[344] Kessler, G.C.: An overview of cryptography. Handbook on Local Area Networks, Auerbach (1998)
[345] Kho, A.N., Cashy, J.P., Jackson, K.L., Pah, A.R., Goel, S., Boehnke, J., Humphries, J.E., Kominers, S.D., Hota, B.N., Sims, S.A., et al.: Design and implementation of a privacy preserving electronic health record linkage tool in Chicago. Journal of the American Medical Informatics Association **22**(5), 1072–1080 (2015)
[346] Kifer, D., Machanavajjhala, A.: Pufferfish: A framework for mathematical privacy definitions. ACM Transactions on Database Systems **39**(1), 1–36 (2014)
[347] Kim, D., Solomon, M.G.: Fundamentals of Information Systems Security, 3 edn. Jones and Bartlett Learning, Burlington (2018)
[348] Kim, H., Lee, D.: Parallel linkage. In: ACM Conference on Information and Knowledge Management, pp. 283–292. Lisboa (2007)
[349] Kim, H., Lee, D.: HARRA: fast iterative hashed record linkage for large-scale data collections. In: Conference on Extending Database Technology, pp. 525–536. Lausanne (2010)
[350] Kim, J.J.: A method for limiting disclosure in microdata based on random noise and transformation. In: Proceedings of the Section on Survey Research Methods, pp. 303–308. American Statistical Association (1986)
[351] Kirielle, N., Christen, P., Ranbaduge, T.: Outlier detection based accurate geocoding of historical addresses. In: Australasian Data Mining Conference, pp. 41–53. Adelaide (2019)
[352] Kirsch, A., Mitzenmacher, M.: Less hashing, same performance: building a better Bloom filter. In: European Symposium on Algorithms, pp. 456–467. Zürich (2006)
[353] Kirsten, T., Kolb, L., Hartung, M., Gross, A., Köpcke, H., Rahm, E.: Data partitioning for parallel entity matching. VLDB Endowment **3**(2) (2010)
[354] Klingwort, J., Buelens, B., Schnell, R.: Capture-recapture techniques for transport survey estimate adjustment using permanently installed highway-sensors. Social Science Computer Review (2019)
[355] Knuth, D.E.: The Art of Computer Programming: Seminumerical Algorithms, vol. 2. Addison Wesley Publishing Company (1969)
[356] Knuth, D.E.: Big omicron and big omega and big theta. ACM SIGACT News **8**(2), 18–24 (1976)
[357] Knuth, D.E.: Efficient balanced codes. Transactions on Information Theory **32**(1), 51–53 (1986)

[358] Kolb, L., Thor, A., Rahm, E.: Dedoop: Efficient deduplication with Hadoop. VLDB Endowment **5**(12), 1878–1881 (2012)
[359] Kong, C., Gao, M., Xu, C., Qian, W., Zhou, A.: Entity matching across multiple heterogeneous data sources. In: Conference on Database Systems for Advanced Applications, pp. 133–146. Dallas (2016)
[360] Köpcke, H., Rahm, E.: Frameworks for entity matching: A comparison. Data and Knowledge Engineering **69**(2), 197–210 (2010)
[361] Köpcke, H., Thor, A., Rahm, E.: Evaluation of entity resolution approaches on real-world match problems. VLDB Endowment **3**(1-2), 484–493 (2010)
[362] Körner, T., Krause, A., Ramsauer, K., Ullmann, P.: Registernutzung in Zensus und Bevölkerungsstatistik in Österreich und der Schweiz. Destatis, Wiesbaden (2017)
[363] Kosa, T.A., El-Khatib, K., Marsh, S.: Measuring privacy. Journal of Internet Services and Information Security **1**(4), 60–73 (2011)
[364] Krawczyk, H., Bellare, M., Canetti, R.: HMAC: Keyed-hashing for message authentication. In: Internet RFCs (1997)
[365] Krieger, J.P., Cabaset, S., Pestoni, G., Rohrmann, S., Faeh, D., Swiss National Cohort Study Group, et al.: Dietary patterns are associated with cardiovascular and cancer mortality among Swiss adults in a census-linked cohort. Nutrients **10**(3), 313 (2018)
[366] Kristensen, T.G., Nielsen, J., Pedersen, C.N.: A tree-based method for the rapid screening of chemical fingerprints. Algorithms for Molecular Biology **5**(1), 9 (2010)
[367] Kroll, M., Steinmetzer, S.: Who is 1011011111...1110110010? Automated cryptanalysis of Bloom filter encryptions of databases with several personal identifiers. In: International Joint Conference on Biomedical Engineering Systems and Technologies, pp. 341–356. Lisbon (2015)
[368] Krumm, J.: A survey of computational location privacy. Personal and Ubiquitous Computing **13**(6), 391–399 (2009)
[369] Kukich, K.: Techniques for automatically correcting words in text. ACM Computing Surveys **24**(4), 377–439 (1992)
[370] Kum, H.C., Krishnamurthy, A., Machanavajjhala, A., Reiter, M.K., Ahalt, S.: Privacy preserving interactive record linkage (PPIRL). Journal of the American Medical Informatics Association **21**(2), 212–220 (2014)
[371] Kuzu, M., Kantarcioglu, M., Durham, E., Malin, B.: A constraint satisfaction cryptanalysis of Bloom filters in private record linkage. In: Privacy Enhancing Technologies Symposium, pp. 226–245. Waterloo, Canada (2011)
[372] Kuzu, M., Kantarcioglu, M., Durham, E., Toth, C., Malin, B.: A practical approach to achieve private medical record linkage in light of public resources. Journal of the American Medical Informatics Association **20**(2), 285–292 (2013)

[373] Kuzu, M., Kantarcioglu, M., Inan, A., Bertino, E., Durham, E., Malin, B.: Efficient privacy-aware record integration. In: Conference on Extending Database Technology. Genoa (2013)
[374] Lai, P., Yiu, S., Chow, K., Chong, C., Hui, L.: An efficient Bloom filter based solution for multiparty private matching. In: Security and Management, p. 7. Las Vegas (2006)
[375] Lait, A., Randell, B.: An assessment of name matching algorithms. Tech. rep., Department of Computer Science, University of Newcastle upon Tyne (1993)
[376] Lambert, D.: Measures of disclosure risk and harm. Journal of Official Statistics **9**, 313–313 (1993)
[377] Larsen, M.D.: Record linkage, nondisclosure, counterterrorism, and statistics. In: Survey Methods Section, Canadian Statistical Society. London (2006)
[378] Lateral Economics: Valuing the Australian census (2019). URL `https://www.abs.gov.au/websitedbs/D3310114.nsf/home/Value+of+the+Australian+Census`
[379] Laufer, R.S., Wolfe, M.: Privacy as a concept and a social issue: A multidimensional developmental theory. Journal of Social Issues **33**(3), 22–42 (1977)
[380] Lauter, K., Naehrig, M., Vaikuntanathan, V.: Can homomorphic encryption be practical? In: ACM Cloud Computing Security Workshop, pp. 113–124. Chicago (2011)
[381] Lawrence, G., Dinh, I., Taylor, L.: The centre for health record linkage: a new resource for health services research and evaluation. Health Information Management Journal **37**(2), 60–62 (2008)
[382] Lazrig, I., Moataz, T., Ray, I., Ray, I., Ong, T., Kahn, M., Cuppens, F., Cuppens, N.: Privacy preserving record matching using automated semi-trusted broker. In: IFIP Data and Applications Security and Privacy, pp. 103–118. Fairfax, Virginia (2015)
[383] Lazrig, I., Ong, T., Ray, I., Ray, I., Jiang, X., Vaidya, J.: Privacy preserving probabilistic record linkage without trusted third party. In: Privacy, Security and Trust, pp. 1–10. Belfast (2018)
[384] Lee, Y., Pipino, L., Funk, J., Wang, R.: Journey to Data Quality. The MIT Press (2009)
[385] Leskovec, J., Rajaraman, A., Ullman, J.D.: Mining of Massive Datasets. Cambridge University Press (2014)
[386] Levenshtein, V.I.: Binary codes capable of correcting deletions, insertions, and reversals. Soviet Physics Doklady **10**(8), 707–710 (1966)
[387] Leyland, A., Barnard, M., McKeganey, N.: The use of capture-recapture methodology to estimate and describe covert populations: an application to female street-working prostitution in Glasgow. Bulletin de Methodologie Sociologique **38**, 52–73 (1993)

[388] Li, F., Chen, Y., Luo, B., Lee, D., Liu, P.: Privacy preserving group linkage. In: Scientific and Statistical Database Management, pp. 432–450. Portland (2011)
[389] Li, J., Baig, M.M., Sattar, A.S., Ding, X., Liu, J., Vincent, M.W.: A hybrid approach to prevent composition attacks for independent data releases. Information Sciences **367**, 324–336 (2016)
[390] Li, N., Li, T., Venkatasubramanian, S.: t-closeness: Privacy beyond k-anonymity and l-diversity. In: IEEE International Conference on Data Engineering, pp. 106–115. Istanbul (2007)
[391] Li, N., Lyu, M., Su, D., Yang, W.: Differential Privacy: From Theory to Practice. Synthesis Lectures on Information Security, Privacy, and Trust. Morgan and Claypool Publishers (2017)
[392] Li, P., Dong, X., Maurino, A., Srivastava, D.: Linking temporal records. VLDB Endowment **4**(11) (2011)
[393] Lin, J.: Scalable language processing algorithms for the masses: A case study in computing word co-occurrence matrices with mapreduce. In: Empirical Methods in Natural Language Processing, pp. 419–428. Honolulu (2008)
[394] Lindell, Y., Pinkas, B.: Secure multiparty computation for privacy-preserving data mining. Journal of Privacy and Confidentiality **1**(1), 5 (2009)
[395] Little, R.J.: A test of missing completely at random for multivariate data with missing values. Journal of the American Statistical Association **83**(404), 1198–1202 (1988)
[396] Little, R.J.A., Rubin, D.B.: Statistical Analysis with Missing Data, 3 edn. Wiley, Hoboken (2020)
[397] Liu, H., Wang, H., Chen, Y.: Ensuring data storage security against frequency-based attacks in wireless networks. In: Distributed Computing in Sensor Systems, pp. 201–215. Santa Barbara (2010)
[398] Lohr, S.: For big-data scientists,'janitor work' is key hurdle to insights. New York Times **17**, B4 (2014)
[399] Lyons, R.A., Ford, D.V., Moore, L., Rodgers, S.E.: Use of data linkage to measure the population health effect of non-health-care interventions. The Lancet **383**(9927), 1517–1519 (2014)
[400] Machanavajjhala, A., Gehrke, J., Kifer, D.: l-density: Privacy beyond k-anonymity. In: IEEE International Conference on Data Engineering. Atlanta (2006)
[401] Manning, C., Schütze, H.: Foundations of Statistical Natural Language Processing. MIT Press (1999)
[402] Martini, M., Kienle, T., Wagner, D., Weinzierl, Q., Wenzel, M.: Rechtliche Rahmenbedingungen für ein nationales Bildungsregister. Legal Expertise on Behalf of the Federal Ministry of Education and Research (BMBF), Speyer (2019)

[403] Martini, M., Wagner, D., Wenzel, M.: Rechtliche Grenzen einer Personen- bzw. Unternehmenskennziffer in staatlichen Registern. Speyer (2017)
[404] Massey, C.G., Genadek, K.R., Alexander, J.T., Gardner, T.K., O'Hara, A.: Linking the 1940 U.S. Census with modern data. Historical Methods: A Journal of Quantitative and Interdisciplinary History **51**(4), 246–257 (2018)
[405] Matwin, S., Nin, J., Sehatkar, M., Szapiro, T.: A review of attribute disclosure control. In: G. Navarro-Arribas, v. Torra (eds.) Advanced Research in Data Privacy, pp. 41–61. Springer (2015)
[406] Maxfield, M.G., Weiler, B.L., Widom, C.S.: Comparing self-reports and official records of arrests. Journal of Quantitative Criminology **16**(1), 87–110 (2000)
[407] McCallum, A., Nigam, K., Ungar, L.H.: Efficient clustering of high-dimensional data sets with application to reference matching. In: ACM Conference on Knowledge Discovery and Data Mining, pp. 169–178. Boston (2000)
[408] McCreight, E.: A space-economical suffix tree construction algorithm. Journal of the ACM **23**(2) (1976)
[409] McKeganey, N., Barnard, M., Leyland, A., Coote, I., Follet, E.: Female streetworking prostitution and HIV infection in Glasgow. British Medical Journal **305**(6857), 801–804 (1992)
[410] Meier, J., Jakscha, T., Schnell, R., Heller, G.: Verknüpfung der Module Geburtshilfe und Neonatologie des QS-Verfahrens Perinatalmedizin. technical report, IQTIG, Berlin (2017). URL `https://iqtig.org/downloads/spezifikation/2018/v01/TechDok_Verknuepfung_Peri_Neo_V03.pdf`
[411] Menard, S. (ed.): Handbook of Longitudinal Research: Design, Measurement, and Analysis. Elsevier, Amsterdam (2007)
[412] Mészáros, J., Ho, C.h.: Big data and scientific research: The secondary use of personal data under the research exemption in the GDPR. Hungarian Journal of Legal Studies **59**(4), 403–419 (2018)
[413] Meyer, B.D., Mok, W.K., Sullivan, J.X.: Household surveys in crisis. Journal of Economic Perspectives **29**(4), 199–226 (2015)
[414] Mitchell, R.J., Cameron, C.M., McClure, R.J., Williamson, A.M.: Data linkage capabilities in Australia: Practical issues identified by a Population Health Research Network 'proof of concept project'. Australian and New Zealand Journal of Public Health **39**(4), 319–325 (2015)
[415] Mitchell, W., Dewri, R., Thurimella, R., Roschke, M.: A graph traversal attack on Bloom filter-based medical data aggregation. International Journal of Big Data Intelligence **4**(4), 217–226 (2017)
[416] Mittelstadt, B., Benzler, J., Engelmann, L., Prainsack, B., Vayena, E.: Is there a duty to participate in digital epidemiology? Life Sciences, Society and Policy **14**(1), 9 (2018)

[417] Mitzenmacher, M., Upfal, E.: Probability and Computing: Randomized Algorithms and Probabilistic Analysis. Cambridge University Press, Cambridge (2005)
[418] Mohammed, N., Fung, B.C., Debbabi, M.: Anonymity meets game theory: secure data integration with malicious participants. VLDB Journal **20**(4), 567–588 (2011)
[419] Mondschein, C.F., Monda, C.: The EU's General Data Protection Regulation (GDPR) in a Research Context. In: P. Kubben, M. Dumontier, A. Dekker (eds.) Fundamentals of Clinical Data Science, pp. 55–71. Springer (2019)
[420] Monge, A.E.: Matching algorithms within a duplicate detection system. IEEE Data Engineering Bulletin **23**(4), 14–20 (2000)
[421] Monge, A.E., Elkan, C.P.: The field-matching problem: Algorithm and applications. In: ACM Conference on Knowledge Discovery and Data Mining, pp. 267–270. Portland (1996)
[422] Moore, H.C., Guiver, T., Woollacott, A., de Klerk, N., Gidding, H.F.: Establishing a process for conducting cross-jurisdictional record linkage in Australia. Australian and New Zealand Journal of Public Health **40**(2), 159–164 (2016)
[423] Mulder, T., Tudorica, M.: Privacy policies, cross-border health data and the GDPR. Information and Communications Technology Law **28**(3), 261–274 (2019)
[424] Muralidhar, K., Sarathy, R.: Data shuffling – a new masking approach for numerical data. Management Science **52**(5), 658–670 (2006)
[425] Nanayakkara, C., Christen, P., Ranbaduge, T.: Robust temporal graph clustering for group record linkage. In: Pacific-Asia Conference on Knowledge Discovery and Data Mining, pp. 526–538. Macau (2019)
[426] Nanayakkara, C., Christen, P., Ranbaduge, T., Garrett, E.: Evaluation measure for group-based record linkage. International Journal of Population Data Science **4**(1) (2019)
[427] National Academies of Sciences, Engineering, and Medicine: Reducing Response Burden in the American Community Survey. National Academies Press, Washington, D.C. (2016)
[428] National Academies of Sciences, Engineering, and Medicine: Improving the American Community Survey. Washington, DC (2019)
[429] National Data Guardian: Review of data security, consent and opt-outs. Document 2904918 (2016)
[430] National Health and Medical Research Council: National statement on ethical conduct in human research. `www.nhmrc.gov.au/file/9131` (2018)
[431] National Research Council: Protecting Individual Privacy in the Struggle Against Terrorists: a Framework for Program Assessment. National Academy of Sciences, Washington, DC (2008)

[432] Naumann, F., Herschel, M.: An Introduction to Duplicate Detection. Synthesis Lectures on Data Management. Morgan and Claypool Publishers (2010)
[433] Navarro, G.: A guided tour to approximate string matching. ACM Computing Surveys **33**(1), 31–88 (2001)
[434] Neubauer, T., Heurix, J.: A methodology for the pseudonymization of medical data. International Journal of Medical Informatics **80**(3), 190–204 (2011)
[435] Newcombe, H., Kennedy, J.: Record linkage: making maximum use of the discriminating power of identifying information. Communications of the ACM **5**(11), 563–566 (1962)
[436] Newcombe, H., Kennedy, J., Axford, S., James, A.: Automatic linkage of vital records. Science **130**(3381), 954–959 (1959)
[437] Newcombe, H.B.: Handbook of Record Linkage: Methods for Health and Statistical Studies, Administration, and Business. Oxford University Press, Inc., New York (1988)
[438] Niedermeyer, F., Steinmetzer, S., Kroll, M., Schnell, R.: Cryptanalysis of basic Bloom filters used for privacy preserving record linkage. Journal of Privacy and Confidentiality **6**(2), 59–79 (2014)
[439] Nisbet, M.C., Nisbet, E.C.: The Public Face of Science Across the World. American Academy of Arts and Sciences (2019)
[440] Nissenbaum, H.: Privacy in Context: Technology, Policy, and the Integrity of Social Life. Stanford University Press, Stanford (2010)
[441] Nissenbaum, H.: Contextual integrity up and down the data food chain. Theoretical Inquiries in Law **20**(1), 221–256 (2019)
[442] Nissim, K., Steinke, T., Wood, A., Altman, M., Bembenek, A., Bun, M., Gaboardi, M., O'Brien, D.R., Vadhan, S.: Differential privacy: A primer for a non-technical audience. In: Privacy Law Scholars Conference. Berkeley (2017)
[443] Norberg, P.A., Horne, D.R., Horne, D.A.: The privacy paradox: Personal information disclosure intentions versus behaviors. Journal of Consumer Affairs **41**(1), 100–126 (2007)
[444] Obar, J.A., Oeldorf-Hirsch, A.: The biggest lie on the Internet: ignoring the privacy policies and terms of service policies of social networking services. Information, Communication and Society pp. 1–20 (2018)
[445] OECD: Health Data Governance: Privacy, Monitoring and Research. OECD Publishing, Paris (2015)
[446] Office for National Statistics: Beyond 2011 matching anonymous data (2013). Methods and Policies Report M9
[447] Office for National Statistics: Beyond 2011 safeguarding data for research: Our policy (2013). Methods and Policies Report M10
[448] Office for National Statistics: 2011 census England and Wales general report (2015)
[449] Office for National Statistics: 2011 census benefits evaluation report (2019). URL `https://www.ons.`

gov.uk/census/2011census/2011censusbenefits/2011censusbenefitsevaluationreport
[450] O'Hare, W.P.: Differential Undercounts in the U.S. Census: Who is Missed? Springer, Cham (2019)
[451] Ohm, P.: Broken promises of privacy: Responding to the surprising failure of anonymization. UCLA Law Review **57**, 1701 (2009)
[452] Ohm, P.: Sensitive information. Southern California Law Review **88**, 1125–1196 (2014)
[453] O'Keefe, C., Yung, M., Gu, L., Baxter, R.: Privacy-preserving data linkage protocols. In: ACM Workshop on Privacy in the Electronic Society, pp. 94–102. Washington DC (2004)
[454] O'Keefe, C.M., Rubin, D.B.: Individual privacy versus public good: protecting confidentiality in health research. Statistics in Medicine **34**(23), 3081–3103 (2015)
[455] Ong, T.C., Mannino, M.V., Schilling, L.M., Kahn, M.G.: Improving record linkage performance in the presence of missing linkage data. Journal of Biomedical Informatics **52**, 43–54 (2014)
[456] Paillier, P.: Public-key cryptosystems based on composite degree residuosity classes. In: Theory and Application of Cryptographic Techniques, pp. 223–238. Prague (1999)
[457] Panel for the Future of Science and Technology: How the General Data Protection Regulation changes the rules for scientific research. European Parliamentary Research Service (2019)
[458] Pang, C., Gu, L., Hansen, D., Maeder, A.: Privacy-preserving fuzzy matching using a public reference table. In: S. McClean, P. Millard, E. El-Darzi, C. Nugent (eds.) Intelligent Patient Management, pp. 71–89. Springer (2009)
[459] Papadakis, G., Alexiou, G., Papastefanatos, G., Koutrika, G.: Schema-agnostic vs schema-based configurations for blocking methods on homogeneous data. VLDB Endowment **9**(4), 312–323 (2015)
[460] Papadakis, G., Ioannou, E., Palpanas, T., Niederee, C., Nejdl, W.: A blocking framework for entity resolution in highly heterogeneous information spaces. Transactions on Knowledge and Data Engineering **25**(12), 2665–2682 (2012)
[461] Papadakis, G., Palpanas, T.: Blocking for large-scale entity resolution: challenges, algorithms, and practical examples. In: IEEE International Conference on Data Engineering, pp. 1436–1439. Helsinki (2016)
[462] Papadakis, G., Skoutas, D., Thanos, E., Palpanas, T.: Blocking and filtering techniques for entity resolution: A survey. ACM Computing Surveys **53**(2), 1–42 (2020)
[463] Papadakis, G., Svirsky, J., Gal, A., Palpanas, T.: Comparative analysis of approximate blocking techniques for entity resolution. VLDB Endowment **9**(9), 684–695 (2016)

[464] Park, J., Sandhu, R.: Towards usage control models: beyond traditional access control. In: ACM Symposium on Access Control Models and Technologies, pp. 57–64. Monterey (2002)
[465] Parker, M.: Humble Pi – A Comedy of Maths Errors. Penguin Random House (2019)
[466] Paterson, M., McDonagh, M.: Data protection in an era of Big data: The challenges posed by big personal data. Monash University Law Review **44**(1) (2018)
[467] Patman, F., Thompson, P.: Names: A new frontier in text mining. In: Intelligence and Security Informatics, pp. 27–38. Tuscon (2003)
[468] Patrascu, M., Thorup, M.: The power of simple tabulation hashing. In: ACM Symposium on Theory of Computing, pp. 1–10. San Jose (2011)
[469] Paul, C., Noel, H., Charles, A., Jeffrey, W., Daniel, E.: Options for encoding name information for use in record linkage. Tech. Rep. 1351.0.55.162, Australian Bureau of Statistics, Canberra (2018)
[470] Perlman, R.: An overview of PKI trust models. IEEE Network **13**(6), 38–43 (1999)
[471] Petrila, J.: Legal issues in the use of electronic data systems for social science research. In: J. Fantuzzo, D.P. Culhane (eds.) Actionable Intelligence, pp. 39–75. Palgrave Macmillan US, New York (2015)
[472] Philips, L.: The double-metaphone search algorithm. C/C++ User's Journal **18**(6) (2000)
[473] Phua, C., Smith-Miles, K., Lee, V., Gayler, R.: Resilient identity crime detection. IEEE Transactions on Knowledge and Data Engineering **24**(3) (2012)
[474] Pita, R., Mendonça, E., Reis, S., Barreto, M., Denaxas, S.: A machine learning trainable model to assess the accuracy of probabilistic record linkage. In: Big Data Analytics and Knowledge Discovery, pp. 214–227. Lyon (2017)
[475] Pita, R., Pinto, C., Sena, S., Fiaccone, R., Amorim, L., Reis, S., Barreto, M., Denaxas, S., Barreto, M.: On the accuracy and scalability of probabilistic data linkage over the Brazilian 114 million cohort. Journal of Biomedical and Health Informatics **22**(2), 346–353 (2018)
[476] Pöge, A.: Persönliche Codes bei Längsschnittuntersuchungen III. Methoden – Daten – Analysen **5**(1), 109–134 (2011)
[477] Pollock, J.J., Zamora, A.: Automatic spelling correction in scientific and scholarly text. Communications of the ACM **27**(4), 358–368 (1984)
[478] Porter, E.H., Winkler, W.E.: Approximate string comparison and its effect on an advanced record linkage system. Tech. Rep. RR97/02, US Bureau of the Census (1997)
[479] Postel, H.J.: Die Kölner Phonetik: Ein Verfahren zur Identifizierung von Personennamen auf der Grundlage der Gestaltanalyse. IBM-Nachrichten **19**, 925–931 (1969)
[480] Pow, C., Iron, K., Boyd, J., Brown, A., Thompson, S., Chong, N., Ma, C.: Privacy-preserving record linkage: an international collabora-

tion between Canada, Australia and Wales. International Journal for Population Data Science **1**(1) (2017)
[481] Preisendörfer, P., Wolter, F.: Who is telling the truth? a validation study on determinants of response behavior in surveys. Public Opinion Quarterly **78**(1), 126–146 (2014)
[482] Prewitt, K.: Why it matters to distinguish between privacy and confidentiality. Journal of Privacy and Confidentiality **3**(2), 41–47 (2011)
[483] Productivity Commission: Data availability and use. Report No. 82, Canberra (2017)
[484] Pyle, D.: Data Preparation for Data Mining. Morgan Kaufmann (1999)
[485] Quantin, C., Benzenine, E., Allaert, F., Guesdon, M., Gouyon, J., Riandey, B.: Epidemiological and statistical secured matching in France. Statistical Journal of the IAOS **30**(3), 255–261 (2014)
[486] Quantin, C., Bouzelat, H., Allaert, F., Benhamiche, A., Faivre, J., Dusserre, L.: How to ensure data quality of an epidemiological follow-up: Quality assessment of an anonymous record linkage procedure. International Journal of Medical Informatics **49**(1), 117–122 (1998)
[487] Quantin, C., Bouzelat, H., Allaert, F.A., Benhamiche, A.M., Faivre, J., Dusserre, L.: Automatic record hash coding and linkage for epidemiological follow-up data confidentiality. Methods of Information in Medicine **37**(3), 271–277 (1998)
[488] Quantin, C., Bouzelat, H., Dusserre, L.: Irreversible encryption method by generation of polynomials. Medical Informatics and the Internet in Medicine **21**(2), 113–121 (1996)
[489] Rabin, M.O.: How to exchange secrets with oblivious transfer. Tech. Rep. TR-81, Aiken Computation Lab, Harvard University (1981)
[490] Ragan, E.D., Kum, H.C., Ilangovan, G., Wang, H.: Balancing privacy and information disclosure in interactive record linkage with visual masking. In: Human Factors in Computing Systems, pp. 1–12. Montreal (2018)
[491] Rahm, E., Do, H.H.: Data cleaning: Problems and current approaches. IEEE Data Engineering Bulletin **23**(4), 3–13 (2000)
[492] Ramadan, B., Christen, P.: Unsupervised blocking key selection for real-time entity resolution. In: Pacific-Asia Conference on Knowledge Discovery and Data Mining, pp. 574–585. Ho Chi Minh City (2015)
[493] Ramadan, B., Christen, P., Liang, H., Gayler, R.W.: Dynamic sorted neighborhood indexing for real-time entity resolution. ACM Journal of Data and Information Quality **6**(4), 15 (2015)
[494] Ranbaduge, T.: A scalable blocking framework for multidatabase privacy-preserving record linkage. Ph.D. thesis, Research School of Computer Science, The Australian National University (2018)
[495] Ranbaduge, T., Christen, P.: Privacy-preserving temporal record linkage. In: IEEE International Conference on Data Mining, pp. 377–386. Singapore (2018)

[496] Ranbaduge, T., Christen, P.: A scalable privacy-preserving framework for temporal record linkage. Knowledge and Information Systems pp. 1–34 (2018)
[497] Ranbaduge, T., Christen, P., Schnell, R.: Secure and accurate two-step hash encoding for privacy-preserving record linkage. In: Pacific-Asia Conference on Knowledge Discovery and Data Mining. Singapore (2020)
[498] Ranbaduge, T., Christen, P., Vatsalan, D.: Tree based scalable indexing for multi-party privacy-preserving record linkage. In: Australasian Data Mining Conference, vol. 158. Brisbane (2014)
[499] Ranbaduge, T., Christen, P., Vatsalan, D.: Clustering-based scalable indexing for multi-party privacy-preserving record linkage. In: Pacific-Asia Conference on Knowledge Discovery and Data Mining. Hanoi (2015)
[500] Ranbaduge, T., Schnell, R.: Securing Bloom filters for privacy-preserving record linkage. In: ACM Conference on Information and Knowledge Management. Galway (2020)
[501] Ranbaduge, T., Vatsalan, D., Christen, P.: Scalable block scheduling for efficient multi-database record linkage. In: IEEE International Conference on Data Mining, pp. 1161–1166. Barcelona (2016)
[502] Ranbaduge, T., Vatsalan, D., Christen, P.: Secure multi-party summation protocols: Are they secure enough under collusion? Transactions on Data Privacy **13**(1), 25–60 (2020)
[503] Ranbaduge, T., Vatsalan, D., Christen, P., Verykios, V.S.: Hashing-based distributed multi-party blocking for privacy-preserving record linkage. In: Pacific-Asia Conference on Knowledge Discovery and Data Mining, pp. 415–427. Auckland (2016)
[504] Randall, S.M., Boyd, J.H., Ferrante, A.M., Bauer, J.K., Semmens, J.B.: Use of graph theory measures to identify errors in record linkage. Computer Methods and Programs in Biomedicine **115**(2), 55–63 (2014)
[505] Randall, S.M., Brown, A.P., Ferrante, A.M., Boyd, J.H.: Privacy preserving linkage using multiple dynamic match keys. International Journal of Population Data Science **4**(1) (2019)
[506] Randall, S.M., Brown, A.P., Ferrante, A.M., Boyd, J.J., Semmens, J.B.: Privacy preserving record linkage using homomorphic encryption. In: Workshop Population Informatics for Big Data, held at ACM SIGKDD. Sydney (2015)
[507] Randall, S.M., Ferrante, A.M., Boyd, J.H., Bauer, J.K., Semmens, J.B.: Privacy-preserving record linkage on large real world datasets. Journal of Biomedical Informatics **50**, 205–212 (2014)
[508] Randall, S.M., Ferrante, A.M., Boyd, J.H., Brown, A.P., Semmens, J.B.: Limited privacy protection and poor sensitivity: Is it time to move on from the statistical linkage key-581? Health Information Management Journal **45**(2), 71–79 (2016)

[509] Randall, S.M., Ferrante, A.M., Boyd, J.H., Semmens, J.B.: The effect of data cleaning on record linkage quality. BMC Medical Informatics and Decision Making **13**(1), 64 (2013)
[510] Rao, F.Y., Cao, J., Bertino, E., Kantarcioglu, M.: Hybrid private record linkage: Separating differentially private synopses from matching records. ACM Transactions on Privacy and Security **22**(3), 1–36 (2019)
[511] Rastogi, V., Dalvi, N., Garofalakis, M.: Large-scale collective entity matching. VLDB Endowment **4**, 208–218 (2011)
[512] Rastogi, V., Suciu, D., Hong, S.: The boundary between privacy and utility in data publishing. In: Conference on Very Large Data Bases, pp. 531–542. Vienna (2007)
[513] Rat für Sozial- und Wirtschaftsdaten: Handreichung Datenschutz. RatSWD, Berlin (2017)
[514] Reid, A., Davies, R., Garrett, E.: Nineteenth-century Scottish demography from linked censuses and civil registers: A 'sets of related individuals' approach. History and Computing **14**(1–2), 61–86 (2002)
[515] Reiter, M.K., Rubin, A.D.: Crowds: Anonymity for web transactions. ACM Transactions on Information and System Security **1**(1), 66–92 (1998)
[516] Richterich, A.: The Big Data Agenda: Data Ethics and Critical Data Studies. University of Westminster Press, London (2018)
[517] Rivera Drew, J.A., Flood, S., Warren, J.R.: Making full use of the longitudinal design of the current population survey: Methods for linking records across 16 months. Journal of Economic and Social Measurement **39**(3), 121–144 (2014)
[518] Rivest, R.L.: Chaffing and winnowing: Confidentiality without encryption. MIT Lab for Computer Science (1998). Available at: `http://theory.lcs.mit.edu/~rivest/chaffing.txt`
[519] Rivest, R.L., Shamir, A., Adleman, L.: A method for obtaining digital signatures and public-key cryptosystems. Communications of the ACM **21**(2), 120–126 (1978)
[520] Roos, L.L., Walld, R., Wajda, A., Bond, R., Hartford, K.: Record linkage strategies, outpatient procedures, and administrative data. Medical Care pp. 570–582 (1996)
[521] Royal Society and British Academy: Data governance: Landscape review. `https://royalsociety.org/-/media/policy/projects/data-governance/data-governance-landscape-review.pdf`, London (2017)
[522] Ruggles, S., Fitch, C.A., Roberts, E.: Historical census record linkage. Annual Review of Sociology **44**(1), 19–37 (2018)
[523] Rumbold, J.M.M., Pierscionek, B.: The effect of the general data protection regulation on medical research. Journal of Medical Internet Research **19**(2), e47 (2017)
[524] Russell, R.: The Soundex coding system. US patent 1261167 (1918)

[525] Sadinle, M.: Bayesian estimation of bipartite matchings for record linkage. Journal of the American Statistical Association **112**(518), 600–612 (2017)
[526] Samarati, P.: Protecting respondents identities in microdata release. Transactions on Knowledge and Data Engineering **13**(6), 1010–1027 (2001)
[527] Samarati, P., de Vimercati, S.C.: Access control: Policies, models, and mechanisms. In: Foundations of Security Analysis and Design, pp. 137–196. Springer (2000)
[528] Sandhu, R.S., Samarati, P.: Access control: principle and practice. IEEE Communications Magazine **32**(9), 40–48 (1994)
[529] Sarathy, R., Muralidhar, K.: Secure and useful data sharing. Decision Support Systems **42**(1), 204–220 (2006)
[530] Sarawagi, S.: Information extraction. Foundations and Trends in Databases **1**(3), 261–377 (2008)
[531] Saris, W.E., Gallhofer, I.N.: Design, Evaluation, and Analysis of Questionnaires for Survey Research, 2 edn. Wiley, Hoboken (2014)
[532] Sayers, A., Ben-Shlomo, Y., Blom, A.W., Steele, F.: Probabilistic record linkage. International Journal of Epidemiology **45**(3), 954–964 (2016)
[533] Scannapieco, M., Figotin, I., Bertino, E., Elmagarmid, A.: Privacy preserving schema and data matching. In: ACM Conference on Management of Data, pp. 653–664. Beijing (2007)
[534] Schmidlin, K., Clough-Gorr, K.M., Spoerri, A.: Privacy preserving probabilistic record linkage (P3RL): a novel method for linking existing health-related data and maintaining participant confidentiality. BMC Medical Research Methodology **15**(1), 46 (2015)
[535] Schneier, B.: Applied Cryptography: Protocols, Algorithms, and Source Code in C, 2 edn. John Wiley and Sons, Inc., New York (1996)
[536] Schnell, R.: An efficient privacy-preserving record linkage technique for administrative data and censuses. Journal of the International Association for Official Statistics **30**(3), 263–270 (2014)
[537] Schnell, R.: Linking surveys and administrative data. In: U. Engel, B. Jann, P. Lynn, A. Scherpenzeel, P. Sturgis (eds.) Improving Survey Methods: Lessons from Recent Research, pp. 273–287. Routledge, New York (2014)
[538] Schnell, R.: Privacy-preserving record linkage. In: K. Harron, H. Goldstein, C. Dibben (eds.) Methodological Developments in Data Linkage, pp. 201–225. John Wiley & Sons (2015)
[539] Schnell, R.: 'Big Data' aus wissenschaftssoziologischer Sicht: Warum es kaum sozialwissenschaftliche Studien ohne Befragungen gibt. Tech. Rep. WP-GRLC-2018-01, German Record Linkage Center (2018). URL `https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3548537`
[540] Schnell, R.: 'Big Data' aus wissenschaftssoziologischer Sicht: Warum es kaum sozialwissenschaftliche Studien ohne Befragungen gibt. In:

D. Baron, O. Arránz Becker, D. Lois (eds.) Erklärende Soziologie und soziale Praxis, pp. 101–125. Springer, Wiesbaden (2019)
[541] Schnell, R.: Survey-Interviews: Methoden standardisierter Befragungen, 2 edn. Springer VS, Wiesbaden (2019)
[542] Schnell, R., Bachteler, T., Bender, S.: A toolbox for record linkage. Austrian Journal of Statistics **33**(1 & 2), 125–133 (2004)
[543] Schnell, R., Bachteler, T., Reiher, J.: Privacy-preserving record linkage using Bloom filters. BMC Medical Informatics and Decision Making **9**(1) (2009)
[544] Schnell, R., Bachteler, T., Reiher, J.: Improving the use of self-generated identification codes. Evaluation Review **34**(5), 391–418 (2010)
[545] Schnell, R., Bachteler, T., Reiher, J.: A novel error-tolerant anonymous linking code. German Record Linkage Center (WP-GRLC-2011-02) (2011)
[546] Schnell, R., Borgs, C.: Randomized response and balanced Bloom filters for privacy preserving record linkage. In: Workshop on Data Integration and Applications, held at IEEE ICDM. Barcelona (2016)
[547] Schnell, R., Borgs, C.: XOR-folding for Bloom filter-based encryptions for privacy-preserving record linkage. German Record Linkage Center (WP-GRLC-2016-03) (2016)
[548] Schnell, R., Borgs, C.: Hardening encrypted patient names against cryptographic attacks using cellular automata. In: Workshop on Data Integration and Applications, held at IEEE ICDM. Singapore (2018)
[549] Schnell, R., Borgs, C.: Protecting record linkage identifiers using a language model for patient names. Studies in Health Technology and Informatics **253**, 91–95 (2018)
[550] Schnell, R., Borgs, C.: Abschlussbericht des Record Linkage für die Leistungsbereiche Geburtshilfe und Neonatologie. Tech. rep., IQTIG, Berlin (2019)
[551] Schnell, R., Borgs, C.: Encoding hierarchical classification codes for privacy-preserving record linkage using Bloom filters. In: Workshop on Data Integration and Applications, held at ECML/PKDD, pp. 142–156. Springer, Würzburg (2019)
[552] Schnell, R., Richter, A., Borgs, C.: Performance of different methods for privacy preserving record linkage with large scale medical data sets. In: International Health Data Linkage Conference. Vancouver (2014)
[553] Schnell, R., Rukasz, D., Borgs, C., Brumme, S., et al.: R PPRL toolbox. `https://cran.r-project.org/web/packages/PPRL/` (2018)
[554] Schröder, D.: Transcript of the public hearing, health committee, 16.10.2019. Tech. rep., Deutscher Bundestag (2019)
[555] Sehili, Z., Kolb, L., Borgs, C., Schnell, R., Rahm, E.: Privacy preserving record linkage with PPJoin. In: BTW Conference. Hamburg (2015)

[556] Sehili, Z., Rahm, E.: Speeding up privacy preserving record linkage for metric space similarity measures. Datenbank-Spektrum **16**(3), 227–236 (2016)
[557] Settles, B.: Active learning. Synthesis Lectures on AI and ML (2012)
[558] Shamir, A.: How to share a secret. Communications of the ACM **22**(11), 612–613 (1979)
[559] Shannon, C.E.: A mathematical theory of communication. Bell System Technical Journal **27**(3), 379–423 (1948)
[560] Shimizu, K., Nuida, K., Rätsch, G.: Efficient privacy-preserving string search and an application in genomics. Bioinformatics **32**(11) (2016)
[561] Shlomo, N.: Probabilistic record linkage for disclosure risk assessment. In: Privacy in Statistical Databases, pp. 269–282. Ibiza (2014)
[562] Singh, S.: The Code Book: The Secret History of Codes and Codebreaking. Fourth Estate (2000)
[563] Smalheiser, N.R., Torvik, V.I.: Author name disambiguation. Annual Review of Information Science and Technology **43**(1), 1–43 (2009)
[564] Smith, D.: Secure pseudonymisation for privacy-preserving probabilistic record linkage. Journal of Information Security and Applications **34**, 271–279 (2017)
[565] Smith, H.J., Dinev, T., Xu, H.: Information privacy research: An interdisciplinary review. MIS Quarterly **35**(4), 989–1015 (2011)
[566] Smith, T.F., Waterman, M.S.: Identification of common molecular subsequences. Journal of Molecular Biology **147**(1), 195–197 (1981)
[567] Smith, T.T.: Examining data privacy breaches in healthcare. Ph.D. thesis, School of Business Administration, Walden University (2016)
[568] Snae, C.: A comparison and analysis of name matching algorithms. International Journal of Applied Science, Engineering and Technology **4**(1), 252–257 (2007)
[569] Sowey, E.R.: A chronological and classified bibliography on random number generation and testing. International Statistical Review pp. 355–371 (1972)
[570] Spindler, G., Schmechel, P.: Personal data and encryption in the European General Data Protection Regulation. Journal of Intellectual Property, Information Technology and Electronic Commerce Law **7**(2), 163–177 (2016)
[571] Spoerri, A., Zwahlen, M., Egger, M., Bopp, M.: The Swiss National Cohort: a unique database for national and international researchers. International Journal of Public Health **55**(4), 239 (2010)
[572] Stalla-Bourdillon, S., Knight, A.: Anonymous data v. personal data – a false debate: An EU perspective on anonymization, pseudonymization and personal data. Wisconsin International Law Journal **34**(2), 284–322 (2017)
[573] Stallings, W.: Information Privacy Engineering and Privacy by Design: Understanding Privacy Threats, Technology, and Regulations Based on Standards and Best Practices. Pearson (2020)

[574] Stegmaier, C., Hentschel, S., Hofstädter, F., Katalinic, A., Tillack, A., Klinkhammer-Schalke, M.: Das Manual der Krebsregistrierung. Zuckschwerdt, Munich (2018)
[575] Stenberg, S.Å.: Born in 1953: The Story about a Post-war Swedish Cohort, and a Longitudinal Research Project. Stockholm University Press (2018)
[576] Stenberg, S.Å., Vågerö, D., Österman, R., Arvidsson, E., Von Otter, C., Janson, C.G.: Stockholm birth cohort study 1953—2003: A new tool for life-course studies. Scandinavian Journal of Public Health **35**(1), 104–110 (2007)
[577] Stiles, P.G., Boothroyd, R.A.: Ethical use of administrative data for research purposes. In: J. Fantuzzo, D.P. Culhane (eds.) Actionable Intelligence, pp. 125–155. Palgrave Macmillan US, New York (2015)
[578] Stinson, D.R.: Cryptography: Theory and Practice. Chapman and Hall/CRC (2005)
[579] Stocks, P.: The measurement of morbidity. Proceedings of the Royal Society of Medicine **37**(10), 593 (1944)
[580] Sun, L., Zhang, L., Ye, X.: Randomized bit vector: Privacy-preserving encoding mechanism. In: ACM Conference on Information and Knowledge Management, pp. 1263–1272. Turin (2018)
[581] Sušelj, M., Marčun, T., Trček, D., Kandus, G.: Application of PKI in health care—needs, ambitions, prospects. In: Medical Informatics Europe. St Malo, France (2003)
[582] Sweeney, L.: Weaving technology and policy together to maintain confidentiality. The Journal of Law, Medicine & Ethics **25**(2-3), 98–110 (1997)
[583] Sweeney, L.: Computational disclosure control: A primer on data privacy protection. Ph.D. thesis, Massachusetts Institute of Technology, Dept. of Electrical Engineering and Computer Science (2001)
[584] Sweeney, L.: K-anonymity: A model for protecting privacy. International Journal of Uncertainty Fuzziness and Knowledge Based Systems **10**(5), 557–570 (2002)
[585] Tai, X.H., Eddy, W.F.: Automatically matching topographical measurements of cartridge cases using a record linkage framework. arXiv Preprint (2020)
[586] Takhshid, Z.: Retrievable images on social media platforms: A call for a new privacy tort. Buffalo Law Review **68**(1) (2020)
[587] Talburt, J.: Entity Resolution and Information Quality. Morgan Kaufmann (2011)
[588] Taylor, L., Zhou, X.H., Rise, P.: A tutorial in assessing disclosure risk in microdata. Statistics in Medicine **37**(25), 3693–3706 (2018)
[589] Tejada, S., Knoblock, C.A., Minton, S.: Learning domain-independent string transformation weights for high accuracy object identification. In: ACM Conference on Knowledge Discovery and Data Mining, pp. 350–359. Edmonton (2002)

[590] Templ, M.: Statistical Disclosure Control for Microdata: Methods and Applications in R. Springer, Cham (2017)
[591] Templ, M., Meindl, B., Kowarik, A., Chen, S.: Introduction to statistical disclosure control (SDC). Working paper 7, https://ihsn.org/sites/default/files/resources/ihsn-working-paper-007-Oct27.pdf, International Household Survey Network (2014)
[592] Thomas, R., Walport, M.: Data sharing review report. Ministry of Justice (2008)
[593] Thompson, S.A., Warzel, C.: Twelve million phones, one dataset, zero privacy. New York Times (2019)
[594] Tinabo, R., Mtenzi, F., O'Shea, B.: Anonymisation vs. pseudonymisation: Which one is most useful for both privacy protection and usefulness of e-healthcare data. In: IEEE Internet Technology and Secured Transactions, pp. 1–6. London (2009)
[595] Torra, V.: Data Privacy: Foundations, new Developments and the Big Data Challenge. Springer, Cham (2017)
[596] Toth, C., Durham, E., Kantarcioglu, M., Xue, Y., Malin, B.: SOEMPI: A secure open enterprise master patient index software toolkit for private record linkage. In: AMIA Annual Symposium Proceedings, vol. 2014, p. 1105. American Medical Informatics Association, Washington DC (2014)
[597] Tourangeau, R., Yan, T.: Sensitive questions in surveys. Psychological Bulletin **133**(5), 859–883 (2007)
[598] Tran, K.N., Vatsalan, D., Christen, P.: GeCo: an online personal data generator and corruptor. In: ACM Conference on Information and Knowledge Management, pp. 2473–2476. San Francisco (2013)
[599] Trepetin, S.: Privacy-preserving string comparisons in record linkage systems: a review. Information Security Journal: A Global Perspective **17**(5), 253–266 (2008)
[600] Trinckes, J.J.: The Definitive Guide to Complying with the HIPAA/HITECH Privacy and Security Rules. CRC Press, Boca Raton (2013)
[601] UNECE: Main results of the UNECE-UNSD survey on the 2010 round of population and housing censuses. Technical Report ECE/CES/GE41/2009/25, United Nations Economic Commission for Europe (2009)
[602] UNECE: Census methodology: Key results the UNECE survey on national census practices, and first proposals about the ces recommendations for the 2020 census round. Technical Report ECE/CES/GE41/2013/3, United Nations Economic Commission for Europe (2013)
[603] United Nations Secretariat: Post enumeration surveys: Operational guidelines. Department Of Economic And Social Affairs, Statistics Division, United Nations, New York (2010)

[604] U.S. Census Bureau: Availability of census records about individuals. www.census.gov/prod/2000pubs/cff-2.pdf (2008)
[605] Vaidya, J., Clifton, C.: Secure set intersection cardinality with application to association rule mining. Journal of Computer Security **13**(4), 593–622 (2005)
[606] Vaidya, J., Clifton, C., Zhu, M.: Privacy Preserving Data Mining. Springer (2006)
[607] Vaiwsri, S., Ranbaduge, T., Christen, P.: Reference values based hardening for Bloom filters based privacy-preserving record linkage. In: Australasian Data Mining Conference, pp. 189–202. Bathurst (2018)
[608] Van Eycken, E., Haustermans, K., Buntinx, F., Ceuppens, A., Weyler, J., Wauters, E., Van, O.: Evaluation of the encryption procedure and record linkage in the Belgian National Cancer Registry. Archives of Public Health **58**(6), 281–294 (2000)
[609] Van Oorschot, P.C., Menezes, A.J., Vanstone, S.A.: Handbook of Applied Cryptography. CRC Press (1996)
[610] Vatsalan, D.: Scalable and approximate privacy-preserving record linkage. Ph.D. thesis, Research School of Computer Science, The Australian National University (2014)
[611] Vatsalan, D., Christen, P.: An iterative two-party protocol for scalable privacy-preserving record linkage. In: Australasian Data Mining Conference, vol. 134. Sydney (2012)
[612] Vatsalan, D., Christen, P.: Sorted nearest neighborhood clustering for efficient private blocking. In: Pacific-Asia Conference on Knowledge Discovery and Data Mining, pp. 341–352. Gold Coast, Australia (2013)
[613] Vatsalan, D., Christen, P.: Scalable privacy-preserving record linkage for multiple databases. In: ACM Conference on Information and Knowledge Management. Shanghai (2014)
[614] Vatsalan, D., Christen, P.: Privacy-preserving matching of similar patients. Journal of Biomedical Informatics **59**, 285–298 (2016)
[615] Vatsalan, D., Christen, P., O'Keefe, C.M., Verykios, V.S.: An evaluation framework for privacy-preserving record linkage. Journal of Privacy and Confidentiality **6**(1) (2014)
[616] Vatsalan, D., Christen, P., Rahm, E.: Scalable privacy-preserving linking of multiple databases using counting Bloom filters. In: IEEE ICDM Workshops. Barcelona (2016)
[617] Vatsalan, D., Christen, P., Rahm, E.: Incremental clustering techniques for multi-party privacy-preserving record linkage. Data and Knowledge Engineering (2020)
[618] Vatsalan, D., Christen, P., Verykios, V.S.: Efficient two-party private blocking based on sorted nearest neighborhood clustering. In: ACM Conference on Information and Knowledge Management, pp. 1949–1958. San Francisco (2013)

[619] Vatsalan, D., Christen, P., Verykios, V.S.: A taxonomy of privacy-preserving record linkage techniques. Information Systems **38**(6), 946–969 (2013)
[620] Vatsalan, D., Sehili, Z., Christen, P., Rahm, E.: Privacy-preserving record linkage for Big Data: Current approaches and research challenges. In: A.Y. Zomaya, S. Sakr (eds.) Handbook of Big Data Technologies, pp. 851–895. Springer (2017)
[621] van Veen, E.B.: Observational health research in Europe: understanding the General Data Protection Regulation and underlying debate. European Journal of Cancer **104**, 70–80 (2018)
[622] Verknüpfungsstelle: Verknüpfungsrichtlinen, Version 1.1. Tech. rep., Bundesamt für Statistik BFS, Bern (2017)
[623] Vernica, R., Carey, M.J., Li, C.: Efficient parallel set-similarity joins using mapreduce. In: ACM Conference on Management of Data, pp. 495–506. Indianapolis (2010)
[624] Verykios, V.S., Christen, P.: Privacy-preserving record linkage. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery **3**(5), 321–332 (2013)
[625] Verykios, V.S., George, M.V., Elfeky, M.G.: A Bayesian decision model for cost optimal record matching. VLDB Journal **12**(1), 28–40 (2003)
[626] Verykios, V.S., Karakasidis, A., Mitrogiannis, V.K.: Privacy preserving record linkage approaches. International Journal of Data Mining, Modelling and Management **1**(2), 206–221 (2009)
[627] Vidanage, A., Christen, P., Ranbaduge, T., Schnell, R.: A graph matching attack on privacy-preserving record linkage. In: ACM Conference on Information and Knowledge Management. Galway (2020)
[628] Vidanage, A., Ranbaduge, T., Christen, P., Randall, S.: A privacy attack on multiple dynamic match-key based privacy-preserving record linkage. International Journal of Population Data Science **5**(1) (2020)
[629] Vidanage, A., Ranbaduge, T., Christen, P., Schnell, R.: Efficient pattern mining based cryptanalysis for privacy-preserving record linkage. In: IEEE International Conference on Data Engineering. Macau (2019)
[630] Voigt, P., von dem Bussche, A.: The EU General Data Protection Regulation (GDPR): A Practical Guide. Springer, Cham (2017)
[631] de Waal, T., Pannekoek, J., Scholtus, S.: Handbook of Statistical Data Editing and Imputation. John Wiley & Sons, Hoboken (2011)
[632] Wagner, D., Lane, M.: The person identification validation system (PVS): Applying the center for administrative records research and applications' (CARRA) Record Linkage Software. Working Paper 2014-01, Center for Administrative Records Research and Applications, U.S. Census Bureau, Washington (2014)
[633] Wainwright, C., Fallon, L.: Evidence in policy making – Scottish Government sponsored data linkage projects. Government Statistical Service Conference (2018)

[634] Waldman, A.E.: Privacy as Trust: Information Privacy for an Information Age. Cambridge University Press, Cambridge (2018)
[635] Wan, Z., Vorobeychik, Y., Xia, W., Clayton, E.W., Kantarcioglu, M., Ganta, R., Heatherly, R., Malin, B.A.: A game theoretic framework for analyzing re-identification risk. PLOS One **10**(3), e0120,592 (2015)
[636] Wang, B., Song, W., Lou, W., Hou, Y.T.: Privacy-preserving pattern matching over encrypted genetic data in cloud computing. In: IEEE Computer Communications, pp. 1–9. Atlanta (2017)
[637] Weber, S.C., Lowe, H., Das, A., Ferris, T.: A simple heuristic for blindfolded record linkage. Journal of the American Medical Informatics Association **19**(1), 157–161 (2012)
[638] Weisberg, H.F.: The Total Survey Error Approach: A Guide to the New Science of Survey Research. The University of Chicago Press, Chicago (2005)
[639] Wen, Z., Dong, C.: Efficient protocols for private record linkage. In: ACM Symposium On Applied Computing, pp. 1688–1694. Gyeongju, Korea (2014)
[640] Westphal, C.: Data Mining for Intelligence, Fraud, and Criminal Detection: Advanced Analytics and Information Sharing Technologies. CRC Press, Boca Raton (2009)
[641] Whang, S.E., Menestrina, D., Koutrika, G., Theobald, M., Garcia-Molina, H.: Entity resolution with iterative blocking. In: ACM Conference on Management of Data, pp. 219–232. Providence, Rhode Island (2009)
[642] Winkler, W.E.: Using the EM algorithm for weight computation in the Fellegi-Sunter model of record linkage. Tech. Rep. RR2000/05, US Bureau of the Census, Washington, DC (2000)
[643] Winkler, W.E.: Record linkage. In: D. Pfeffermann, C. Rao (eds.) Handbook of Statistics, vol. 29, pp. 351–380. Elsevier (2009)
[644] Winkler, W.E.: Quality and analysis of sets of national files. In: Proceedings of the Section on Survey Research Methods, pp. 1432–1442. American Statistical Association (2014)
[645] Winkler, W.E., Thibaudeau, Y.: An application of the Fellegi-Sunter model of record linkage to the 1990 U.S. decennial census. Tech. Rep. RR1991/09, US Bureau of the Census, Washington, DC (1991)
[646] Winterleitner, A.D., Spichiger, A.: Personenidentifikatoren: Analyse der gesamtschweizerischen Kosten. In: J. Stember, W. Eixelsberger, A. Spichiger (eds.) Wirkungen von E-Government: Impulse für eine wirkungsgesteuerte und technikinduzierte Verwaltungsreform, pp. 383–424. Springer, Wiesbaden (2018)
[647] Witten, I.H., Moffat, A., Bell, T.C.: Managing Gigabytes, 2 edn. Morgan Kaufmann (1999)
[648] Wolfram, S.: A New Kind of Science. Wolfram Media, Champaign (2002)

[649] Woo, M.J., Reiter, J., Oganian, A., Karr, A.: Global measures of data utility for microdata masked for disclosure limitation. Journal of Privacy and Confidentiality **1**, 111–124 (2009)
[650] World Health Organization: WHO Guidelines on Ethical Issues in Public Health Surveillance. WHO, Geneva (2017)
[651] Xiao, C., Wang, W., Lin, X.: Ed-join: an efficient algorithm for similarity joins with edit distance constraints. VLDB Endowment **1**(1), 933–944 (2008)
[652] Xiao, C., Wang, W., Lin, X., Yu, J.X., Wang, G.: Efficient similarity joins for near-duplicate detection. ACM Transactions on Database Systems **36**(3), 1–41 (2011)
[653] Yakout, M., Atallah, M., Elmagarmid, A.: Efficient private record linkage. In: IEEE International Conference on Data Engineering, pp. 1283–1286. Shanghai (2009)
[654] Yancey, W.E.: Evaluating string comparator performance for record linkage. Tech. Rep. RR2005/05, US Bureau of the Census (2005)
[655] Yancey, W.E., Winkler, W.E., Creecy, R.H.: Disclosure risk assessment in perturbative microdata protection. In: J. Domingo-Ferrer (ed.) Inference Control in Statistical Databases, pp. 135–152. Springer (2002)
[656] Yao, A.C.C.: How to generate and exchange secrets. In: IEEE Symposium on Foundations of Computer Science, pp. 162–167. Toronto (1986)
[657] Yi, X., Paulet, R., Bertino, E.: Homomorphic Encryption and Applications. Springer (2014)
[658] Young, A., Flack, F.: Recent trends in the use of linked data in Australia. Australian Health Review **42**(5), 584–590 (2018)
[659] Yu, V.Y.: Promotion of global perinatal health. In: J. Ehiri (ed.) Maternal and Child Health, pp. 43–51. Springer, Boston (2009)
[660] Zezula, P., Amato, G., Dohnal, V., Batko, M.: Similarity Search: the Metric Space Approach. Springer Science & Business Media (2006)
[661] Zhang, L.C., Chambers, R.L.: Analysis of Integrated Data. CRC Press, Boca Raton (2019)
[662] Zheng, X., Cai, Z., Li, Y.: Data linkage in smart internet of things systems: A consideration from a privacy perspective. Communications Magazine **56**(9), 55–61 (2018)
[663] Zingmond, D., Ye, Z., Ettner, S., H., L.: Linking hospital discharge and death records – accuracy and sources of bias. Journal of Clinical Epidemiology **57**, 21–29 (2004)
[664] Zobel, J., Dart, P.: Phonetic string matching: Lessons from information retrieval. In: ACM Conference on Research and Development in Information Retrieval, pp. 166–172. Zürich (1996)
[665] Zomaya, A.Y., Sakr, S.: Handbook of Big Data Technologies. Springer (2017)

Index

Page numbers shown in bold refer to entries in the glossary.