UCLouvain
Institute for Information
and Communication Technologies,
Electronics and Applied Mathematics

Anonymization in the 21st Century: A Critical Examination of De-identification for Modern Large-Scale Data

Luc Rocher

Thesis submitted in partial fulfillment
of the requirements for the degree of
Ph.D. Degree in Engineering Sciences

Dissertation committee

Prof. Julien Hendrickx (UCLouvain, advisor)
Prof. Yves-Alexandre de Montjoye (Imperial College London, advisor)
Prof. Roland Keunings (UCLouvain, chairperson)
Prof. François-Xavier Standaert (UCLouvain)
Prof. Renaud Lambiotte (University of Oxford)
Prof. Bart Preneel (KU Leuven)

Louvain-la-Neuve, December 20th, 2019.

CONTENTS

1 ACKNOWLEDGMENTS 9

2 INTRODUCTION 11

I BACKGROUND AND PRINCIPLES

3 WHY ANONYMITY MATTER? 21

3.1 Information collected online and offline 22

3.1.1 Web search 22

3.1.2 Web browsing 22

3.1.3 Anonymous communications 23

3.1.4 Cryptocurrencies 24

3.1.5 Online fingerprinting 25

3.1.6 Mobile devices 25

3.1.7 DNA tests 26

3.2 Summary 26

4 FROM DE-IDENTIFICATION TO RE-IDENTIFICATION 29

4.1 Risk assessment 30

4.1.1 Data similarity metrics 30

4.1.2 Uncertainty metrics 31

4.2 How anonymity can be breached 32

4.2.1 Membership attacks 32

4.2.2 Inference attacks 33

4.3 Data transformation techniques 33

4.4 Summary 34

II MEASURING ANONYMITY WITH FEW OR NO DATA

5 ESTIMATING THE SUCCESS OF RE-IDENTIFICATIONS 37

5.1 Setting and model 38

5.1.1 A Gaussian copula model 38

5.1.2 Inferring marginals distributions 40

5.1.3 Inferring the copula parameters 41

5.1.4 Modeling the association structure 41

5.1.5 Theoretical and empirical uniqueness 42

5.1.6 Individual uniqueness and correctness 43

5.2 Empirical validation 45

5.2.1 Data availability 45

5.2.2 Model validation 46
5.2.3 Subsampling experiments 46
5.2.4 Measuring the model calibration 47
5.3 Results 49
5.3.1 Likelihood of successful re-identification 49
5.3.2 Appropriateness of the de-identification model 54
5.3.3 Correcting individual scores 57
5.3.4 Using exact marginals improves predictions 58
5.4 Related work 58
5.4.1 Population uniqueness methods 62
5.4.2 Log-linear models 65
5.5 Discussion 66
5.5.1 Testing for bias and heteroskedasticity 68
5.5.2 Sample unique records and uniqueness 69
5.5.3 Multivariate mutual information 71
5.5.4 Estimating population uniqueness at extremely small sampling fractions for large datasets 71
5.5.5 Evaluation of the model error 72
5.6 Summary 72
6 A THEORETICAL MODEL FOR ANONYMITY 77
6.1 Introduction 77
6.2 Modeling uniqueness and correctness 78
6.2.1 A Pitman-Yor model 78
6.2.2 Combining information theory and tail distributions 81
6.2.3 Anonymity bounds from summary statistics 85
6.2.4 Application to device fingerprints 87
6.2.5 Application to set-valued data 89
6.3 Summary 92
6.4 Additional methods 93
6.4.1 Uniqueness and correctness for Pitman-Yor processes 93
6.4.2 Entropy, uniqueness, and correctness for uniform distributions 95
6.4.3 Maximum a-posteriori (MAP) estimates from Pitman-Yor priors 95
6.4.4 Audit scenarios 97

III ENGINEERING PRIVACY

7 MODERN PRIVACY-PRESERVING PLATFORMS 101
7.1 Four models to release data 102
7.1.1 Limited Release 103
7.1.2 Precomputed Statistics 103
7.1.3 Remote Access 105
7.1.4 Interactive and decentralized computations 105
7.2 Privacy moving forward 107
7.2.1 Provable guarantees and heuristics 108
7.2.2 Practical defenses 109

8 CASE STUDY: EXPLOITING DIFFIX'S STICKY NOISE 111
8.1 The Diffix framework 112
8.1.1 Noise-exploitation attacks 116
8.1.2 Differential noise-exploitation attack 116
8.1.3 Cloning noise-exploitation attack 122
8.2 Experiments 129
8.2.1 Description of the datasets 130
8.2.2 Differential noise-exploitation attack 131
8.2.3 Cloning noise-exploitation attack 133
8.3 Discussion 135
8.3.1 Attribute predictability 135
8.3.2 Producing and detecting dummy conditions 136
8.3.3 Improving the attacks 136
8.4 Related work 137
8.5 Appendices 139
8.5.1 Likelihood ratio test 139
8.5.2 Cutoff threshold for the cloning attack 141
8.5.3 Size of value-unique classes and cloning attack on randomized dataset 143
8.5.4 Improving the accuracy and generalizing to non-binary attributes 145
8.5.5 Reducing the number of queries 147
8.5.6 Attacks on implementations of differential privacy 150
8.6 Summary 150

9 CASE STUDY: AGGREGATED MOBILE PHONE METADATA 153
9.1 Using the bandicoot toolbox 154
9.2 Project focus 155

9.3 Visualizing patterns in user data 157
9.4 Performance for large-scale datasets 157
9.5 Conclusion 158
10 CONCLUSION 161
Bibliography 163

LIST OF FIGURES

Figure 5.1 Estimating the population uniqueness of the USA corpus. 51

Figure 5.2 The model predicts correct re-identifications with high confidence. 54

Figure 5.3 The average individual uniqueness increases fast with the number of collected attributes 56

Figure 5.4 Cumulative distribution of the deviation between average individual likelihood and predicted population uniqueness. 58

Figure 5.5 False-discovery rate with original (left) and corrected (right) scores. 59

Figure 5.6 Reliability diagrams for copula methods. 60

Figure 5.7 Comparing the performance of our model against frequency-based re-identification risk models from the literature. 64

Figure 5.8 Poor calibration achieved by log-linear models. 66

Figure 5.9 Marginals determine most of the error at small sampling fraction. 73

Figure 5.10 KL divergence between empirical and estimated distributions. 74

Figure 5.11 The copula parameters converge quickly. 75

Figure 6.1 Pitman-Yor processes fit a wide range of discrete count distributions, and correctly model their uniqueness. 82

Figure 6.2 Using the Pitman-Yor model to quantify the criticality and regimes of uniqueness and correctness. 84

Figure 6.3 Little prior knowledge sufficiently bounds the resulting uniqueness of device fingerprints. 90

Figure 6.4 Empirical validation of the PY-SV model on the APPS dataset. 92

Figure 7.1 Four models for the privacy-conscientious use of data. 102

Figure 8.1 Diffix privacy-preserving architecture. 114
Figure 8.2 Accuracy of the differential attack. 132
Figure 8.3 Results of the differential attack 133
Figure 8.4 Results of the cloning attack 135
Figure 8.5 Influence of the variance cutoff σ^* on the accuracy of the cloning attack 143
Figure 8.6 Size of value-unique classes for users predicted attackable 143
Figure 8.7 Randomized results of the cloning attack. 144
Figure 8.8 Randomized results of the differential attack. 144
Figure 9.1 bandicoot's aggregation methods. 156
Figure 9.2 bandicoot's interactive visualization tool. 157
Figure 9.3 bandicoot's runtime performance. 158

1

ACKNOWLEDGMENTS

What makes a doctorate a doctorate? What may have I learned during all these years? There is, of course, the research. The scientific articles, conferences, meetings, presentations, talks, pauses café, coding coding coding experiments, typing typing typing on the keyboard. Nights writing equations by the flickering light of the LEDs to find them wrong in the morning. Rejection emails. Acceptance emails.

Il y a aussi ces rencontres qui pavent le chemin ou offrent des sentiers et détours fortuits. Ce doctorat et les années qui l'ont précédé n'auraient bien entendu pas été les mêmes sans celles et ceux que j'ai rencontrés durant et avant cette traversée.

Therefore, please find below this fuzzy list of names, people who taught me how to move forward and brought me joy every passing year. Thank you. If you are desperately searching for your name but do not find it here, I was limited not by space but by my memory. My imprecise recollection of the past may have forgotten your name as I write this page, and I am sorry for that.

Sans ordre d'apparition.

I would like to thank Julien Hendrickx and Yves-Alexandre de Montjoye, without whom none of this would have, of course, happened. So much accomplishment these past years thanks to you two. I learned a lot with you and keep learning. Renaud Lambiotte and François-Xavier Standaert, participating in my doctoral committee, and Bart Preneel, who joined them for the dissertation jury, brought their key expertise and helped shape this thesis with fruitful discussions and feedback. Finally, I am also grateful to Roland Keunings who chaired this jury with a great professionalism during the dissertation process.

À La Rochelle. Michel et Laurence pour votre soutien sans faille et vos encouragements, pour m'avoir éveillé le cœur, l'esprit et les yeux. *Damlaya damlaya göl olur.*

À Nantes. François Sauvageot, Bernard Luron et Yannick Alméras pour avoir, avec patience, tenté de forger en moi, et malgré toutes mes réticences, une pratique rigoureuse de la recherche scientifique. Cette thèse n'aurait pu aboutir sans vos enseignements précieux. Adrien, Édouard, Étienne, Gabriel, Orianne, Romain, Solen, Sophie, Vincent, et tant d'autres pour m'avoir chaleureusement soutenu dans cette aventure commune.

À Lyon. Julie et Céline bien sûr, que d'aventures à vos côtés. Jocelyn et Guillaume, votre présence me fut précieuse. Jill-Jênn, à la croisée des chemins. Alice, Alexandre, Amélie, Aurore, Bertrand, Léo, Margaux, Pierre, Rémi et tant d'autres, les années passent, mais les souvenirs de Lyon restent gravés profondément.

In Boston. Sandy, Yves-Alexandre, Bjarke, Dan, Dhaval, Florent, Oren and everyone at the Media Lab for making me feel welcome. A truly wonderful and vibrant environment. The Camberville diaspora, Molly, Val, Brian, Colin, Julian, Noah; Jesse, Sigurður, and everyone at Burnside; every instant with you is an eye-opening experience. Everyone at MITOC, for the wonderful distractions from the sea to the summits.

In London. To the best office mates: Ali, Ana-Maria, Andrea, Arnaud, Florimond, Shubham, and Thibaut.

À Louvain-la-Neuve. Matthew et Corentin bien entendu, acolytes au grand cœur et lurons de la première heure. Étienne, François W., Marie-Christine et Nathalie pour leur aide au quotidien. Mes collègues Adeline, Adrien S., Adrien T., Allan, Benjamin, Benoit, Corentin, Émilie, Estelle, François G., Matthew, Nicolas, Pierre-Alexandre, Pierre-Yves et Romain, il n'y a pas d'Euler sans vous.

À Bruxelles. Arielle, of course, you were always here for me. Arsène et Matthias au magnifique Liber, havre de calme au cœur de la cité. Delphine et Marie, pour m'avoir laissé le soin d'écrire loin des marteaux-piqueurs. Charly, pour m'avoir laisse le soin d'écrire près des marteaux-piqueurs. Kuntaka as Emmanuel as the Marquis of Monticello as =e=man>, your joyful curiosity is a balm to the soul; Julia for countless productive distractions; Nico for the life lessons. Corentin a third time, for countless coffee sessions. Eileen, Ella, Hugo, Julia, Koen, Nikita et tant d'autres pour cette communauté bienveillante et enjouée qui se reconnaitra !

2

INTRODUCTION

Calling and texting friends, searching online, running with smart watches, talking to virtual assistants, and tracking sleep patterns: Digital devices gather and share an increasingly detailed picture of our lives.

In the last decade, the ability to collect and store personal data has exploded. This collection is growing exponentially: 90% of all the data circulating on the Internet were created less than two years ago [1]. With two thirds of the world population having access to the Internet [2], electronic medical records becoming the norm [3], and the rise of the Internet of Things[1], this is unlikely to stop anytime soon. Collected at scale from financial or medical services, when filling in online surveys or liking pages, this data has an incredible potential for good. It drives scientific advancements in medicine [4], social science [5, 6], and AI [7], and promises to revolutionize the way businesses and governments function [8, 9].

At the same time, the large-scale collection and use of detailed individual-level data raise legitimate privacy concerns. The recent backlashes against the sharing of NHS [UK National Health Service] medical data with DeepMind [10] and the collection and subsequent sale of Facebook data to Cambridge Analytica [11] are the latest evidences that people are concerned about the confidentiality, privacy, and ethical use of their data. In a recent survey, more than 72% of U.S. citizens reported being worried about sharing personal information online [12]. In the wrong hands, sensitive data can be exploited for blackmailing, mass surveillance, social engineering[2], or identity theft. Misuse of data, and privacy breaches are on the rise, increasingly damage public trust in this economy. 91% of Americans and 80% of Europeans believe that they lost control over their data [13, 14].

[1] *The Internet of Things (IoT) refers to the billions of physical devices connected to the Internet worldwide, collecting and sharing data. These devices range from simple sensors to smartphones and wearables.*

[2] *Social engineering is the use of deception to manipulate individuals into performing actions or divulging confidential information, usually through technology.*

DIGITAL PRIVACY While the speed and amount of personal information collected nowadays is unprecedented, this collection is

nevertheless regulated through a growing series of privacy laws and cases worldwide.

The right to privacy is a fundamental right, upheld both by international and national human right laws. The Article 12 of the Universal Declaration of Human Rights, adopted in 1948, proclaims:

> "*No one shall be subjected to arbitrary interference with his privacy, family, home or correspondence, nor to attacks upon his honour and reputation. Everyone has the right to the protection of the law against such interference or attacks.*"

with similar provisions in the International Covenant on Civil and Political Rights[3] and the European Convention on Human Rights (ECHR)[4].

Yet privacy is a broad term, historically rooted at the intersection of personal rights, social rights, and technology, delineated by studies ranging from law reviews [15–18] to technical research in secure communications and security [19–21], and not a modern concern. In response to privacy-invading practices and technologies, Warren and Brandeis, *e.g.*, argued in 1890 for a broad legal protection of privacy, instituting "*the right to be left alone* as the foundations of individual freedom [22]:

> "*Recent inventions and business methods call attention to the next step which must be taken for the protection of the person, and for securing to the individual […] the right 'to be let alone'. […] Numerous mechanical devices threaten to make good the prediction that what is whispered in the closet shall be proclaimed from the house-tops.*"

At the US Supreme Court in 1928, Brandeis further argued for a constitutional right to privacy in Olmstead v. United States [23], dissenting from the majority opinion. The court later overturned the ruling in 1967, in the case Katz v. United States, and shaped the legal rights to privacy by defining a "*reasonable expectation of privacy*" and concluding that what an individual "*seeks to preserve as private, even in an area accessible to the public, may be constitutionally protected*" [24].

While these laws provide a general framework for privacy, there is a growing need for modern principles, standards and practices

[3] *The Article 17 of the ICCPR provides* "no one shall be subjected to arbitrary or unlawful interference with his or her privacy, family, home or correspondence".

[4] *The Article 8 of the ECHR provides* "Everyone has the right to respect for his private and family life, his home and his correspondence.".

to support the right to privacy in the digital age [25]. Recent data protection regulations specifically aim to strengthen our rights regarding the use of electronic devices and services as well as how digital personal data are shared and to whom. In Europe, the Data Protection Directive from 1995, followed by the E-Privacy Directive in 2002, and the recent General Data Protection Regulation in 2018 directly target digital privacy and created specific regimes for processing and disclosing personal data. In the US, digital privacy is not regulated at the federal level and lacks a comprehensive data protection framework. Few privacy laws govern specific sectors, such as health data[5] or data for children under thirteen[6]. Some states have passed further regulations such as the recent California Consumer Privacy Act (CCPA), aiming to protect privacy rights.

[5] *The Health Insurance Portability and Accountability Act (HIPAA), enacted in 1996, regulates health insurance and medical data sharing.*

[6] *The Children's Online Privacy Protection Act (COPPA), enacted in 1998, regulates to the online collection of personal information for children under thirteen.*

PERSONAL OR ANONYMOUS DATA? A core legal and technical issue beneath modern data protection regulations is: what data are *personal* data? Some regulations explicitly propose a ready-made list of information that could identify individuals. The HIPAA privacy rule, *e.g.*, lists 18 identifiers: name, address, dates, phone and fax numbers, email address, social security number, medical record number, health plan beneficiary number, account number, certificate or license number, any vehicle or other device serial number, Web URL, IP address, fingerprint or voice print, photographic image, and any other unique identifying number, characteristic, or code [26].

Regulators instead often opt for a conservative definition. In Europe, the recent General Data Protection Regulation (GDPR) provides in Article 4: "*an identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person*" [27].

Technically, the answer is continuously revisited as new types of data emerge and new methods to identify and track individuals are developed. Some data, such as the total population of the United States, are obviously non-personal. Some data, such as a social security number or a passport number, are obviously personal. The frontier between obvious anonymous and personal data is

thin and changes over time. For example, IP addresses, unique number assigned statistically or dynamically to every device online, have long been considered public non-personal information. The European Court of Justice instead ruled in 2017 that IP addresses constitute personal data, since ISPs or websites can combine an IP address with additional data to identify the device owner.

Drawing a line between anonymous and personal data is therefore a crucial and challenging task. The GDPR, *e.g.*, only regulates personal data, "*any information relating to an identified or identifiable natural person*" (Article 4). It does not apply to non-personal data, concluding in Recital 26 that "*The principles of data protection should therefore not apply to anonymous information, namely information which does not relate to an identified or identifiable natural person or to personal data rendered anonymous in such a manner that the data subject is not or no longer identifiable.*"

In 2015, Cambridge Analytica's detailed psychometric profiles, later used to target US voters, were collected "*anonymously*" from Facebook [28]. In 2016, journalists re-identified German public officials and exposed their medical information and sexual preferences, in the deemed-anonymous browsing habits of more than three million German citizens, sold by the Web of Trust (WoT) browser extension[29]. In 2017, a dataset of 123 million U.S. households was accidentally put online by Alteryx. The data comprised 248 attributes, including street addresses, demographics, finances and information pertaining to participants' children. Alteryx later claimed that the breached files were anonymous and did not "*pose a risk of identity theft to any consumers*" [30].

While
certain data are undisputedly not personal, researchers have now exposed a large body of seemingly innocuous, "anonymous" data that can actually be re-identified [31]. For example, demographics such as postal code, date of birth, or gender alone are not unique enough to identify individuals, yet together often form a unique fingerprint [32, 33]. In the recent field of *re-identification science* [18], multiple empirical studies have demonstrated how few external information can help identify individuals inside demographics data, mobility traces, credit card transactions, surveys of installed mobile apps and movie ratings, or genome wide association studies [34–41]. Researchers have also showed how to train statistical models, from

disclosed sample of records, to estimate the population uniqueness, the fraction of uniquely identifiable individuals in the complete population [32, 33, 42–44].

Legally, there should be little ambiguity. Modern data protection regulations clearly distinguish simple data pseudonymization from true anonymization. Three conditions for a dataset to be anonymous have been set forth by the EU's Working Party 29: one cannot (*i*) single out an individual, (*ii*) link records relating to an individual, and (*iii*) infer information concerning an individual [45]. Pseudonymization, simply replacing all identifiers (*e.g.*, phone number, names, email address) by pseudonyms so that "*the data can no longer be attributed to a specific data subject without the use of additional information*" [27], does not necessarily yield anonymous data.

In response to these concerns, this thesis constitutes, firstly, a critical review of how data anonymization practices are currently evaluated, spanning from disclosed medical and demographics microdata to high-dimensional datasets, as well as methods to release aggregated statistics instead of the raw data. This thesis builds, secondly, new models and frameworks to estimate the anonymity of data collections and suggests modern ways to release sensitive personal data while limiting privacy breaches.

OUTLINE AND CONTRIBUTION The present introduction is followed by an overview of the research on anonymity (Chapter 3), with a specific focus on de-identification and re-identification of personal data (Chapter 4).

Chapter 5 provides a critical response to the current framework used to de-identify and anonymize personal data. Anonymizing datasets through de-identification and sampling before sharing them has been the main tool used to address the privacy concerns associated with sharing sensitive socio-demographic, medical, or financial data. We propose a generative copula-based method that can accurately estimate the likelihood of a specific person to be correctly re-identified, even in a heavily incomplete dataset. Our results suggest that even heavily sampled anonymized datasets are unlikely to satisfy the modern standards for anonymization set forth

by GDPR and seriously challenge the technical and legal adequacy of the de-identification release-and-forget model.

Chapter 6 proposes a theoretical model for anonymity. Without legal or practical access to the data, auditing data collections to verify whether the data are really anonymous is impractical. This is especially true for present and future collections where no collected data is available. First, using a Bayesian nonparametrics approach, we propose a general model, the PY model, that governs the uniqueness of individuals and the correctness of re-identifications. We validate its universality across a wide range of data collections, from tabular data (census and surveys) to high-dimensional set-valued data (mobile app usage, mobile phone and shopping patterns) showing that, in each case, we accurately predict the re-identification patterns. Second, combining few information-based statistics and linear optimization methods on the entropic cone, we show how data owners, data protection authorities, and policymakers can now accurately evaluate whether specific data collections will be anonymous or not.

Chapter 7 discusses the present and future privacy-conscientious solutions for data controllers to collect and disclose insights from personal data, without breaching participants' anonymity, highlighting the research directions requiring further work. In Chapter 8, we demonstrate the concerns associated with such modern privacy-preserving platforms, using Diffix, a privacy-preserving database, as an example, calling for better theoretical and practical guarantees for releasing personal data. In Chapter 9, we finally show how aggregated insights can be derived from sensitive personal data using bandicoot, a toolbox for mobile phone metadata, as an example.

This thesis contains work from publications I co-authored during the past four years, including:

- Yves-Alexandre de Montjoye, Luc Rocher, and Alex Sandy Pentland. bandicoot: a Python Toolbox for Mobile Phone Metadata, *J. Mach. Learn. Res.* **17**, n. 175 (2016), pp. 1–5.

- Yves-Alexandre de Montjoye, Ali Farzanehfar, Julien M Hendrickx, and Luc Rocher. Solving artificial intelligence's

privacy problem, *Field Actions Science Reports* n. 17 (2017), pp. 80–83.

- Yves-Alexandre de Montjoye, Sébastien Gambs, Vincent Blondel, Geoffrey Canright, Nicolas de Cordes, Sébastien Deletaille, Kenth Engø-Monsen, Manuel Garcia-Herranz, Jake Kendall, Cameron Kerry, Gautier Krings, Emmanuel Letouzé, Miguel Luengo-Oroz, Nuria Oliver, Luc Rocher, Alex Rutherford, Zbigniew Smoreda, Jessica Steele, Erik Wetter, Alex Sandy Pentland, and Linus Bengtsson. On the privacy-conscientious use of mobile phone data. en, *Sci Data* **5** (Dec. 2018), p. 180286.

- Andrea Gadotti, Florimond Houssiau, Luc Rocher, Benjamin Livshits, and Yves-Alexandre de Montjoye. "When the Signal is in the Noise: Exploiting Diffix's Sticky Noise", *USENIX Security Conference Proceedings*. 2019.

- Luc Rocher, Julien M Hendrickx, and Yves-Alexandre de Montjoye. Estimating the success of re-identifications in incomplete datasets using generative models. en, *Nat. Commun.* **10**, n. 3069 (July 2019).

Andrea Gadotti, Florimond Houssiau, and myself contributed equally to the research.

as well as current research projects not published at the moment.

Part I

BACKGROUND AND PRINCIPLES

3

WHY ANONYMITY MATTER?

In the digital age, an ever increasing portion of our lives is being routinely recorded [51]. While in 1790, the first U.S. census asked only four questions [52], modern data collection from governmental and private-sector organizations now produce detailed individual profiles of our behaviours. US governmental agencies maintain more than 2,000 databases of personal data [53]. The data broker Acxiom claims to have files on 10% of the population worldwide, with about 1,500 pieces of information per consumer [54]; Experian, an international credit rating agency, is also known to have gathered fine-grained demographic data on over 90% of the US population and over 80% of the UK population [55]. This large scale hoarding of personal data is driven by an ever growing desire to a fine grained understanding of customers and citizens, allowing targeted intervention and marketing [51].

Personal and behavioural information can now be collected through a variety of means. Data can be aggregated from third parties, such as banks, credit rating agencies, and health care or insurance companies. Data can be collected actively, by answering surveys and census, or uploading photos and texts to social media. It can also be collected passively—typically without the person's knowledge or explicit consent—, such as with location traces from the metadata of photos uploaded to social media, mobile phone records (CDR and xDR), shopping patterns in loyalty or credit cards, or voting profiles when using apps, "liking" pages or tweets, or simply online browsing patterns available to ISPs, websites, or eavesdroppers [54, 56].

The unbridled collection of our personal data by a multitude of agents is not without its failures. In the recent years, the number of data leaks either incidental or provoked has exploded (3 billion user accounts from Yahoo, 500 million passport and credit card details from Marriott, 143 million demographic profiles of US residents from Equifax, 540 million records from Facebook users, and many more) [57–59]. With the ever increasing granularity of the data,

every leak leads to an increasing risk of blackmail, discrimination, harassment, and financial or identity theft. Joined together, this mass of data—the "database of ruin" as coined by Paul Ohm—gives malicious entities access to detailed digital fingerprints of our behaviours, leading to easy re-identification of personal data [18].

3.1 INFORMATION COLLECTED ONLINE AND OFFLINE

Past and current research on tracking and fingerprinting have exposed how our digital behaviours leave countless breadcrumbs behind, which together can form a detailed profile of virtually everyone [18]. In this section, we review some of the most important sources and collections of sensitive behavioural information. We also discuss the mitigation strategies that have been proposed to reduce, but often not suppress, the risk of exposure.

3.1.1 *Web search*

Barbaro and Zeller showed in 2006 how to link few AOL search queries to individuals and released publicly the identity of a correctly re-identified user, Thelma Arnold [60]. Collected at scale, search queries can reveal the demographics, financial situation, and medical conditions of users [61, 62]. A number of techniques have been proposed to address the problem of (online) search privacy, including proxies and anonymizing networks such as Tor [63], or query obfuscation [64, 65]. However, none of the obfuscators proposed in the literature currently provide an acceptable privacy and some can even be thwarted by simple adversarial approaches [64, 66].

3.1.2 *Web browsing*

The collection of personal information online does not stop at search queries. Modern civilian communications network are by design very open: Internet is a public network, and routing information is easily accessible to passive observers. Eavesdroppers, such as governments and ISPs, can monitor who exchanges packets with whom, at scale using, for instance, Deep Packet Inspection [67].

Secure encrypted protocols, such as HTTPS (a encrypted layer over HTTP) have been designed to hide the content of communications. However, researchers have shown that it is still possible to infer information about the web page or the content. Specific webpage often generate a unique fingerprint in term of page size, network packets, *etc*. This has been used to identify web pages with high likelihood and is still hard to protect against [68–71]. Side-channel attacks on HTTPS compression can further help identify protocols, parts of the communication [72–75], or even VoIP conversations [76].

3.1.3 *Anonymous communications*

Protocols such as HTTP(S) do not hide the identity of the sender and receiver of a communication, therefore revealing sensitive information to an eavesdropper (e.g., who connects to which online platform, when, where, *etc*.). In response, systems to achieve anonymous communications have been developed to protect not only the content, but also the route of the communication, and conceal the identity of the sender and receiver of a network packet. They include a wide range of techniques, from proxies and VPNs (one trusted intermediary) to mix network (a complex network of relays obfuscating the path of packets, such that no single relay knows both source and destination of a packet) [63, 77, 78]. Based on onion routing (a variation of mix networks, using circuit-based routing), Tor is for instance used by 2-3M people daily[1] to connect to the Internet anonymously.

However, different techniques have been designed to identify one's communications, even if they use such anonymizing systems. Proxies and VPNs, besides abiding to the legal framework of their registration country, are also vulnerable to the same side-channel attacks we mentioned above: identifying websites or web pages using unique fingerprints, or decoding parts of the content.

For mix networks, the first entry points, the gateways, are often public or easy to identify. Once gateways are known, different side channel attacks have been proposed to track packets, either using Sybil relays [79], sets of relays controlled by the same malicious

1 The Tor Project reports the estimated number of directly-connecting clients each day at https://metrics.torproject.org/userstats-relay-country.html.

entity, or leveraging routing information from outside the network. An attacker can observe or manipulate latency and congestion to link packets on both side of the network [80–82]. They can correlate TCP traffic based on few compromised nodes in or outside the network [83–85]. Furthermore, while DNS requests are now routed through the Tor network, entities such as Google's DNS resolver, observing almost 40% of all DNS requests exiting the Tor network, could use DefecTor attacks [86] to determine the website that a Tor user is visiting. Finally, passive powerful attackers, with control of a large portion of the network, will always be able to identify users of mix networks [84]. While running large-scale Sybil attacks on the Tor network—and breaking its anonymity—is supposedly hard and costly, attacks to identify[7] and compromise[8] Tor users have been successfully launched by organizations such as the NSA in the past [87].

[7] *XKEYSCORE, a secret system developed by the NSA, continually monitors and analyzes global Internet data, including tracking Tor users worldwide.*

[8] *FOXACID, a software developed by the NSA, send exploits from a compromised server to a targeted web browser.*

3.1.4 *Cryptocurrencies*

Another class of anonymity networks, cryptocurrencies, have raised concerns about the confidentiality of transactions. By design, cryptocurrencies such as Bitcoin log the complete history of all transactions in a public registry. By linking addresses with known identities, it becomes possible to identify individuals' transactions or purchases, by using online transactions [88] or crawlers that search social media for BTC addresses [89].

To protect against such easy re-identifications, cryptocurrency users may employ multiple addresses for different purposes, but researchers have shown that heuristics can easily identify all the addresses used by the same individual [90, 91]. It is even possible to link the Bitcoin users' public keys to their IP addresses with an accuracy ranging from 11 to 60% depending on the attack scenario [92].

Coin mixing services have been proposed to prevent Bitcoins from being traced, but have shown to be easily thwarted [88]. In response to this concerns, cryptocurrencies such as ZeroCoin or Monero have been developed to provides anonymity by design using zero-knowledge proofs [93] yet a recent study showed that Monero's "mixin sampling strategy" might be vulnerable to "chain-reaction" de-anonymization attacks [94].

3.1.5 *Online fingerprinting*

On a digital device, private browsing mode and extensions were thought to effectively block or delete cookies, and prevent websites from assigning a unique identifier to a device [95]. Yet many browser fingerprinting techniques have been devised to uniquely track devices without cookies. These techniques use the rich Javascript APIs, HTTP headers, WebGL and GPU features, or installed plugins [96–99]. Disabling scripts, audio, or canvas fingerprinting [100], spoofing the behaviour of other devices [99], or techniques to randomize fingerprints [101] have been proposed to limit web profiling. Vastel et al. showed that state-of-the-art fingerprinting countermeasures can still be classified as genuine or altered, potentially reducing the impact of such measures [95]. However, Gómez-Boix et al. also showed that browser fingerprints are actually less unique than reported in the past, reducing the effectiveness of device identification [102], and questioning the effectiveness of such methods.

3.1.6 *Mobile devices*

Even without a cookie or device fingerprint, individuals can still be identified by their behavioural usage of digital tools or services. The way we move around [36], purchase goods [37], or use mobile apps [41] render us unique in very large populations. de Montjoye et al. showed that four data points—approximate places and times where an individual was present—were enough to uniquely re-identify 95% of users in a mobile phone dataset of 1.5M people [36] and 90% from the credit card transaction records of 1.1M individuals [37]. The simple combination of home and work location is often sufficient to correctly identify individuals [103]. Researchers have shown that mobile phones pose further privacy concerns: installed Android or iOS apps uniquely identify most mobile devices [41, 104]; the devices leak many unique identifiers, such as their MAC address (even with MAC address randomization) [105], and broadcast a partial list of previously connected WiFi access points, which can be used to uniquely identify them and track their past locations [106].

Information can be further collected from a wide range of other electronic devices: unique driving style can help identify drivers [107]; poor security practices let Internet of Things (IoT) devices, home appliances, and CCTV cameras openly accessible online [108]; smart meters, providing high resolution usage data to electricity or gas providers, reveal sensitive details about users [109, 110] and require complex privacy-preserving methods to protect participants' data [111, 112].

3.1.7 *DNA tests*

Finally, tracking and fingerprinting can even happen physically. Consumer ancestry tests have become a common exercise, with more people taking tests in 2017 than in all previous years combined. An estimated 12 million tests having been performed by companies such as Ancestry.com or 23andMe [113]. While this has opened the door to new ventures for paternity tests or personalised medicine, it has also led to increasing concerns for the privacy risks associated with the data [114]. Recent research has shown that, contrary to commonly held beliefs, common aggregation techniques still leave the data vulnerable to linkage attacks [115, 116]. Even advanced mechanisms to release data, which we discuss below, can suffer from scalability or usability issues when applied to genome-wide association studies [117].

3.2 SUMMARY

Overall, the recent years have seen an increased awareness of the risks associated with collecting personal data, both from the public, policymakers, and researchers. Seemingly innocuous data, combined together, form precise profiles of individuals and help uniquely identify them in future data collections.

However, solutions exist to protect anonymity while collecting data. At the collection level, limited amount of data can be collected and data can be obfuscated or aggregated to thwart re-identification. To protect against eavesdroppers and control who access the data, better privacy-by-design mechanisms can be employed such as anonymity network. Finally, there is no such thing as risk zero in

privacy; solutions also come from trust in secure platforms, privacy developers and engineers, and technologies.

4

FROM DE-IDENTIFICATION TO RE-IDENTIFICATION

De-identification is a broad term used by academics, regulators, and industry, often with different interpretations. De-identified or non-personal data generally refers to "*data that is not linkable to either an individual or a device*" [118] or not "*reasonably linkable*", according to the FTC's guidelines [119]. In Europe, the General Data Protection Regulation (GDPR) strictly differentiates anonymized data, where identification is no longer possible, and pseudonymized data, which replaces attributes to reduce the linkability of a dataset [45].

De-identification methods have been extensively used to prevent an individual's information from being linked with them, when collecting, storing, or releasing their data [18]. The US Health Insurance Portability and Accountability Act (HIPAA) exempts researchers who de-identify medical data from numerous strict conditions, since the data is not personal anymore [120]. Similarly, in Europe, anonymized data as ruled by the GDPR require no explicit consent or authorization to be used and disseminated to third parties [27, 45].

De-identification research traditionally falls in two categories: firstly, measuring the likelihood to re-identify an individual or infer information about them from de-identified data; secondly, designing heuristics to transform personal data in order to render it anonymous [121]. These tasks have shown to be challenging, as privacy violations are difficult to quantify. Fully provable anonymization guarantees are typically hard to obtain without destroying a significant part of the usability [122, 123], resulting in heuristics and approximate guarantees.

Note. Modern de-identification techniques, including releasing synthetic data or differential privacy are discussed in Chapter 7. This thesis focuses on the de-identification and re-identification of individual-level data; howevever, a wide range of privacy-enhancing techniques have been proposed, including secure multi-

party computation, homomorphic encryption, federated learning, or trusted execution environment. These topics fall outside of the scope of this thesis.

4.1 RISK ASSESSMENT

When releasing personal data, two approaches are used to measure the risk to re-identify participants:

1. similarity metrics capture the similarity of records in the released data,

2. uncertainty metrics capture the likelihood of false-positive matches in the complete population where the data comes from.

4.1.1 *Data similarity metrics*

Data similarity metrics [124] measure the indistinguishability of individuals in de-identified data. Intuitively, they ensure that for every person in an "anonymous" dataset, enough other people share their characteristics or are close enough to conceal the person's identity.

k-anonymity, introduced by Latanya Sweeney in 1998 [125], measures the highest number *k* such as identifying someone in the data always lead to at least *k* potential matches. The larger *k* is, the most difficult it becomes to identify individuals. While *k*-anonymity is widely used to share medical data [126], location data [127], or census data [128]), researchers have since shown that it still exposes individuals to risks of inference, and consequently developed a large range of alternative metrics to protect against such inference [129]. *k*-anonymity does not forbid inference attacks:

l-diversity [130] extends *k*-anonymity to ensure that in each equivalence class every attribute has at least *l* different other values. *t*-closeness [131] refines *l*-diversity, aiming at obtaining equivalence classes that match the initial distribution of other attributes in the data.

Different attacks have been proposed against such heuristics:

HOMOGENEITY ATTACKS Assume that three records share Bob's characteristics in a medical dataset, all three are associated with an HIV positive diagnostic. An adversary might not re-identify Bob's record yet will correctly identify Bob's diagnostic. Here, 3-anonymity does not provide any guarantees against an homogeneity attack.

BACKGROUND KNOWLEDGE ATTACK Males cannot be diagnosted with endometriosis. In a medical dataset, Bob's identity matches three potential records, two with endometriosis and one with a stage IV cancer. An adversary, armed with Bob's characteristics, will therefore learn that Bob has a stage IV cancer. Here, neither 3-anonymity nor 2-diversity will protect Bob's medical diagnostic against a background knowledge attack.

Overall, these measures are widely used in practice [132], including in de-identification softwares [133]. Yet researchers argue that they are only heuristics, falling short of providing any provable privacy protection [122] and cannot anonymize high-dimensional data without significant deterioration of the information content [36, 37, 134].

4.1.2 *Uncertainty metrics*

Uncertainty metrics, on the other hand, are probabilistic measures of anonymity, attempting to quantify the anonymity of an individual, their likelihood to be hidden in the overall population. They measure the probability to correctly identify someone based on the uniqueness of their characteristics.

Pfitzmann and Hansen define anonymity as "*the state of being not identifiable within a set of subjects*", the anonymity set [135]. Multiple metrics aim to capture this notion, by using the frequency distribution[9] of the anonymity set [136, 137]: large frequency classes yield high uncertainty, small frequency classes high risk of re-identification. The entropy of the distribution of anonymity set size is used to measure the uncertainty of a re-identification [138–140], as well as the degree of anonymity, the worst-case entropy [136].

[9] *These frequencies are also called anonymity set sizes.*

Another widely used metric, the uniqueness ε, measures the fraction of individuals with an anonymity set size of one, *i.e.* who can

be uniquely identified in the population. It can be calculated exactly if all data are available, or estimated using statistical methods [32, 44]. The uniqueness has been used to measure the re-identification of demographics data [32, 33], home and work location data [103], health data [141], or genealogical trees [142]. Finally, for set-valued datasets, where each record is associated with a set of data points (*e.g.*, mobility traces, shopping cart items, credit card transactions), a relaxation of uniqueness—the unicity ε_p—measures the fraction of records for which p points from their trace uniquely identify them. The unicity has been used to measure risks in mobile-phone location data, shopping patterns, or mobile app usage [36, 37, 41].

While both the uniqueness and the unicity can be used to make precise comparisons, there is no theoretically justified threshold for acceptable values, though arbitrary values such as $\varepsilon = 0.05$ or $\varepsilon = 0.1$ are often used [122].

4.2 HOW ANONYMITY CAN BE BREACHED

Most breaches of anonymity when collecting, using, or sharing personal data fall in two categories: membership attacks and inference attacks.

4.2.1 *Membership attacks*

In a membership attack, an adversary is able to test whether a specific person participated in a data collection or belongs to a database, hence the term "membership".

Membership attacks are significant privacy issues, *e.g.*, if the data comprises medical health record with one specific condition or records from an abortion clinic. In 2008, researchers proved how to test the presence or absence of individuals in a public genome-wide association study (GWAS), from aggregated allele frequency statistics [38]. Their watershed work, based on hypothesis testing on exact statistics, was then validated [143–145] and improved by taking into account individual genotypes [39] as well as correlations within the human genome [146]. It was also extended to microRNA data [147] or smart-meter data [148].

Dwork et al. have shown that membership attacks are essentially always possible as long as enough attributes are released, even if these are perturbed with a significant amount of noise [149].

4.2.2 *Inference attacks*

Inference or reconstruction attacks, where the adversary is able to single out an individual and infer their sensitive attributes, *e.g.*, political affiliation or medical condition, represent a further failure of aggregate data. For instance, the single-nucleotide polymorphism (SNP) sequences of the participants in the GWAS database studied by Homer et al. can actually be identified for hundreds of participants [146] from the published statistics; this attack works even with rounded and partial statistics.

More recently, researchers developed membership and inference attacks on aggregate mobility data. Using bespoke human mobility models, researchers managed to reconstruct the mobility patterns of single individuals [150, 151]. They also inferred whether a user's trace contributed to the aggregates based on simple machine learning classifiers such as the logistic regression or random forests [150, 152]. In doing this, they showed how different types of background knowledge available to the adversary impact the attack accuracy. In their latest work, the authors study the effectiveness of several defense strategies including differential privacy and show that while they can be effective, they usually come at a heavy price on the utility.

4.3 DATA TRANSFORMATION TECHNIQUES

To reduce the measured re-identification risks, the process of de-identifying data typically involves a combination of data transformation techniques to redact or coarsen the data [141, 153].

In 2011, Matthews and Harel reviewed the standard methods used to de-identify data [121]. They include the suppression of directly or indirectly identifying attributes (*e.g.*, deemed if too sensitive or informative); the suppression of records at random (subsampling) or following risk heuristics [154]; the suppression of specific cells [155, 156]. They also include perturbation methods: perturbing values by adding noise, rounding, or masking [141,

157, 158]; coarsening the collected data [159]; swapping cells or partially shuffling variables to add uncertainty while keeping the marginal distributions unchanged. Finally, hashing can be used to mask variables and create pseudonyms [160].

Combinations of these techniques are often used in order to achieve a low re-identification risk, for instance a given k for k-anonymity, deemed sufficient. Since achieving k-anonymity yet maximizing utility is NP-hard, greedy and stochastic heuristics have been developed to obtain k-anonymity while minimizing changes to the data [126, 161–163].

4.4 SUMMARY

The research on membership and inference attacks has had an immediate impact on data sharing practices for aggregate data. In 2009, considering the vulnerabilities of genomic statistics [38], research institute worldwide, such as the US National Institutes of Health, the Broad Institute, and the Wellcome Trust, decided to apply more restrictive access policies to previously publicly available summary data [164]. These measures encountered criticism from the research community, concerned that they "*could hamper data sharing, which has facilitated so many discoveries*" [164].

Strong concerns have been raised about the efficacy in general of the de-identification framework [18, 31, 122, 123, 165]. Numerous de-identified datasets have been re-identified in the past years, *e.g.*, the NIGMS and NIH genetic data in 2013 [40]; the Washington State Health Data in 2013 [34]; the NYC Taxicab dataset in 2014 [166]; the Transport For London bike sharing dataset in 2014 [167]; the Australian de-identified Medicare Benefits Schedule (MBS) and Pharmaceutical Benefits Schedule (PBS) datasets in 2017 [168]. In an internal study, the US Census Bureau was able to re-identify a large fraction of its de-identified data and urgently called for new provable methods to protect participants' data [169]. Even proponents of the traditional de-identification framework admit its failure to protect modern high-dimensional data [170]. Overall, the lack of fundamental guarantees for de-identification techniques, the unprecented scale of data being collected and released, and the wide range of re-identified data challenge the technical and legal adequacy of the de-identification release-and-forget model.

Part II

MEASURING ANONYMITY WITH FEW OR NO DATA

5

ESTIMATING THE SUCCESS OF RE-IDENTIFICATIONS

This chapter includes work from our article "Estimating the success of re-identifications in incomplete datasets using generative models" [50].

De-identification, the process of anonymizing datasets before sharing them, has been the main paradigm used in research and elsewhere to share data while preserving people's privacy [120, 171, 172]. Yet numerous supposedly anonymous datasets have recently been released and re-identified [18, 29, 36, 37, 122, 167, 168, 173–177].

Statistical disclosure control researchers and some companies are disputing the validity of these re-identifications: as datasets are always incomplete, journalists and researchers can never be sure they have re-identified the right person even if they found a match [121, 178–180]. They argue that this provides strong plausible deniability to participants and reduce the risks, making such de-identified datasets anonymous including according to GDPR [153, 181–183]. De-identified datasets can be intrinsically incomplete, *e.g.*, because the dataset only covers patients of one of the hospital networks in a country, or because they have been subsampled as part of the de-identification process. For example, the U.S. Census Bureau releases only 1% of their decennial census and sampling fractions for international census range from 0.07% in India to 10% in South American countries [184]. Companies are adopting similar approaches with, *e.g.*, the Netflix Prize dataset including less than 10% of their users [185].

Imagine a health insurance company who decides to run a contest to predict breast cancer and publishes a de-identified dataset of 1,000 people, 1% of their 100,000 insureds in California, including people's birth date, gender, ZIP code, and breast cancer diagnosis. John Doe's employer downloads the dataset and finds one (and only one) record matching Doe's information: male living in Berkeley, CA (94720), born on January 2nd, 1968, and diagnosed with breast cancer (self-disclosed by John Doe). This record also contains the details of his recent (failed) stage IV treatments. When contacted, the insurance company argues that matching does not equal re-

identification: the record could belong to one of the 99,000 other people they insure or, if the employer does not know whether Doe is insured by this company or not, to anyone else of the 39.5M people living in California.

In this chapter, we show how the likelihood of a specific individual to have been correctly re-identified can be estimated with high accuracy even when the anonymized dataset is heavily incomplete. We propose a generative graphical model that can be accurately and efficiently trained on incomplete data. Using socio-demographic, survey, and health datasets, we show that our model exhibits a mean absolute error (MAE) of 0.018 on average in estimating population uniqueness [33] and an MAE of 0.041 in estimating population uniqueness when the model is trained on only a 1% population sample. Once trained, our model allows us to predict whether the re-identification of an individual is correct with an average false-discovery rate of less than 6.7% for a 95% threshold ($\widehat{\xi_x} > 0.95$), and an error rate 39% lower than the best achievable population-level estimator. With population uniqueness increasing fast with the number of attributes available, our results show that the likelihood of a re-identification to be correct, even in a heavily sampled dataset, can be accurately estimated and is often high. Our results reject the claims that, first, re-identification is not a practical risk and, second, sampling or releasing partial datasets provide plausible deniability. Moving forward, they question whether current de-identification practices satisfy the anonymization standards of modern data protection laws such as GDPR and CCPA, and emphasize the need to move, from a legal and regulatory perspective, beyond the de-identification release-and-forget model.

5.1 SETTING AND MODEL

5.1.1 *A Gaussian copula model*

We consider a dataset $\mathcal{D}$, released by an organization, and containing a sample of $n_{\mathcal{D}}$ individuals extracted at random from a population of n individuals, *e.g.*, the US population. Each row $x^{(i)}$ is an individual record, containing d nominal or ordinal attributes (*e.g.*, demographic variables, survey responses) taking values in a discrete sample space $\mathcal{X}$. We consider the rows $x^{(i)}$ to be independent

and identically distributed (i.i.d.), drawn from the probability distribution X with $\mathbb{P}(X = x)$, abbreviated $p(x)$.

Our model quantifies, for any individual x, the likelihood ξ_x for this record to be unique in the complete population, and therefore always successfully re-identified when matched (nullifying any claim of anonymity in any released "anonymous" data). From ξ_x, we derive the likelihood κ_x for x to be correctly re-identified when matched, which we call correctness. If Doe's record $x^{(d)}$ is unique in $\mathcal{D}$, he will always be correctly re-identified ($\kappa_{x^{(d)}} = 1$ and $\xi_{x^{(d)}} = 1$). However, if two other people share the same attribute ($x^{(d)}$ not unique, $\xi_{x^{(d)}} = 0$), Doe would still have one chance out of three to have been successfully re-identified ($\kappa_{x^{(d)}} = 1/3$). We model ξ_x as:

$$\xi_x \equiv \mathbb{P}\big(x \text{ unique in } (x^{(1)}, \dots, x^{(n)}) \mid \exists i, x^{(i)} = x\big) \tag{5.1}$$

$$= \big(1 - p(x)\big)^{n-1} \tag{5.2}$$

and κ_x as:

$$\kappa_x \equiv \mathbb{P}\big(x \text{ correctly matched in } (x^{(1)}, \dots, x^{(n)}) \mid \exists i, x^{(i)} = x\big) \tag{5.3}$$

$$= \frac{1}{n} \frac{1 - \xi_x^{n/(n-1)}}{1 - \xi_x^{1/(n-1)}} \tag{5.4}$$

with proofs in Section 5.1.6.

We model the joint distribution of $X_1, X_2, \dots X_d$ using a latent Gaussian copula [186]. Copulas have been used to study a wide range of dependence structures in finance [187], geology [188], and biomedicine [189] and allow us to model the density of X by specifying separately the marginal distributions, easy to infer from limited samples, and the dependency structure. For a large sample space $\mathcal{X}$ and a small number $n_{\mathcal{D}}$ of available records, Gaussian copulas provide a good approximation of the density using only $d(d-1)/2$ parameters for the dependency structure and no hyperparameter.

The density of a Gaussian copula C_Σ is expressed as:

$$c_\Sigma(u) = \frac{1}{\sqrt{\det \Sigma}} \exp\Big(-\frac{1}{2}\, \Phi^{-1}(u)^T \cdot \big(\Sigma^{-1} - \mathrm{I}\big) \cdot \Phi^{-1}(u)\Big) \tag{5.5}$$

with a covariance matrix Σ, $u \in [0,1]^d$, and Φ the CDF of a standard univariate normal distribution.

Notably, Gaussian copulas take into account marginal distributions and pairwise associations but ignore higher order association patterns. While other generative model could be used, our experiments below show that Gaussian copulas provide a robust model of the discrete distributions studied here in order to estimate uniqueness.

We estimate from $\mathcal{D}$ the marginal distributions Ψ (marginal parameters) for $X_1, \ldots, X_d$ and the copula distribution Σ (covariance matrix), such that $p(x)$ is modeled by

$$q(x|\Sigma, \Psi) = \int_{F_1^{-1}(x_1-1|\Psi)}^{F_1^{-1}(x_1|\Psi)} \cdots \int_{F_d^{-1}(x_d-1|\Psi)}^{F_d^{-1}(x_d|\Psi)} c_\Sigma(u)\, \mathrm{d}u \qquad (5.6)$$

with F_j the CDF of the discrete variable X_j. In practice, the copula distribution is a continuous distribution on the unit cube, and $p(x)$ its discrete counterpart on $\mathcal{X}$.

We select, using maximum likelihood estimation, the marginal distributions from categorical, logarithmic, and negative binomial count distributions (see Section 5.1.2). Sampling the complete set of covariance matrices to estimate the association structure of copulas is computationally expensive for large datasets. We rely instead on a fast two-steps approximate inference method: we infer separately each pairwise correlation factor Σ_{ij} and then project the constructed matrix Σ on the set of symmetric positive definite matrices to accurately recover the copula covariance matrix (see below).

5.1.2 *Inferring marginals distributions*

Marginals can be either (*i*) unknown and are estimated from the marginals of the population sample X_S, this is the assumption used in the main text, or (*ii*) known with their exact distribution and cumulative density function directly available.

In the first case, we fit marginal counts to categorical (naive plug-in estimator), negative binomial, and logarithmic distributions using maximum log-likelihood. We compare the obtained distributions and select the best likelihood according to its Bayesian information criterion (BIC):

$$BIC = -2\log(\widehat{L}) + k\log(n_{\mathcal{D}}) \qquad (5.7)$$

where $\widehat{L}$ is the maximized value of the likelihood function, $n_{\mathcal{D}}$ the number of individuals in the sample $\mathcal{D}$, and k the number of parameters in the fitted marginal distribution.

5.1.3 *Inferring the copula parameters*

Each cell Σ_{ij} of the Σ covariance matrix of a multivariate copula distribution is the correlation parameter of a pairwise copula distribution. Hence, instead of inferring Σ from the set of all covariance matrices, we separately infer every cell $\Sigma_{ij} \in [0,1]$ from the joint sample of $\mathcal{D}_i$ and $\mathcal{D}_j$. We first measure the mutual information $I(\mathcal{D}_i;\mathcal{D}_j)$ between the two attributes and select $\sigma = \widehat{\Sigma_{ij}}$ minimizing the Euclidean distance between the empirical mutual information and the mutual information of the inferred joint distribution.

In practice, since the cdf. of a Gaussian copula is not tractable, we use a bounded Nelder-Mead minimization algorithm. For a given $(\sigma, (\Psi_i, \Psi_j))$, we sample from the distribution $q(\cdot|\sigma, (\Psi_i, \Psi_j))$ and generate a discrete bivariate sample Y from which we measure the objective:

$$f(\sigma) = \begin{cases} \left\| I(\mathcal{D}_i;\mathcal{D}_j) - I(Y_1;Y_2) \right\|_2 & \text{for } \sigma \in [0,1] \\ +\infty & \text{otherwise} \end{cases} \tag{5.8}$$

We then project the obtained $\widehat{\Sigma}$ matrix on the set of SDP matrices by solving the following optimization problem:

$$\begin{aligned} \min_{A} \quad & \left\| A - \widehat{\Sigma} \right\|_2 \\ \text{s.t.} \quad & A \succeq 0 \end{aligned} \tag{5.9}$$

5.1.4 *Modeling the association structure*

We use the pairwise mutual information to measure the strength of association between attributes. For a dataset $\mathcal{D}$, we denote by $I_{\mathcal{D}}$ the mutual information matrix where each cell $I(\mathcal{D}_i;\mathcal{D}_j)$ is the mutual information between attributes $\mathcal{D}_i$ and $\mathcal{D}_j$. When evaluating mutual information from small samples, obtained scores are often

overestimating the strength of association. We apply a correction for randomness using a permutation model [190]:

$$AI(\mathcal{D}_i;\mathcal{D}_j) = \frac{I(\mathcal{D}_i;\mathcal{D}_j) - \mathbb{E}\left[I(\mathcal{D}_i;\mathcal{D}_j)\right]}{\max\left\{\mathbb{H}(\mathcal{D}_i),\mathbb{H}(\mathcal{D}_j)\right\} - \mathbb{E}\left[I(\mathcal{D}_i;\mathcal{D}_j)\right]} \tag{5.10}$$

In practice, we estimate the expected mutual information between $\mathcal{D}_i$ and $\mathcal{D}_j$ with successive permutations of $\mathcal{D}_j$. We found that the adjusted mutual information provides significant improvement for small samples and large support size $|\mathcal{X}|$ compared to the naive estimator.

5.1.5 *Theoretical and empirical uniqueness*

For n individuals $x^{(1)}, x^{(2)}, \ldots, x^{(n)}$ drawn from X, the uniqueness Ξ_X is the expected percentage of unique individuals. It can be estimated either (*i*) by computing the mean of individual uniqueness or (*ii*) by sampling a synthetic population of n individuals from the copula distribution. In the former case, we have

$$\Xi_X \equiv \frac{1}{n}\,\mathbb{E}\Big[\sum_{i=1}^{n}\left[x^{(i)} \text{ unique in } (x^{(1)},\ldots,x^{(n)})\right]\Big] \tag{5.11}$$

$$= \frac{1}{n}\,\mathbb{E}\Big[\sum_{x\in\mathcal{X}} T_x\Big] \tag{5.12}$$

$$= \frac{1}{n}\sum_{x\in\mathcal{X}} \mathbb{E}[T_x] \tag{5.13}$$

where $T_x = [\exists! i, x^{(i)} = x]$ equals one if there exists a single individual i such as $x^{(i)} = x$ and zero otherwise. T_x follows a binomial distribution $B(p(x), n)$. Therefore

$$\mathbb{E}[T_x] = n\,p(x)\,(1-p(x))^{n-1} \tag{5.14}$$

and

$$\Xi_X = \sum_{x\in\mathcal{X}} p(x)\big(1-p(x)\big)^{n-1} \tag{5.15}$$

This requires to iterate over all combinations of attributes, whose number grows exponentially as the number of attributes increases, and quickly becomes computationally intractable. The

second method is therefore often more tractable and we use it to estimate population uniqueness.

For cumulative marginal distributions $F_1, F_2, \ldots, F_d$ and copula correlation matrix Σ, the procedure ValidationMAE samples n individuals from $q(\cdot|\Sigma, \Psi)$ using the latent copula distribution. From the n generated records $(y^{(1)}, y^{(2)}, \ldots, y^{(n)})$, we compute the empirical uniqueness

$$\Xi_X = \frac{1}{n}\left|\{i \in [1, n] \,/\, \forall j \neq i, y^{(i)} \neq y^{(j)}\}\right| \tag{5.16}$$

Procedure ValidationMAE(Σ, Ψ, n)

$L \leftarrow \text{Cholesky}(\Sigma)$ ▷ `Lower matrix decomposition`
for $i \leftarrow 1$ **to** n **do**
  for $j \leftarrow 1$ **to** d **do**
    Draw $z_j \sim \mathcal{N}(0, 1)$
  end for
  $x \leftarrow Lz$
  $u \leftarrow (\varphi(x_1), \ldots, \varphi(x_M))$
  $y^{(i)} \leftarrow (F_1(u_1|\Psi), \ldots, F_M(u_M|\Psi))$
end for
return $(y^{(1)}, y^{(2)}, \ldots, y^{(n)})$

5.1.6 *Individual uniqueness and correctness*

The probability distribution $q(\cdot|\Sigma, \Psi)$ can be computed by integrating over the latent copula density. Note that the marginal distributions X_1 to X_d are discrete, causing the inverses F_1^{-1} to F_d^{-1} to have plateaus. When estimating $p(x)$, we integrate over the latent copula distribution inside the hypercube $[x_1 - 1, x_1] \times [x_2 - 1, x_2] \times \ldots \times [x_d - 1, x_d]$:

$$q(x|\Sigma, \Psi) = \mathbb{P}(x_1 - 1 < X_1 \leq x_1, \ldots, x_d - 1 < X_d \leq x_d|\Sigma, \Psi) \tag{5.17}$$

$$= \int_{F_1^{-1}(x_1-1|\Psi)}^{F_1^{-1}(x_1|\Psi)} \cdots \int_{F_d^{-1}(x_d-1|\Psi)}^{F_d^{-1}(x_d|\Psi)} c_\Sigma(u)\, \mathrm{d}u \tag{5.18}$$

$$= \int_{\varphi^{-1}(F_1^{-1}(x_1-1|\Psi))}^{\varphi^{-1}(F_1^{-1}(x_1|\Psi))} \cdots \int_{\varphi^{-1}(F_d^{-1}(x_d-1|\Psi))}^{\varphi^{-1}(F_d^{-1}(x_d|\Psi))} \varphi_\Sigma(z)\, \mathrm{d}z \tag{5.19}$$

with φ_Σ the density of a zero-mean multivariate normal (MVN) of correlation matrix Σ. Several methods have been proposed in the literature to estimate MVN rectangle probabilities. Genz and Bretz [191, 192] proposed a randomized quasi Monte Carlo method which we use to estimate the discrete copula density.

The likelihood ξ_x for an individual's record x to be unique in a population of n individuals can be derived from $p_X(X = x)$:

$$\begin{aligned}
\xi_x &\equiv p_X(x \text{ unique in } (x^{(1)}, \ldots, x^{(n)}) \mid \exists i, x^{(i)} = x) && (5.20)\\
&= p_X(x \text{ unique in } (x^{(1)}, \ldots, x^{(n)}) \mid x^{(1)} = x) && (5.21)\\
&= p_X(\forall i \in [2, n], x^{(i)} \neq x) && (5.22)\\
&= \big(1 - p(x)\big)^{n-1} && (5.23)
\end{aligned}$$

which the copula model then estimates as $\widehat{\xi_x} = \big(1 - q(x|\Sigma, \Psi)\big)^{n-1}$.

Similarly, the likelihood κ_x for an individual's record x to be correctly matched in a population of n individuals can be derived from $p_X(X = x)$. With $T \equiv \sum_{i=1}^{n} \left[x^{(i)} = x\right] - 1$, the number of potential false positives in the population, we have:

$$\kappa_x \equiv \mathbb{P}(x \text{ correctly matched in } (x^{(1)}, \ldots, x^{(n)}) \mid \exists i, x^{(i)} = x) \tag{5.24}$$

$$= \sum_{k=0}^{n-1} \frac{1}{k+1} \mathbb{P}(T = k) \tag{5.25}$$

$$= \sum_{k=0}^{n-1} \frac{1}{k+1} \binom{n-1}{k} p(x)^k (1 - p(x))^{(n-1-k)} \tag{5.26}$$

$$= \frac{1}{n\,p(x)} \left(1 - (1 - p(x))^n\right) \tag{5.27}$$

Note that, since records are independent, T follows a Binomial distribution $B(n-1, p(x))$.

We substitute the expression for ξ_x in the last formula and obtain:

$$\begin{aligned}
\kappa_x &= \frac{1}{n\,p(x)} \left(1 - (1 - p(x))^n\right) && (5.28)\\
&= \frac{1}{n} \frac{1 - \xi_x^{n/(n-1)}}{1 - \xi_x^{1/(n-1)}} && (5.29)
\end{aligned}$$

5.2 EMPIRICAL VALIDATION

We collect five corpora from publicly available sources: population census (USA and MERNIS) as well as surveys from the UCI Machine Learning repository (ADULT, MIDUS, HDV). From each corpus, we create populations by selecting subsets of attributes (columns) uniformly. The resulting 210 populations cover a large range of uniqueness values (0 to 0.96), numbers of attributes (2 to 47), and records (7108 to 9 million individuals). For readability purposes, we report in here the numerical results for all five corpora but will show figures only for USA. Figures for MERNIS, ADULT, MIDUS, HDV are similar and can be found in the Supplementary Information of our related article [50] at https://nature.com/articles/s41467-019-10933-3.

5.2.1 *Data availability*

The USA datasets are extracted from the 1-Percent Public Use Microdata Sample (PUMS) files, a collection of 3,061,692 individual records from the 2010 US Census of Population and Housing. The PUMS files are available online on the US Census Bureau website and contain 11 attributes we use: state FIP, county, PUMA, number of vehicles, sex, date of birth, marital status, race, educational attainment, employment status, and occupation.

The Adult Income dataset (ADULT) is a canonical Machine Learning dataset, composed of 32,561 individuals from the 1994 US Census database. The ADULT dataset is available in the UCI Machine Learning Repository and contain 10 nominal and ordinal attributes we use: age, workclass, education-num, marital-status, occupation, relationship, race, sex, hours-per-week, and native-country.

MERNIS is a complete population database of virtually all 48 million individuals born before early 1991 in Turkey, that was made available online in April 2016 after a data leak from Turkey's Central Civil Registration System. Our use of this data was approved by Imperial College as it provides a unique opportunity to perform uniqueness estimation on a complete census survey. Due to the sensitivity of the data, we have only analyzed, at Imperial College, a copy of the dataset where every distinct value was replaced by

a unique integer to obfuscate records, without loss of precision for uniqueness modeling. We have analyzed a sample of 8,820,050 individuals (district of Istanbul) using 8 attributes: year, month, and day of birth, home city, district, address, and birthplace city and district.

The Histoire de vie (HDV) dataset is composed of 13,500 individual responses to a 2003 survey from the French National Institute of Statistics and Economic Studies (INSEE), and available on INSEE's website. After pre-processing and removal of null responses, the dataset contains 632 attributes we use for 8403 individuals.

Midlife in the United States (MIDUS) is a longitudinal survey of 7,108 individuals, comprising physical health, psychological well-being, and social variables. The survey files are available on the Inter-university Consortium for Political and Social Research (ICPSR) website. After pre-processing and removal of null responses, the dataset contains 415 attributes.

We also use the 5% PUMS files from 1990 to estimate the correctness of Governor Weld's re-identification and provide population uniqueness estimates in Fig. 5.3, for which we used 15 attributes: ZIP code (inferred from the PUMA code), date of birth (inferred from age), marital status, citizenship status, class, occupation, mortgage, state of work, race, vehicle occupancy, time of departure for work, sex, school, number of vehicles, number of own natural born or adopted children.

5.2.2 *Model validation*

For each surveyed population $\mathcal{D}$, we first infer the marginals distributions and then the correlation matrix of the latent copula distribution. We measure the uniqueness values Ξ_X (ground truth) and $\widehat{\Xi_X}$. For a corpus $\mathcal{C}$ of C populations, we report the uniqueness MAE as described in procedure SampleCopula.

5.2.3 *Subsampling experiments*

For each surveyed population $\mathcal{D}$, for each sampling fraction p, we select $m = 100$ samples $\mathcal{S}_1, \ldots, \mathcal{S}_m$, containing $n_{\mathcal{S}} = n \times p$ individuals. We use an algorithm similar to

Procedure SampleCopula(X, m)

$L \leftarrow \text{Cholesky}(\Sigma) \triangleright$ `Lower matrix decomposition`
for $k \leftarrow 1$ **to** m **do**
  $\Psi \leftarrow$ marginal estimates from X
  $\Sigma \leftarrow$ estimated copula correlation matrix from X
  for $i \leftarrow 1$ **to** n **do**
    Draw $y^{(i)} \sim q(\cdot|\Sigma, \Psi)$
  end for
  $\widehat{\Xi_k} \leftarrow \text{Uniqueness}[y^{(1)}, y^{(2)}, \ldots, y^{(n)}]$
end for
return $\frac{1}{m} \sum_{k=1}^{m} \left|\Xi_X - \widehat{\Xi_k}\right|$

the procedure SampleCopula to estimate the MAE distribution for a corpus $\mathcal{C}$ (procedure SubsampleMAE).

Procedure SubsampleMAE($X, m, n_{\mathcal{S}}$)

$L \leftarrow \text{Cholesky}(\Sigma) \triangleright$ `Lower matrix decomposition`
for $k \leftarrow 1$ **to** m **do**
  $X_{\mathcal{S}} \leftarrow$ sample $n_{\mathcal{S}}$ records from X without replacement.
  $\Psi \leftarrow$ marginal estimates from $X_{\mathcal{S}}$
  $\Sigma \leftarrow$ estimated copula correlation matrix from $X_{\mathcal{S}}$
  for $i \leftarrow 1$ **to** n **do**
    Draw $y^{(i)} \sim q(\cdot|\Sigma, \Psi)$
  end for
  $\widehat{\Xi_k} \leftarrow \text{Uniqueness}[y^{(1)}, y^{(2)}, \ldots, y^{(n)}]$
end for
return $\frac{1}{m} \sum_{k=1}^{m} \left|\Xi_X - \widehat{\Xi_k}\right|$

5.2.4 *Measuring the model calibration*

We use the Brier score [193] to measure the calibration of probabilistic predictions:

$$BS = \frac{1}{n} \sum_{i=1}^{n} \left(\xi_{x^{(i)}} - \widehat{\xi_{x^{(i)}}}\right)^2$$

with, in our case, $\xi_{x^{(i)}}$ the actual uniqueness of the record $x^{(i)}$ (1 if $x^{(i)}$ is unique and 0 if not) and $\widehat{\xi_{x^{(i)}}}$ the estimated likelihood.

Lemma 1. *If the same likelihood $\widehat{\xi_x} = p$ is assigned to every individual, the Brier Score when predicting individual uniqueness is:*

$$BS = \Xi_X (1-p)^2 + (1-\Xi_X) p^2$$

Proof. If the same likelihood $\widehat{\xi_x} = p$ is assigned to every individual, then

$$\begin{aligned} BS &= \text{Mean}\left[\left(\widehat{\xi_x} - \xi_x\right)^2\right] && (5.30)\\ &= \Xi_X \,\text{Mean}\left[\left(1-\widehat{\xi_x}\right)^2\right] + (1-\Xi_X)\,\text{Mean}\left[\left(0-\widehat{\xi_x}\right)^2\right] && (5.31)\\ &= \Xi_X \,\text{Mean}\left[(1-p)^2\right] + (1-\Xi_X)\,\text{Mean}\left[(0-p)^2\right] && (5.32)\\ &= \Xi_X (1-p)^2 + (1-\Xi_X) p^2 && (5.33) \end{aligned}$$

□

The lowest Brier score for a uniform likelihood is obtained when $p = \Xi_X$. In that case, we have $BS = \Xi_X (1-\Xi_X)$.

Lemma 2. *If $\widehat{\xi_x} \sim U(0,1)$ is uniformly drawn at random, the Brier Score when predicting individual uniqueness is:*

$$BS = 1/3$$

Proof. We similarly compute BS by dividing the positive and negative cases:

$$\begin{aligned} BS &= \text{Mean}\left[\left(\widehat{\xi_x} - \xi_x\right)^2\right] && (5.34)\\ &= \Xi_X \,\text{Mean}\left[\left(1-\widehat{\xi_x}\right)^2\right] + (1-\Xi_X)\,\text{Mean}\left[\left(0-\widehat{\xi_x}\right)^2\right] && (5.35)\\ &= \Xi_X \,\text{Mean}\left[\widehat{\xi_x}^2\right] + (1-\Xi_X)\,\text{Mean}\left[\widehat{\xi_x}^2\right] && (5.36)\\ &= \text{Mean}\left[\widehat{\xi_x}^2\right] && (5.37)\\ &= \int_0^1 u^2 \,\mathrm{d}u = 1/3 && (5.38) \end{aligned}$$

□

5.3 RESULTS

Fig. 5.1A shows that, when trained on the entire population, our model correctly estimates population uniqueness

$$\Xi_X = \sum_{x \in \mathcal{X}} p(x)\big(1 - p(x)\big)^{n-1}$$

i.e. the expected percentage of unique individuals in $\big(x^{(1)}, x^{(2)}, \ldots, x^{(n)}\big)$. The mean absolute error (MAE) between the empirical uniqueness of our population Ξ_X and the estimated uniqueness $\widehat{\Xi_X}$ is 0.028 ± 0.026 [mean ± s.d.] for USA and 0.018 ± 0.019 on average across every corpus (see Table 5.1). Fig. 5.1A and Supplementary Figure 1[10] furthermore show that our model correctly estimates uniqueness across all values of uniqueness, with low within-population s.d. (Supplementary Table 3).

Fig. 5.1B shows that our model estimates population uniqueness very well even when the dataset is heavily sampled (see Supplementary Figure 2, for other populations). For instance, our model achieves an MAE of 0.029 ± 0.015 when the dataset only contains 1% of the USA population and an MAE of 0.041 ± 0.053 on average across every corpus. Table 5.1 shows that our model reaches a similarly low MAE, usually below 0.050, across corpora and sampling fractions.

[10] *These supplementary Figures and Tables, not included here, are available in the online Supplementary Information of our corresponding article at https://nature.com/articles/s41467-019-10933-3.*

5.3.1 *Likelihood of successful re-identification*

Once trained, we can use our model to estimate the likelihood of his employer having correctly re-identified John Doe, our 50 year-old male from Berkeley with breast cancer. More specifically, given an individual record x, we can use the trained model to compute the likelihood $\widehat{\xi_x} = \big(1 - q(x|\Sigma, \Psi)\big)^{n-1}$ for this record x to be unique in the population. Our model takes into account information on both marginal prevalence (*e.g.*, breast cancer prevalence) and global attribute association (*e.g.*, gender and breast cancer). Since the cdf. of a Gaussian copula distribution has no close-form expression, we evaluate $q(x|\Sigma, \Psi)$ with a numerical integration of the latent continuous joint density inside the hyperrectangle defined by the d components $(x_1, x_2, \ldots, x_d)$ [191, 192]. We assume no prior

		MERNIS	USA	ADULT	HDV	MIDUS
Corpus	n	8,820,049	3,061,692	32,561	8,403	7,108
	c	10	40	50	50	60
	[min Ξ, max Ξ]	[0.087, 0.844]	[0.000, 0.961]	[0.000, 0.794]	[0.002, 0.941]	[0.052, 0.944]
Sampling fraction	100%	0.029 ± 0.019	0.028 ± 0.026	0.018 ± 0.016	0.006 ± 0.009	0.018 ± 0.014
	10%	0.030 ± 0.019	0.028 ± 0.016	0.022 ± 0.020	0.011 ± 0.009	0.035 ± 0.044
	5%	0.029 ± 0.019	0.027 ± 0.016	0.027 ± 0.023	0.015 ± 0.012	0.037 ± 0.055
	1%	0.029 ± 0.019	0.029 ± 0.015	0.027 ± 0.014	0.045 ± 0.050	0.055 ± 0.079
	0.5%	0.028 ± 0.019	0.029 ± 0.015	0.048 ± 0.039		
	0.1%	0.026 ± 0.017	0.058 ± 0.037			

TABLE 5.1: Mean Absolute Error (mean ± s.d.) when estimating population uniqueness (100 trials per population). Our model correctly estimates population uniqueness even when only a small to very small fraction of the population is available. n denotes the population size and c the corpus size (the total number of populations considered per corpus). We do not estimate population uniqueness when the sampled dataset contains less than 50 records.

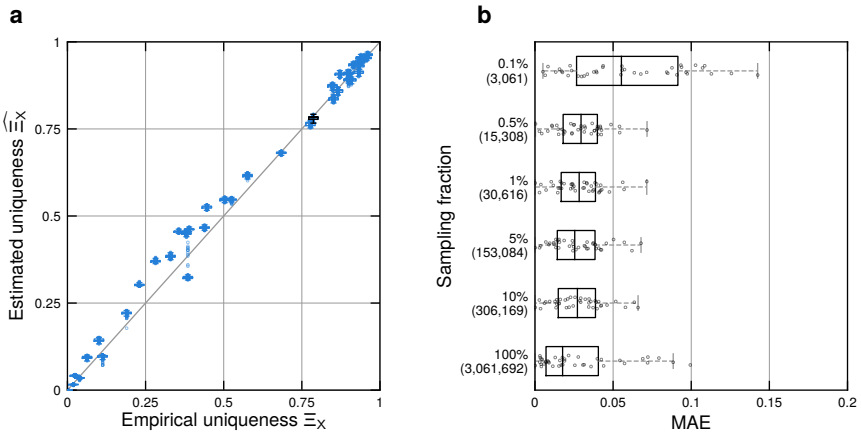


FIG. 5.1: Estimating the population uniqueness of the USA corpus. (a) We compare, for each population, empirical and estimated population uniqueness (boxplot with median, 25th and 75th percentiles, maximum 1.5 interquartile range (IQR) for each population, with 100 independant trials per population). For example, date of birth, location (PUMA code), marital status, and gender uniquely identify 78.7% of the 3 million people in this population (empirical uniqueness) which our model estimates to be 78.2 ± 0.5% (boxplot in black). (b) Absolute error when estimating USA's population uniqueness when the disclosed dataset is randomly sampled from 10% to 0.1%. The boxplots (25, 50, and 75th percentiles, 1.5 IQR) show the distribution of MAE for population uniqueness, at one subsampling fraction across all USA populations (100 trials per population and sampling fraction). The y-axis shows both p, the sampling fraction, and $n_S = p \times n$, the sample size. Our model estimates population uniqueness very well for all sampling fractions with the MAE slightly increasing when only a very small number of records are available ($p = 0.1\%$ or 3,061 records).

knowledge on the order of outcomes inside marginals for nominal attributes and randomize their order.

Fig. 5.2A shows that, when trained on 1% of the USA populations, our model predicts very well individual uniqueness, achieving a mean AUC (area under the receiver-operator characteristic curve) of 0.89. For each population, to avoid overfitting, we train the model on a single 1% sample then select 1,000 records, independent from the training sample, to test the model. For re-identifications that the model predicts to be always correct ($\widehat{\xi_x} > 0.95$, estimated individual uniqueness above 95%), the likelihood of them to be incorrect (false-discovery rate) is 5.26% (see bottom-right inset in Fig. 5.2A). ROC curves for the other populations are available in Supplementary Figure 3 and have overall a mean AUC of 0.93 and mean false-discovery rate of 6.67% for $\widehat{\xi_x} > 0.95$ (see Supplementary Table 5.2).

TABLE 5.2: AUC and FDR for the classification of individual uniqueness. For each population, for 1000 individuals sampled at random in the whole population, we estimate their individual uniqueness (method trained on a 1% sample) and compare the predicted likelihood to the true value. We report the AUC (mean ± s.d.) and the FDR per corpus and overall. We also report the F-scores in Table 5.3.

			False discovery rate (%)		
Corpus	c	AUC	$\xi = 0.90$	$\xi = 0.95$	$\xi = 0.99$
MERNIS	10	0.84 ± 0.05	11.40	7.88	3.80
USA	40	0.89 ± 0.06	7.43	5.26	1.99
ADULT	50	0.91 ± 0.05	15.62	12.37	7.62
HDV	50	0.97 ± 0.03	6.36	3.95	1.20
MIDUS	60	0.96 ± 0.02	5.91	3.91	1.55
Overall	210	0.93 ± 0.06	9.34 ± 4.12	6.67 ± 3.57	3.23 ± 2.65

Finally, Fig. 5.2B shows that our model outperforms even the best theoretically achievable prediction using only population uniqueness, *i.e.* assigning the score $\xi_x^{(\text{pop})} = \Xi_X$ to every individual (ground truth population uniqueness, see Section 5.1.5). We use the Brier score [193] to measure the calibration of probabilistic predictions (see Section 5.2.4). Our model obtains scores on average 39% lower than the best theoretically achievable prediction

TABLE 5.3: F-score for the classification of individual uniqueness. For each population, for 1000 individuals sampled at random in the whole population, we estimate their individual uniqueness (method trained on a 1% sample) and compare the predicted likelihood $\widehat{\xi_x}$ to the true value ξ_x. We report the F-score per corpus and overall. Underlined scores indicate the highest score per corpus. If the model must obtain a good balance between precision and recall, the optimal threshold lies near $\xi = 0.50$. Yet, if the model must obtain a low proportion of false positives, achieved by a small false-discovery rate, Table 5.2 shows that this requires a higher cutoff, above $\xi = 0.90$.

ξ cutoff	MERNIS	USA	ADULT	HDV	MIDUS	Overall
$\xi = 0.10$	0.72	0.87	0.76	0.75	0.81	0.78 ± 0.06
$\xi = 0.20$	0.75	0.88	0.78	0.78	0.83	0.80 ± 0.05
$\xi = 0.30$	0.76	0.89	0.79	0.79	0.84	0.81 ± 0.05
$\xi = 0.40$	0.77	0.89	0.80	<u>0.80</u>	0.85	0.82 ± 0.05
$\xi = 0.50$	<u>0.78</u>	0.90	0.81	0.79	<u>0.86</u>	<u>0.83 ± 0.05</u>
$\xi = 0.60$	0.78	0.90	0.82	0.78	0.85	0.83 ± 0.05
$\xi = 0.70$	0.76	<u>0.91</u>	<u>0.82</u>	0.76	0.84	0.82 ± 0.06
$\xi = 0.80$	0.73	0.90	0.82	0.72	0.82	0.80 ± 0.08
$\xi = 0.90$	0.64	0.87	0.80	0.63	0.77	0.74 ± 0.11
$\xi = 0.95$	0.51	0.83	0.76	0.52	0.71	0.67 ± 0.14
$\xi = 0.99$	0.21	0.69	0.64	0.33	0.55	0.48 ± 0.20

using only population uniqueness, emphasizing the importance of modeling individuals' characteristics.

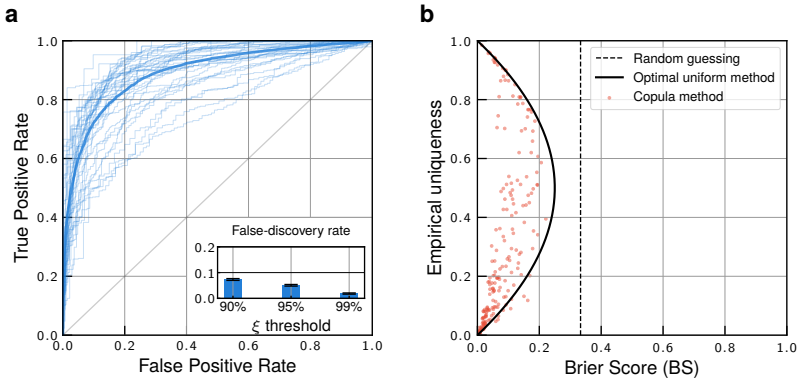


FIG. 5.2: The model predicts correct re-identifications with high confidence. (a) Receiver operating characteristic (ROC) curves for USA populations (light ROC curve for each population and a solid line for the average ROC curve) Our method accurately predicts the (binary) individual uniqueness. (*Inset*) False-discovery rate (FDR) for individual records classified with $\xi > 0.9$, $\xi > 0.95$, and $\xi > 0.99$. For re-identifications that the model predicts are likely to be correct ($\widehat{\xi_x} > 0.95$), only 5.26% of them are incorrect (FDR). (b) Our model outperforms by 39% the best theoretically achievable prediction using population uniqueness across every corpus. A red point shows the Brier Score obtained by our model, when trained on a 1% sample. The solid line represents the lowest Brier Score achievable when using the exact population uniqueness while the dashed line represents the Brier Score of a random guess prediction ($BS = 1/3$).

5.3.2 *Appropriateness of the de-identification model*

Using our model, we revisit the (successful) re-identification of Gov. Weld [174]. We train our model on the 5% Public Use Microdata Sample (PUMS) files using ZIP code, date of birth, and gender, and validate it using the last national estimate [32]. We show that, as a male born on July 31, 1945 and living in Cambridge (02138), the information used by Latanya Sweeney at the time, William Weld was unique with a 58% likelihood ($\xi_x = 0.58$ and $\kappa_x = 0.77$), meaning that Latanya Sweeney's re-identification had 77% chances of being correct. We show that, if his medical records had included number of children—5 for William Weld—, her re-identification would have had 99.8% chances of being correct! Fig. 5.3A shows that the same combinations of attributes (ZIP code, date of birth, gender, and

number of children) would also identify 79.4% of the population in Massachusetts with high confidence ($\widehat{\xi_x} > 0.80$). We finally evaluate the impact of specific attributes on William Weld's uniqueness. We either change the value of one of his baseline attributes (ZIP code, date of birth, or gender) or add one extra attribute, in both cases picking the attribute at random from its distribution. Fig. 5.3C shows for instance that individuals with 3 cars or no car are harder to re-identify than those with 2 cars. Similarly, it shows that it would not take much to re-identify people living in Harwich Port, MA, a city of less than 2,000 inhabitants.

To evaluate the impact of specific attributes on William Weld's uniqueness, we train our copula model on the PUMS corpus for the Massachusetts population. For each baseline attribute, (ZIP code, date of birth, or gender), we then perform 1000 trials, randomly replacing the value of this attribute by a new value sampled at random from its marginal distribution. For each trial, we estimate uniqueness.

For each single additional attribute, we sample from the marginal distribution of this additional attribute, add the new value to the base characteristics (58 year old male from Cambridge, MA), and estimate the uniqueness ξ_x using the 3+1 attributes. We perform 1000 trials for each of the eleven additional attributes.

Fig. 5.3C reports these scores of uniqueness, grouped by attribute: each boxplot shows the distribution of uniqueness obtained after replacing (resp. adding) a baseline (resp. additional) attribute. In order of appearance, the attributes used from the PUMS corpus are ZIP code (ZCTAs extrapolated from PUMA codes), gender (sex variable in the PUMS corpus), date of birth (extrapolated from age with random month and day), Race, citizenship (citizen), school enrollment (school), vehicle occupancy (riders), place of work–state (powstate), mortgage (mortgage), marital status (marital), class of worker (class), number of vehicles (vehicles), and occupation (occup).

Modern datasets contain a large number of points per individuals. For instance, the data broker Experian sold Alteryx access to a de-identified dataset containing 248 attributes per household for 120M Americans [30]; Cambridge university researchers shared anonymous Facebook data for 3M users collected through the myPersonality app and containing, amongst other attributes, users'

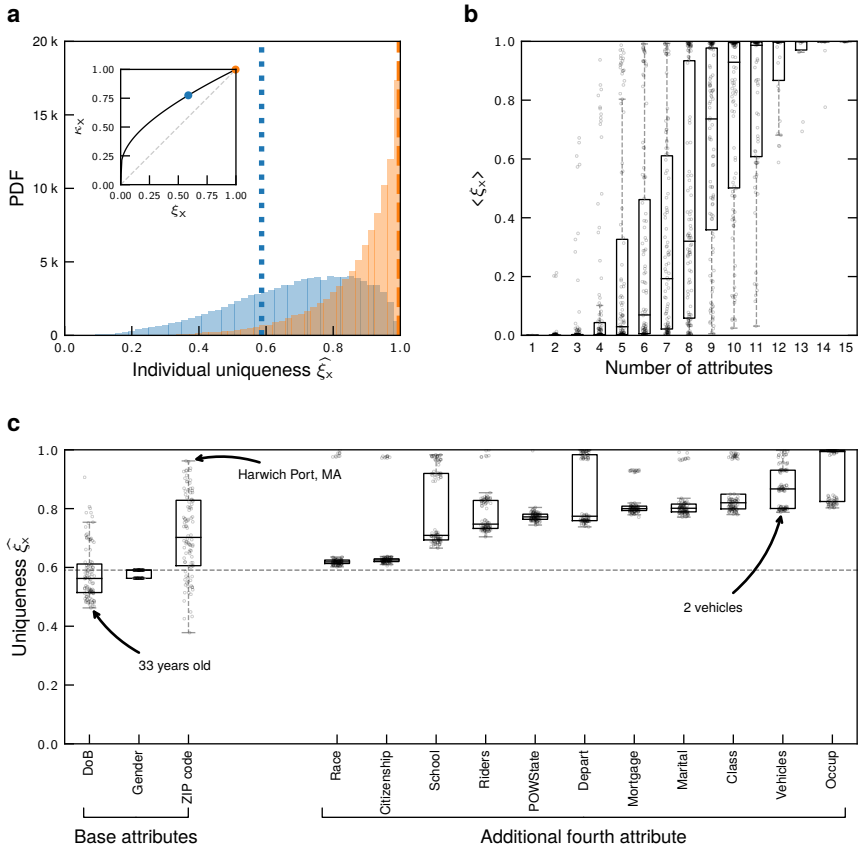


FIG. 5.3: Average individual uniqueness increases fast with the number of collected demographic attributes. (a) Distribution of predicted individual uniqueness knowing ZIP code, date of birth, and gender (resp. ZIP code, date of birth, gender, and number of children) in blue (resp. orange). The dotted blue line at $\widehat{\xi_x} = 0.580$ (resp. dashed orange line at $\widehat{\xi_x} = 0.997$) illustrates the predicted individual uniqueness of Gov. Weld knowing the same combination of attributes. (*Inset*) The correctness κ_x is solely determined by uniqueness ξ_x and population size n (here for Massachusetts). We show individual uniqueness and correctness for William Weld with 3 (in blue) and 4 (in orange) attributes. (b) The boxplots (25, 50, and 75th percentiles, 1.5 IQR) show the average uniqueness $\langle \xi_x \rangle$ knowing k demographic attributes, grouped by number of attributes. The individual uniqueness scores ξ_x are estimated on the complete population in Massachusetts, based on the 5% PUMS files. (c) Uniqueness varies with the specific value of attributes. For instance a 33 year old is less unique than a 58 year old person. We here either (*i*) randomly replace the value of one baseline attribute (ZIP code, date of birth, or gender) or (*ii*) add one extra attribute, both by sampling from its marginal distribution, to the uniqueness of a 58 years old male from Cambridge, MA. The dashed baseline shows his original uniqueness $\widehat{\xi_x} = 0.580$ and the boxplots the distribution of individual uniqueness obtained after randomly replacing or adding one attribute.

age, gender, location, status updates, and results on a personality quiz [194]. These datasets do not necessarily share all the characteristics of the one studied here. Yet, our analysis of the re-identification of Gov. Weld by Latanya Sweeney shows that few combined attributes are often enough to render the likelihood of correct re-identification very high. For instance, Fig. 5.3B shows that the average individual uniqueness increases fast with the number of collected demographic attributes and that 15 demographic attributes would render 99.98% of people in Massachusetts unique. While few attributes might not be sufficient for a re-identification to be correct, collecting a few more attributes will quickly render the re-identification very likely to be successful.

5.3.3 *Correcting individual scores*

In our experiments, we noticed a small numerical discrepancy between the mean of the estimated likelihoods $\widehat{\xi_x}$ and the estimated population uniqueness. Specifically, $\mathbb{E}[\widehat{\xi_x}] > \widehat{\Xi_{\mathcal{X}}}$ in 66% of the tested populations (see Fig. 5.4). This is likely due to numerical errors between (i) sampling a population of n individuals to compute population uniqueness and (ii) sampling 1,000 individuals from the original dataset and computing their estimated likelihood $\widehat{\xi_x}$ from $q(x|\Sigma, \Psi)$ (*e.g.*, for 9M individuals, a uniqueness of $\widehat{\xi_x} = 0.90$ corresponds to a small probability $q(x|\Sigma, \Psi) = 1.1710^{-8}$).

We test here whether the population-level method (ii) can be used to correct (normalize) the individual likelihoods from method (i). Recall that, from a sample $\mathcal{S}$, we estimate both the average uniqueness $\widehat{\Xi_{\mathcal{X}}}$ and the likelihood $\widehat{\xi_x}$ for each individual x. We apply a correction factor α:

$$\mathbb{E}[{\xi_x}^{\alpha}] = \widehat{\Xi_{\mathcal{X}}} \tag{5.39}$$

and let $\xi_x^* = {\xi_x}^{\alpha}$ be the corrected likelihood of uniqueness for the record x.

This correction reduces, for certain corpora (MERNIS, ADULT, HDV), the false-discovery rate (Fig. 5.5) and provides an increased calibration (Fig. 5.6). However, we did not find enough evidence that the calibration helps for the data surveyed here, and do not use here; this correction could be employed for a specific use case where a significant discrepancy is measured between population and invidual uniqueness.

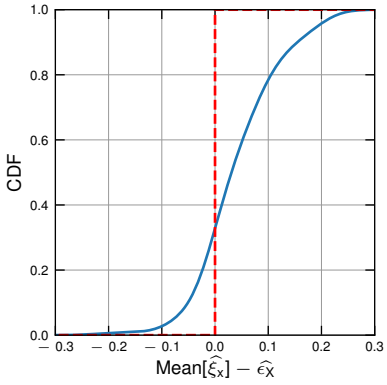


FIG. 5.4: Cumulative distribution of the deviation between average individual likelihood and predicted population uniqueness. For each population (all corpora combined), we select 1000 individuals from the original population, and compare their average likelihood with the estimated population uniqueness (in blue). The dashed red line represents the baseline null deviation after correction. The copula model is inferred on a 1% sample from the original population.

5.3.4 *Using exact marginals improves predictions*

As mentioned in the Discussion section of the main text, marginal distributions can be inadequately modeled when the sample size is small. Our method can incorporate exogenous information such as better estimates for marginals. Table 5.4 compares the performance of our model when using approximate (inferred from the given sample) and exact marginals (prior knowledge of the marginal distributions from the complete population). Using the exact marginals decrease MAE when the sample is small.

5.4 RELATED WORK

While developed to estimate the likelihood of a specific re-identification to be successful, our model can also be used to estimate population uniqueness. We show below that, while not its primary goal, our model performs consistently better than existing methods to estimate population uniqueness on all five corpora (Supplementary Figure 4, $P < 0.05$ in 78 cases out of 80) [42–44, 195–197], and consistently better than previous attempts to

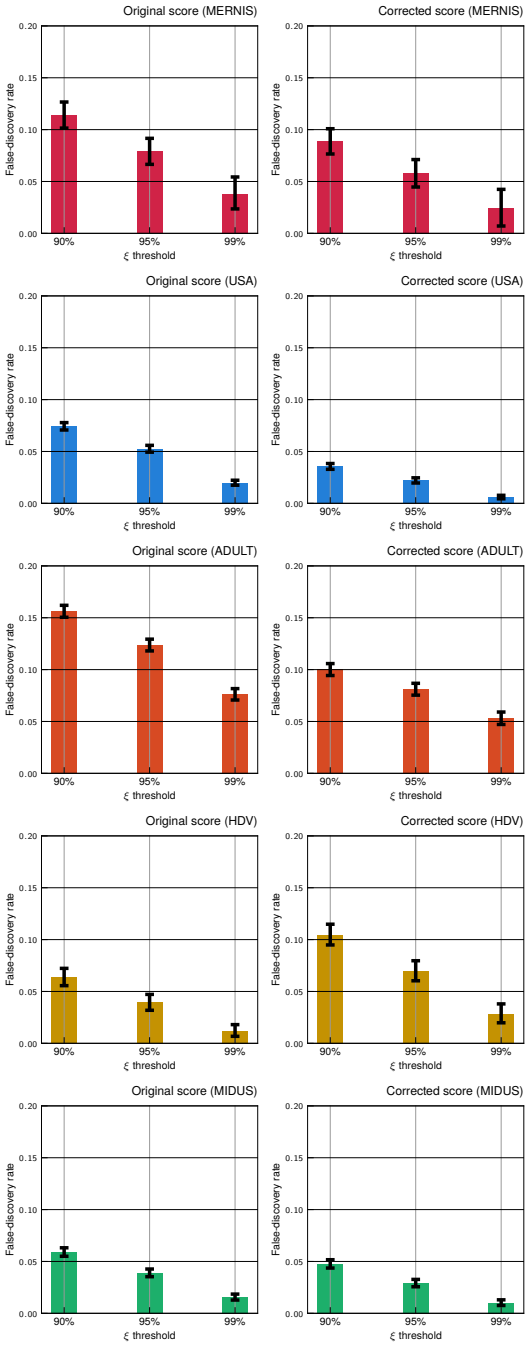


FIG. 5.5: False-discovery rate with original (left) and corrected (right) scores for all five corpora. The copula model is trained on a 1% sample, and the FDR scores are computed on 1000 records from the original population.

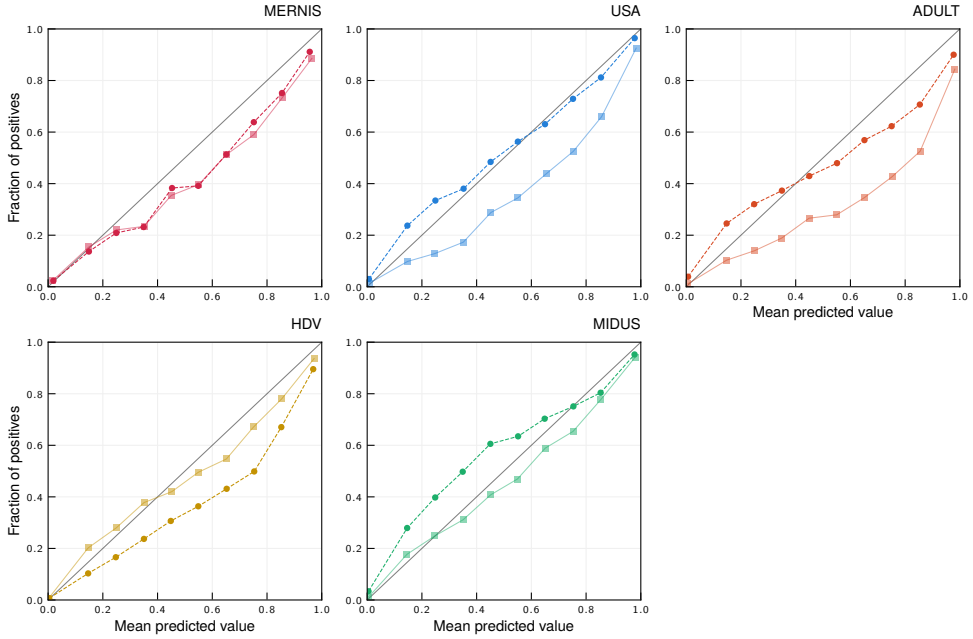


FIG. 5.6: Reliability diagrams for copula methods trained on 1% samples. Calibration plots of mean predicted value vs. fraction of positive outcomes (unique individuals). The solid lines are the copula estimators $\widehat{\xi}_x$, the dashed ones the corrected copula estimators $\widehat{\xi}_x^*$ and the diagonal lines represents an ideal predictor. We train copula methods on 1% samples for each population, and report these measures for 1000 records per population.

TABLE 5.4: Error rates for predicting population uniqueness (exact marginals). Mean absolute error (MAE) [mean ± s.d., in percent] on estimated population uniqueness grouped by corpus. n denotes the population size and c the corpus size (the total number of populations considered per corpus). We do not evaluate the model when samples contain less than 50 records. Compared to Table 5.1, exact marginals provide lower error rates for small sampling fractions.

		MERNIS	USA	ADULT	HDV	MIDUS
Corpus	n	8,820,049	3,061,692	32,561	8,403	7,108
	c	10	40	50	50	60
Sampling fraction	100%	0.029 ± 0.019	0.028 ± 0.026	0.018 ± 0.016	0.006 ± 0.009	0.018 ± 0.014
	10%	0.029 ± 0.018	0.028 ± 0.016	0.020 ± 0.018	0.007 ± 0.008	0.029 ± 0.029
	5%	0.029 ± 0.019	0.027 ± 0.016	0.020 ± 0.018	0.008 ± 0.009	0.030 ± 0.030
	1%	0.029 ± 0.019	0.029 ± 0.016	0.021 ± 0.017	0.016 ± 0.015	0.032 ± 0.030
	0.5%	0.029 ± 0.019	0.029 ± 0.016	0.023 ± 0.016		
	0.1%	0.029 ± 0.018	0.030 ± 0.017			

estimate individual uniqueness [198, 199]. Existing approaches, indeed, exhibit unpredictably large over- and under-estimation errors. Finally, a recent work quantifies the correctness of individual re-identification in incomplete (10%) hospital data using complete population frequencies [168]. Compared to this work, our approach does not require external data nor to assume this external data to be complete.

5.4.1 *Population uniqueness methods*

Previous approaches, based on extrapolations of the contingency table of a disclosed random sample of the dataset, have been proposed to model population uniqueness [42–44, 195–198, 200]. A contingency table is here a d-dimensional table where each cell (or class) counts the number of individuals with a specific combinations of the d attributes.

Using our previous notations, let $|\mathcal{X}|$ denote the number of potential combinations of attributes, *i.e.* the true number of cells in the complete contingency table. F_i (resp. f_i) denotes the size of the i$^{\text{th}}$ cell in the contingency table of X (resp. $X_{\mathcal{S}}$), with an arbitrary order on cells. We define $S_k = \sum_i 1_{F_i=k}$ (the number of cells of size k in X) and $s_k = \sum_i 1_{f_i=k}$ (the number of cells of size k in $X_{\mathcal{S}}$) while the empirical uniqueness of the population is $\Xi_X = S_1/n$. All of those methods then use parametric models, fitting a specific distribution on the unordered frequency distribution of the contingency table, and estimate the population uniqueness from the unordered estimated distribution of cell frequencies.

Following studies comparing these estimators [42–44], we selected the most promising methods to estimate uniqueness: the Ewens model [195], the Slide Negative Binomial (SNB) model [42], the Pitman mosdel [197], and the Zayatz model [200].

The Ewens model, based on the multivariate Ewens distribution, estimates uniqueness as:

$$\widehat{\Xi_X} = \frac{s_1 \left(n_{\mathcal{S}} - 1\right)}{n_{\mathcal{S}} \left(n - 1\right) - s_1 \left(n - n_{\mathcal{S}}\right)} \tag{5.40}$$

The Zayatz model estimates uniqueness as:

$$\widehat{\Xi_X} = \frac{s_1}{n_{\mathcal{S}}} \, \mathbb{P}\left(F = 1 \,|\, f = 1\right) \tag{5.41}$$

where $\mathbb{P}(F = i | f = 1)$ follows an hypergeometric distribution.
The Pitman model assumes:

$$\widehat{\Xi_X} = \frac{\Gamma(\theta + 1)}{\Gamma(\theta + \alpha)} n^{\alpha - 1} \tag{5.42}$$

The α and θ parameters are estimated from the log-likelihood of the sampling distribution, using Newton-Raphton.
The Slide Negative Binomial model assumes a translated negative binomial distribution on the F_j frequencies (without zero count), such as:

$$\widehat{\Xi_X} = \frac{|\mathcal{X}|}{n} \beta^{\alpha} \tag{5.43}$$

The parameters α, β are estimated from the sample by solving a system of two non-linear equations, and $|\mathcal{X}|$ separately estimated from the sample.

We have implemented the original methods and verified the numerical results using the open source ARX toolbox [133]. Sometimes, they do not converge or estimate uniqueness above 1. We therefore limit scores to [0, 1] and, if a method does not converge on a specific sample (such as occasionally with the SNB inference [44]), we do not make a decision and discard the sample. Taking a conservative approach and assuming, *e.g.*, that every individual is unique when the method do not converge gives the same results below ($P < 0.05$ in 78 cases out of 80 on Fig. 5.7). Specifically, the Pitman method estimates uniqueness to be above 1.00 for 12.4% of all trials while the SNB (resp. Ewens) method do not converge in 4.2% (resp. 1.1%) of all trials.

According to the literature, both the Pitman and SNB methods provide accurate estimators of population uniqueness while the Pitman method provides the best estimator for small sampling fractions [44]. We found that, while not its primary goal, our method performs significantly better than all other approaches ($P < 0.05$ in 78 cases out of 80). Fig. 5.7 compares the mean absolute error (MAE) between empirical and estimated uniqueness for the four proposed methods and ours. Several of the methods proposed in the literature severely under- or over-estimate the risk of re-identification, especially when their parameters are fitted on a small population sample. This is likely due to the fact that (*i*) fitting specific distribution to frequency counts can inherently provide

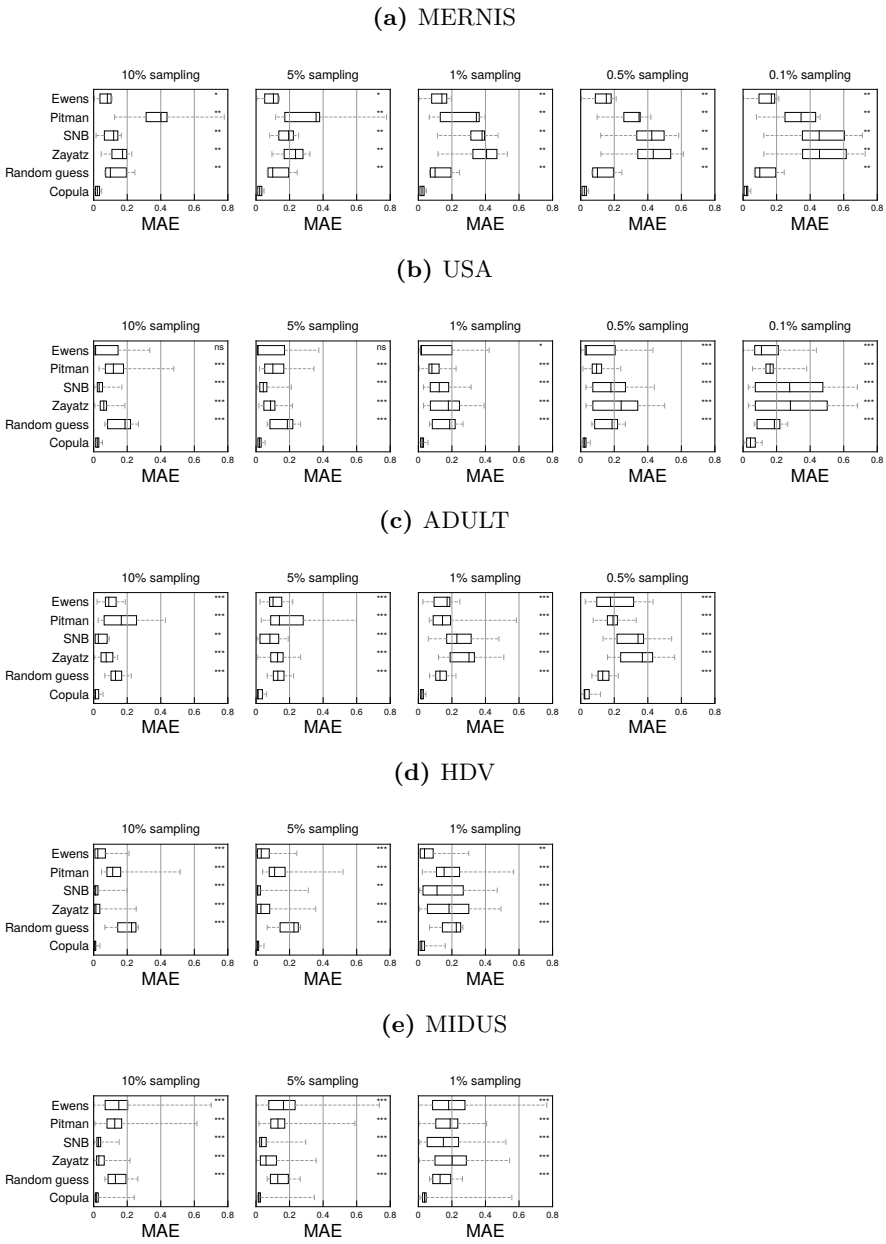


FIG. 5.7: Comparing the performance of our model against frequency-based re-identification risk models from the literature. Our model obtains a significantly lower MAE on every corpus and sampling fraction but two (USA with 10% and 5% sampling compared to the Ewens model). We report the MAE for average uniqueness precision (using 100 trials per population) and the p-value of the Wilcoxon signed-rank test, comparing the performance of the copula model with other approaches (*** for $P < 0.01$, ** for $P < 0.05$, * for $P < 0.1$, and otherwise ns. when non significant). The panels (a) to (e) correspond to the corpora MERNIS, USA, ADULT, HDV, MIDUS.

biased results if inappropriate distributions are selected, (*ii*) fitting the frequency counts of a population requires a lot more samples than for a multivariate model where each marginal distribution is estimated separately.

5.4.2 *Log-linear models*

Finally, Skinner and Holmes [198] and Skinner and Shlomo [199] have proposed to use log-linear models to estimate the likelihood for sample unique records to be population unique. Using the sample contingency table, these models can smooth an estimator of population uniqueness, by taking into account the main effects from the sample marginals.

A log-linear model is a generalized linear model (GLM) used to model count data and contingency tables. It assumes that the response variable, the population count F_x of an equivalence class $x \in \mathcal{X}$ (the number of individuals with the record x in the population), follows a a specific count distribution (*e.g.*, Poisson) with mean λ_x, that is: $F_x \sim Po(\lambda_x)$. For a sampling fraction p, the sample count f_x also follows a Poisson distribution: $f_x \sim Po(p\,\lambda_x)$.

In order to "borrow strength" [198] between equivalence classes, a GLM model is fitted to the sample using the canonical logarithmic link:

$$\log\lambda_x = z_k\beta \tag{5.44}$$

where β is a $1 \times q$ parameter vector, and z_k a $q \times 1$ vector of the main effects of each marginal attribute, and potentially low order interactions between attributes. It yields an estimated individual likelihood of uniqueness:

$$\widehat{\xi_x} = e^{-(1-p)\lambda_x} \tag{5.45}$$

We have implemented log-linear models on all studied corpora. The log-linear models did not converge for 27 of the 210 studied populations (when all sample records are sample unique, or share the same frequency, a GLM cannot be fitted, as there is only one possible outcome). When they converge, they obtain poor calibration with Brier scores, on average, 418% higher (lower is better) than our copula-based method (and strictly worse for 210 out of 210 tested population). Fig. 5.8 shows the calibration of Poisson

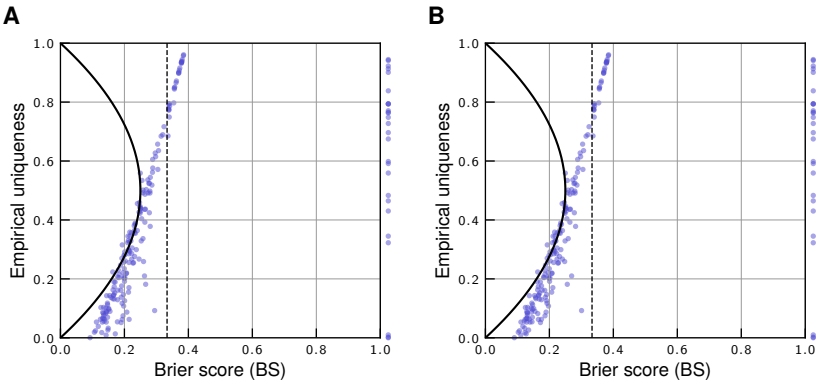


FIG. 5.8: Poor calibration achieved by log-linear models. A blue point shows the Brier Score obtained by a log-linear model, when trained on a 1% sample, and evaluated on sample unique records. The panel (a) uses a Poisson log-linear model and the panel (b) a Negative Binomial model. The solid line represents the lowest Brier Score achievable when using the exact population uniqueness while the dashed line represents the Brier Score of a random guess prediction ($BS = 1/3$). For 11% of all studied populations, the model did not converge: we report these results with dots outside the right margin.

and Negative Binomial log-linear models on a 1% sample. This is to be compared with Fig. 5.2B, showing the calibration of our copula-based approach.

Despite being an individual-level measure of uniqueness, log-linear models do not perform better at predicting individual uniqueness than simply using population uniqueness (assigning every individual x the overall population uniqueness $\widehat{\xi_x} = \widehat{\Xi_X}$). For instance, a Poisson (resp. Negative Binomial) log-linear model obtains Brier scores on average 56.4% (resp. 68.1%) higher than the Zayatz method (see above), and 233% (resp. 291%) higher than the best theoretically achievable prediction using only population uniqueness (for each individual x, assigning $\widehat{\xi_x} = \Xi_X$).

5.5 DISCUSSION

Beyond the claim that the incompleteness of the dataset provides plausible deniability, our method also challenges claims that a low population uniqueness is sufficient to protect people's privacy [141, 170]. Indeed, an attacker can, using our model, correctly re-identify an individual with high likelihood even if the population uniqueness

is low (Fig. 5.3A). While more advanced guarantees such as k-anonymity [201] would give every individual in the dataset further protection, they have been shown to be NP-Hard [202], hard to achieve in modern high-dimensional datasets [134], and not always sufficient [131].

To study the stability and robustness of our estimations, we perform further experiments (see sections infra).

First, we analyze the impact of marginal and association parameters on the model error, and show how to use exogenous information to lower it. In Section 5.5.4, we show that, at very small sampling fraction (below 0.1%), where the error is the largest, the error is mostly determined by the marginals, and converges after few hundred records when the exact marginals are known. The copula covariance parameters exhibit no significant bias and decrease fast when the sample size increases (Section 5.5.5).

As our method separates marginals and association structure inference, exogenous information from larger data sources could also be used to estimate marginals with higher accuracy. For instance, count distributions for attributes such as date of birth or ZIP code could be directly estimated from national surveys. We replicate our analysis on the USA corpus using a subsampled dataset to infer the association structure along with the exact counts for marginal distributions. Incorporating exogenous information reduces, *e.g.*, the mean MAE of uniqueness across all corpora by 48.6% ($P < 0.01$, Mann–Whitney) for a 0.1% sample. Exogenous information become less useful as the sampling fraction increases (Supplementary Table 2).

Second, our model assumes that $\mathcal{D}$ is either uniformly sampled from the population of interest X or, as several census bureaus are doing, released with post-stratification weights to match the overall population. We believe this to be a reasonable assumption as biases in the data would greatly affect its usefulness and affect any application of the data, including our model. To overcome an existing sampling bias, the model can be (*i*) further trained on a random sample from the population $\mathcal{D}$ (*e.g.*, microdata census or survey data) and then applied to a non-uniform released sample (*e.g.*, hospital data, not uniformly sampled from the population), or (*ii*) trained using better, potentially unbiased, estimates for marginals or association structure coming from other sources (see above).

Third, since $\mathcal{D}$ is a sample from the population X, only the records that are unique in the sample can be unique in the population. Hence, we further evaluate the performance on our model only on records that are sample unique and show that it only marginally decrease the AUC (see Section 5.5.2). We therefore prefer to not restrict our predictions to sample unique records as (a) our models need to perform well on non sample unique records for us to be able to estimate correctness and (b) to keep the method robust if oversampling or sampling with replacement were to have been used.

5.5.1 *Testing for bias and heteroskedasticity*

Our estimate of uniqueness might be more accurate at lower or at higher uniqueness. We therefore test for potential biases in our estimates for population uniqueness ($\widehat{\Xi_X}$) and for homoskedasticity of errors. Intuitively, the prediction the model outputs should have the same accuracy (no bias) and the same variance of error (homoskedasticity) across uniqueness values.

We use general linear mixed-effects models with restricted maximum likelihood (REML) on the estimates given by our copula method at 100% sampling size. We control for corpus effect by grouping observations by corpus (5 groups).

TABLE 5.5: Results of mixed-effects REML regressions for $\widehat{\Xi_X}$ and for the RMSE between Ξ_X and $\widehat{\Xi_X}$.

Dependent Variable	Regressors	Coef.	Std. Err.	CI 95%
$\widehat{\Xi_X}$	Intercept	0.003	0.005	(-0.007, 0.012)
	Ξ_X	1.024	0.000	(1.023, 1.025)
	Group effect	0.000	0.004	
RMSE	Intercept	0.015	0.004	(0.008, 0.022)
	Ξ_X	0.014	0.005	(0.004, 0.023)
	Group effect	0.000	0.002	

Overall, we find a statistically significant bias and heteroskedasticity, albeit both with negligible effects on uniqueness. Table 5.5 shows the results of modeling $\widehat{\Xi_X}$ (32000 observations across corpora and repeated trials) and the RMSE between Ξ_X and

$\widehat{\Xi_X}$ (210 observations). Notably, there is not significant difference between corpora.

We further test for potential bias and heteroskedasticity in the predicted *individual* likelihoods of uniqueness. Table 5.6 shows the results of modeling $\widehat{\xi_x}$ (210,000 observations), grouped by corpus and population (210 groups), as well as modeling the RMSE between ξ_x and $\widehat{\xi_x}$, grouped by corpus (5 groups). The results validate the homoscedasticity of errors and exhibit no significant bias.

TABLE 5.6: Results of mixed-effects REML regressions for $\widehat{\xi_x}$ (grouped by corpus and population) and for the RMSE between ξ_x and $\widehat{\xi_x}$ (grouped by corpus).

Dependent Variable	Regressors	Coef.	Std. Err.	CI 95%
$\widehat{\xi_x}$	Intercept	0.040	0.037	(-0.033, 0.112)
	Ξ_X	0.934	0.106	(0.726, 1.141)
	Group effect	0.055	0.018	
RMSE	Intercept	0.159	0.067	(0.028, 0.291)
	Ξ_X	0.328	0.178	(-0.021, 0.677)
	Group effect	0.004	0.035	

5.5.2 *Sample unique records and uniqueness*

As described in the text, we do not take into account whether a record x is unique in $\mathcal{D}$. Our modeled uniqueness

$$\xi_x = (1 - p(x))^{n-1}$$

depends only on the probability to draw x in the population and on n, the number of individuals in the population.

As $\mathcal{D}$ is sampled at random from the population, every record x that is not unique in the sample $\mathcal{D}$ cannot be unique in the population ($\xi_x = 0$). We therefore further evaluate the performances of our model only on records that are sample unique. Table 5.7 shows the AUC and FDR scores when the model is evaluated only on sample unique records, and Supplementary Figure 8 (see online Supplementary Information) the ROC curves for the same experiment.

TABLE 5.7: AUC and FDR for the classification of individual uniqueness (only sample unique records, to be compared with Table 5.2). For each population, for all individuals unique in the 1% training sample, we estimate their individual uniqueness (method trained on a 1% sample) and compare the predicted likelihood to the true value. We report the AUC (mean ± s.d.) and the FDR per corpus and overall.

			False discovery rate (%)		
Corpus	c	AUC	$\xi = 0.90$	$\xi = 0.95$	$\xi = 0.99$
MERNIS	10	0.82 ± 0.05	11.53	8.38	4.83
USA	40	0.88 ± 0.05	7.19	5.01	1.85
ADULT	50	0.89 ± 0.04	15.90	12.80	7.22
HDV	50	0.95 ± 0.03	5.98	3.65	1.09
MIDUS	60	0.95 ± 0.02	6.17	4.22	1.68
Overall	210	0.92 ± 0.05	9.36 ± 4.29	6.81 ± 3.82	3.34 ± 2.61

We therefore prefer to not restrict our predictions to sample unique records. First, this keeps the method more robust *e.g.*, if oversampling or sampling with replacement were to have been used. In that case, we can never rule out that a sample non-unique record is not unique in the population. Second, in order to accurately measure the correctness κ_x of a matched record x, we need to ensure that the model performs well for any record, sample unique or not. Indeed, if 9 other records match x in the population X, and 2 other records in the sample $\mathcal{D}$ (x sample non-unique), x still has a 1 chance out of 10 to be correctly re-identified.

However, potential adversaries running a re-identification attack on a released sample will use this sample to train the model then estimate the uniqueness of the sample records. Therefore, an adversary's success at predicting uniqueness can also be measured on both sample unique and non-unique records. For this reason, we also provide the ROC curves for the five corpora, when the model is trained and tested on the same 1% sample (Supplementary Figure 13,

see online Supplementary Information). We update the definition of ξ_x, the likelihood for an individual x to be unique in the population:

$$\xi_x^{(\text{sample})} = \begin{cases} (1 - p(x))^{n-1} & \text{if } x \text{ is unique in } \mathcal{D} \\ 0 & \text{otherwise} \end{cases} \tag{5.46}$$

This yields a higher accuracy, as sample non-unique records are never population unique (perfect prediction).

5.5.3 *Multivariate mutual information*

Our copula method takes into account the marginals and pairwise association structure. While our model, when trained on the complete population, already performs very well with an average MAE across corpora of 0.018 ± 0.019, more complex models, capturing better the complete association structure between attributes, might perform even better.

To investigate this, we compute the distribution of triplewise information $I(X; Y; Z)$. Positive interaction information values denote redundant combinations (X and Y share the same information about Z) and negative values synergistic combinations (X and Y provide more information about Z together than they do individually).

We perform 100 trials per corpus. For MERNIS, ADULT, and HDV, the interaction factors are synergistic but very small, with an average mean ± s.d. of −0.001 ± 0.035 (MERNIS), −0.017 ± 0.036 (ADULT), and −0.002 ± 0.009 (HDV). For USA and MIDUS, the interactions factors are redundant, with an average mean ± s.d. of 0.309 ± 1.050 (USA) and 0.009 ± 0.152 (MIDUS).

Across all 500 trials, the triplewise interaction factors obtained are often null or redundant, with $I(X; Y; Z) > -0.1$nat in 94% of all trials. This suggest that, at least for the five corpora we study, covering a broad range of uniqueness values and association patterns, pairwise associations capture most of the information.

5.5.4 *Estimating population uniqueness at extremely small sampling fractions for large datasets*

The MERNIS and USA populations contain respectively 8,820,049 and 3,061,692 individuals. Even a 0.1% sample still contains the

records of thousands of individuals. We here study the performance of our method at extremely small sample size for those two datasets.

Fig. 5.9A and C show that our model performs well until a sample size of approximately 6400 records. Fig. 5.9B and D then show that, when using the exact marginals, our models performs well even with as little as 200 records. Indeed, at small sampling fractions, marginals with high entropy and therefore many unique records, such as the "postal address" attribute in MERNIS, are difficult to estimate without a few thousand records. Incorrectly inferring the distribution of "postal address" with only 50 or 100 records can lead to a large variance in population uniqueness estimates.

5.5.5 *Evaluation of the model error*

Our method to estimate individual uniqueness relies on a generative model for $p(x)$, the probability to draw a record x from the joint distribution X which we call $q(x)$. Figure 5.10 shows that the distribution of the Kullback–Leibler divergence, a measure of distance between the true empirical distribution $p(x)$ and the estimated distribution $q(x)$, is very small (1.59nats on average).

Fig. 5.11 furthermore shows that the estimated covariance matrix of the latent copula distribution is not significantly biased, even at small sampling fractions (pairwise correlations higher or lower, on average, than their value when trained on 100% of the population). Similarly, the variance of the covariance error decreases fast, showing signs of convergence after few hundred records.

5.6 SUMMARY

Our results, first, show that few attributes are often sufficient to re-identify with high confidence individuals in heavily incomplete datasets and, second, reject the claim that sampling or releasing partial datasets, *e.g.*, from one hospital network or a single online service, provide plausible deniability. Finally, they show that, third, even if population uniqueness is low—an argument often used to justify that data is sufficiently de-identified to be considered anonymous [141]—, many individuals are still at risk of being successfully re-identified by an attacker using our model.

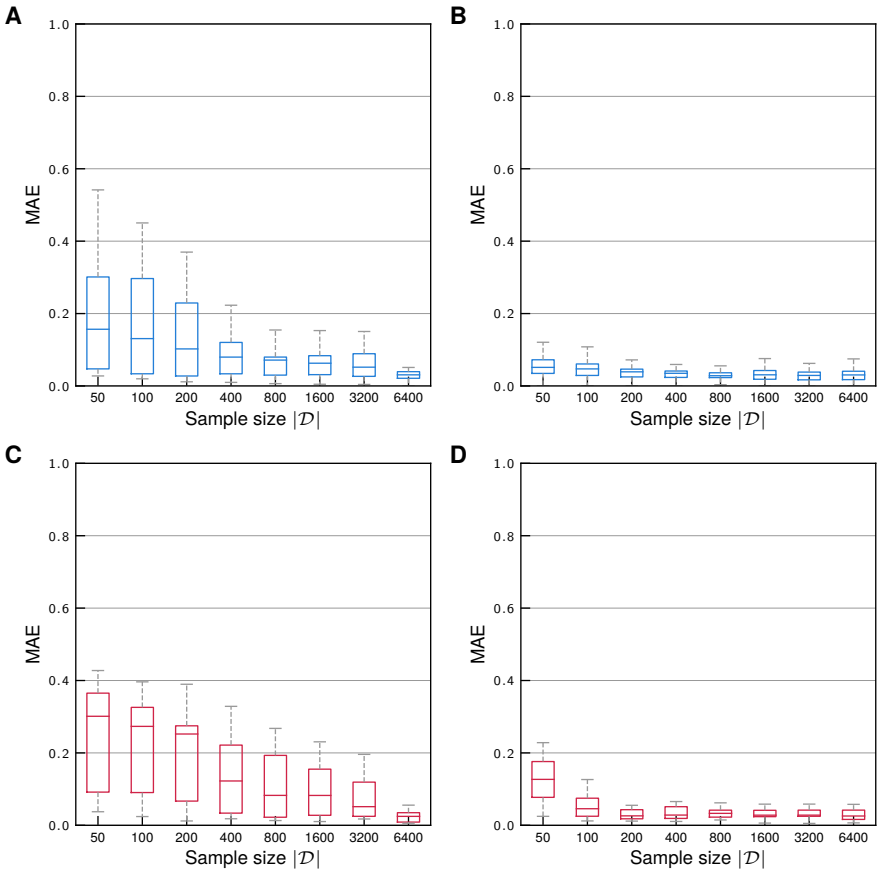


FIG. 5.9: At very small sampling fractions, the error is mostly determined by the marginals. The boxplots show the mean absolute error (MAE) for population uniqueness estimates, grouped by sample size (from 50 to 6400 records). Panels A and C represent the absolute error for the USA and MERNIS populations when the marginals are approximated from the sample, while panels B and D are the absolute error using the exact marginals (for USA and MERNIS respectively).

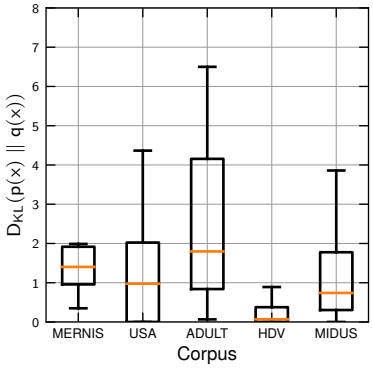


FIG. 5.10: Kullback–Leibler divergence (in nats) between the empirical $p(x)$ and estimated $q(x)$ distributions for each corpus. The copula method is trained on a 1% sample. Overall, the copula method achieves small prediction errors, with a K-L divergence of 1.59 ± 1.78 nats overall.

As standards for anonymization are being redefined, including by national and regional data protection authorities in the EU, it is essential for them to be robust and account for new threats like the one we present in this chapter. They need to take into account the individual risk of re-identification and the lack of plausible deniability—even if the dataset is incomplete—, as well as legally recognize the broad range of provable privacy-enhancing systems and security measures that would allow data to be used while effectively preserving people's privacy [203, 204].

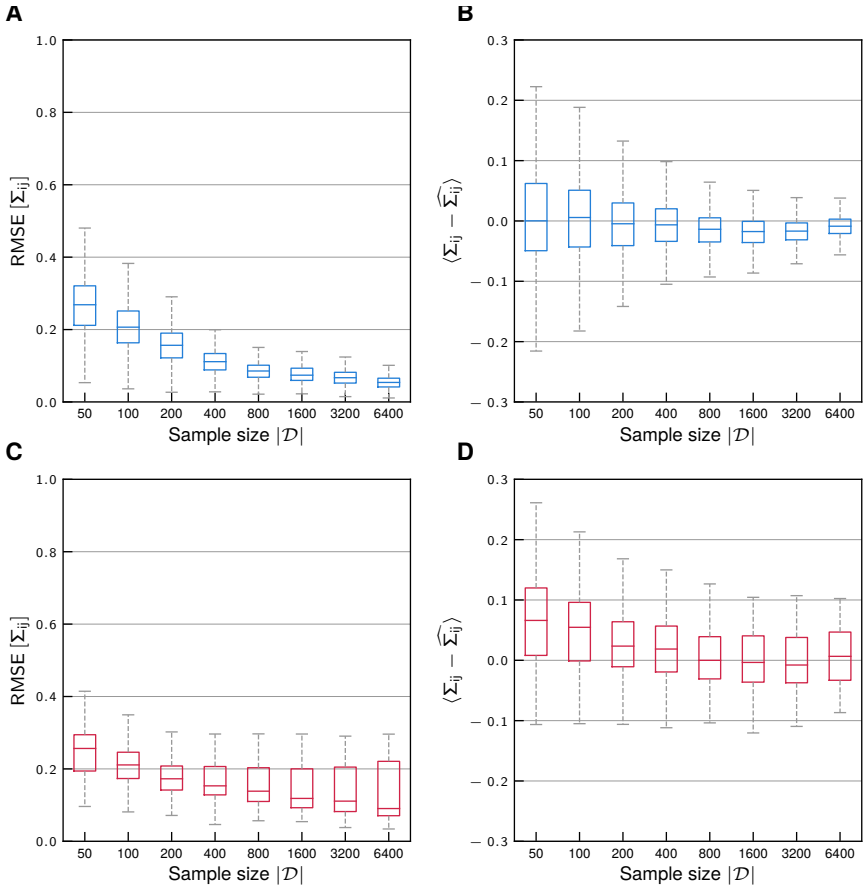


FIG. 5.11: The copula parameters converge quickly. Left and right panels compare the covariance parameter $\widehat{\Sigma}$ estimated on 50 to 6400 records, with Σ estimated on 100% of the population. The boxplots in panel A (resp. C) show the RMSE between $\widehat{\Sigma}$ and Σ for the USA (resp. MERNIS) corpus. The boxplots in panel B (resp. D) show the bias between $\widehat{\Sigma}$ and Σ for the USA (resp. MERNIS) corpus. Both metrics are computed, for each population, on every pair (i, j) of attributes.

6

A THEORETICAL MODEL FOR ANONYMITY

6.1 INTRODUCTION

We showed in Chapter 5 how to estimate the uniqueness of individuals and the correctness of re-identifications from heavily sampled datasets. However, despite the empirical data-driven studies showing the risks associated with anonymized data collections, concerned individuals and data protection authorities can still hardly evaluate claims that deemed-anonymous data will never be re-identified. Without legal or practical access to the data, auditing the anonymity of present or future data collections is a challening task.

Auditing anonymity

To audit anonymity without data, many scholars have suggested to estimate the Shannon entropy, an information theory metric measuring the average *surprise* or uncertainty over all possible information known by an adversary, hence capturing whether there is enough information to re-identify individuals. Entropy is a compelling tool to audit anonymity, with two fundamental principles. The coined "feasibility limit" principle (FLP) [205, 206] links re-identification risk and entropy; the "Aggregation of Total Knowledge" (ATK) [207] or cumulative entropy [124] exploits the additivity of entropy to bound the total risk. The FLP posits that 28 bits of entropy (and 33 bits on Earth) are required to identify one out of 300M Americans [98, 206, 208, 209]. Using the ATK, inferring the gender reduces the entropy to 27 bits, the state of residency and age to 16 bits, the surname to 3 bits [206], resulting in theory with 97% of correct re-identifications. These rough guestimates are widely used to predict whether data collections are anonymous or not.

Despite the plausibility of the FLP and ATK for predicting anonymity, our results below demonstrate how they actually fail capture the structural conditions that govern the re-identification

of individuals. Our results in this paper suggest that their failure is not experimental, but fundamental: they rely upon two implicit assumptions—uniform distributions and independance between collected attributes—, both almost never valid in practice and yielding highly inaccurate over- and under-estimations.

Here, we demonstrate how to correctly estimate which data collections will be anonymous or not from few key statistics but no data. First, we propose a universal model, the PY model, using a Bayesian nonparametrics approach to reason about anonymity. We show that traditional entropy-based approaches do not capture the crucial variations exhibited in our empirical analyses. Second, we derive a general intuitive framework to audit from few statistics the risk of re-identification in any data collections, from tabular data (census and surveys) to high-dimensional datasets (mobile app usage, mobile phone and shopping patterns). Applied to web, mobile apps, and behavioral fingerprinting data collections, our framework shows that current anonymity claims require far more guarantees than what practitioners provide.

6.2 MODELING UNIQUENESS AND CORRECTNESS

We consider the general scenario of a data collection D over a potential population of up to n records. An adversary, armed with auxiliary information about one individual, Alice, aims to either trace back whether Alice participated in the data collection (membership or tracing attack) or identify correctly her record (record linkage attack). An anonymous data collection must protect against both attacks. For example, for an anonymized longitudinal study of 100 HIV-positive UK, no auxiliary information should allow an adversary to test whether a UK resident ($n = 66.1$M) participated in the study or, worse, identify their exact medical information in the study.

6.2.1 *A Pitman-Yor model*

This is traditionally captured by two measures, defined in the literature to quantify the likelihood to re-identify individual records [50]. The *uniqueness* ε measures the overall risk of

membership attacks. It characterizes the probability that, for one individual in the population, the $n-1$ others will not share the same auxiliary information in D. The *correctness* κ measures the likelihood of record linkage attacks. Intuitively, if exactly three individuals share the same auxiliary information in D, they all have one chance out of three to be correctly re-identified if matched in any dataset.

For example, websites may collect visitors' device fingerprint, a persistent tracking identifier which can be derived for all $n = 6\text{B}$ existing smartphones. Alice, browsing an online website, unwillingly discloses her device fingerprint. The website, armed with Alice's fingerprint (auxiliary information), may now trace back whether Alice visited the website in the past and identify her in commercial databases by using her fingerprint.

We denote by x_A the auxiliary information available about an individual record $x \sim X$, with x drawn i.i.d. according to a distribution X, and x_A according to the marginal X_A. Focusing here on discrete data collections, we consider X to be a multivariate discrete distribution and x_A a subset of the vector x ($x_A \in \mathbb{N}^d$ for d auxiliary attributes).

Both the expected uniqueness ε and the correctness κ are controlled by the frequency distribution X_A (see Methods) with

$$\varepsilon \equiv \sum_{x \sim X} \mathbb{P}(X = x) \left[\left(1 - \mathbb{P}(X_A = x_A)\right)^{n-1} \right]$$

and

$$\kappa \equiv \sum_{x \sim X} \mathbb{P}(X = x) \left[\frac{1 - \left(1 - \mathbb{P}(X_A = x_A)\right)^n}{n\, \mathbb{P}\left(X_A = x_A\right)} \right]$$

Fundamentally, what makes a re-identification possible? First, the size of the population: one individual is always unique but, as n increases, both ε and κ decrease, until everyone is hidden in the crowd. Second, the information available to an adversary: the more information, the easier it becomes to successfully identify a target; this is captured by the distribution X_A.

To propose a formal model of ε and κ, we part from the traditional frequentist methods, all requiring access to data samples [42–44, 196]. We instead take a nonparametrics Bayesian approach, building upon the work of Nemenman and Archer [210, 211]. We assume no specific frequency distribution for X_A, but place a prior over all countably-infinite discrete distributions. Notably, this requires no

prior knowledge about the support size of X_A yet is consistent with finite distributions [211].

To model a large range of discrete distributions, we use a Pitman-Yor process (PYP), a two-parameter prior distribution which generalizes the Dirichlet process. These processes offer a flexible nonparametric prior for discrete distributions. They have been used to study how online social networks grow [212], count object frequencies for image segmentation [213], and build species sampling models in ecology [214]. We denote by $PY(d, \alpha)$ the PYP with a discount parameter $d \in [0, 1]$ and a concentration parameter $\alpha \in [-d, +\infty]$. Distributions $\pi \sim PY(d, \alpha)$ can be drawn using a generative "stick-breacking" process, by sampling independent Beta random variables:

$$\beta_i \sim \text{Beta}(1 - d, \alpha + i\,d), \quad \pi_i = \prod_{j=1}^{i-1}(1 - \beta_j)\beta_i$$

The discount parameter d controls the shape of the tail, large values of d (resp. $d = 0$) are associated with heavy tail (resp. exponential tail) distributions. The concentration parameter α controls the probability mass for the first outcomes, with large values of α yielding spread-out distributions.

Setting a PYP prior for the distribution of auxiliary information $X_A \sim PY(d, \alpha)$, we derive two analytical expressions for the expected uniqueness and correctness (see section 6.4.1):

$$\mathbb{E}\left[\varepsilon \,|\, d, \alpha\right] = \frac{\Gamma(1+\alpha)}{\Gamma(d+\alpha)} \frac{\Gamma(n+d+\alpha-1)}{\Gamma(n+\alpha)} \tag{6.1}$$

$$\mathbb{E}\left[\kappa \,|\, d, \alpha\right] = \frac{1}{n\,d}\left(\frac{\Gamma(1+\alpha)}{\Gamma(d+\alpha)} \frac{\Gamma(n+d+\alpha)}{\Gamma(n+\alpha)} - \alpha\right) \tag{6.2}$$

with Γ the standard gamma (factorial) function.

We empirically demonstrate the utility of this model, which we call PY model, on a wide range of auxiliary information from demographics census, surveys, fingerprints, and synthetic data. For each dataset, we select different subsets A of attributes at random and consider them as auxiliary information. Overall, we obtain 483 sets of auxiliary information. For each set A of auxiliary information, from the empirical frequency distribution of auxiliary information, we estimate maximum a-posteriori parameters α^* and d^* (see Section 6.4.3). We sample distributions $\widehat{X_A} \sim PY(d^*, \alpha^*)$,

using stick-breaking representations, to obtain 95% confidence intervals on the inferred probability mass functions, and compute the expected uniqueness and correctness using the analytical expressions 6.1 and 6.2.

Fig. 6.1 compares the true and inferred posterior distributions of auxiliary information, and the resulting uniqueness for a population size from 10^0 to 10^9 individuals. To assess how accurate the PYP models the distributions we consider, we calculate the Kullback–Leibler divergence (a non-parametric discrepancy measure between probability distributions) between the empirical distributions and the inferred ones. The typical divergences, of the order of 0.23 bits, are small compared to the total entropy of the distributions, 12.5 ± 7.73, suggesting absence of overfitting and good model specification, despite using only two parameters. In particular, the PYP correctly models both geometric (Fig. 6.1d; $KL = 0.08 \pm 0.06$ for an entropy of 4.23) and power-law distributions (Fig. 6.1e; $KL = 0.13 \pm 0.17$ for an entropy of 4.57, Table 6.1).

Furthermore, the PY model accurately captures the empirical uniqueness and correctness scores (Fig. 6.1g-h). Over all studied distributions, we obtain low bias and high precision, with a RMSE of 0.028 for uniqueness (Table 6.1), compared to 0.489 for a random estimator. The FLP, assuming a uniform frequency distribution, obtains an RMSE of 0.252, only slightly better than random.

6.2.2 *Combining information theory and tail distributions*

Our results above suggest that the two-parameters PY model correctly captures the uniqueness and correctness of auxiliary information, but also explains why the FLP fails to accurately model the wide range of behaviors of uniqueness and correctness we studied (Fig. 6.2).

The PY model parameters d and α are closely related to the entropy of auxiliary information. Below, we reparametrize the Pitman-Yor process by using the expected entropy h and a

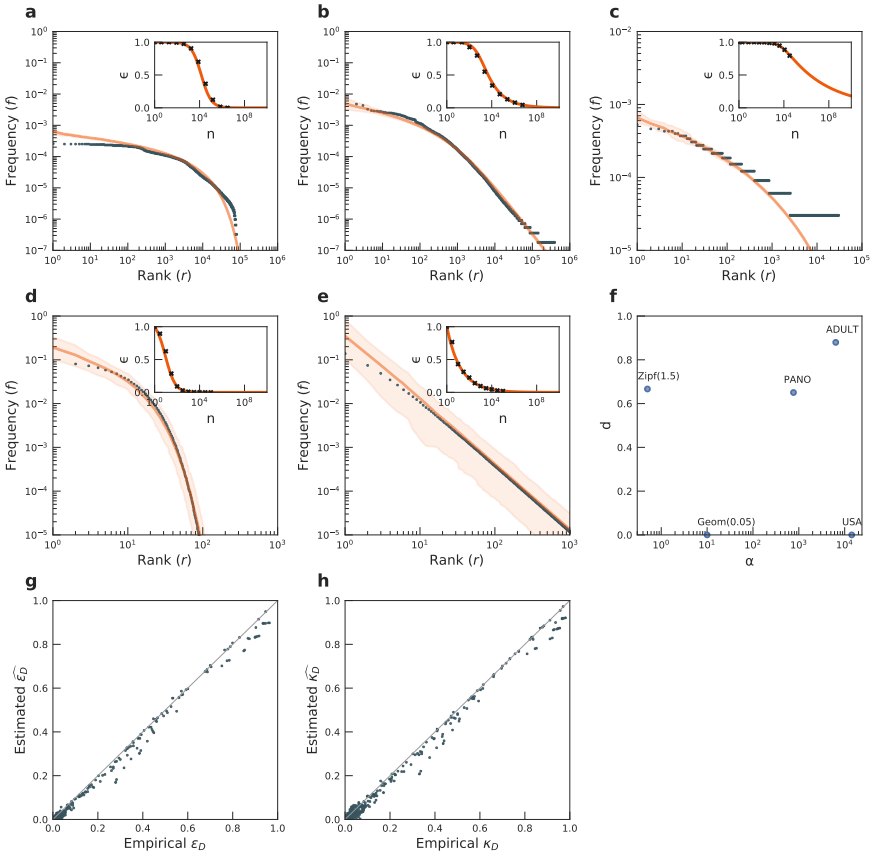


FIG. 6.1: Pitman-Yor processes fit a wide range of discrete count distributions, and correctly model their uniqueness. (a-e) Frequency plot of empirical samples (black) and 95% CI from the MAP estimate of the PYP. (Inset) Empirical and expected uniqueness for a population size ranging from 1 to 9B individuals. (a) Age, sex, location for the 1% ACS census published in 2010 (n=3,061,692). (b) User Agent strings from the Panopticlick survey (n=5,463,165). (c) Adult dataset extracted from the 1994 US census (n=32,561). (d) Synthetic sample from a Geometric distribution (p=0.05) (e) Synthetic sample from a Geometric distribution (p=0.1). (e) Synthetic sample from a Zipf distribution (s=1.5). (f) MAP estimated parameters α and d for each of the five surveyed dataset in panels (a) to (e). (g) Empirical vs. estimated uniqueness ε for all surveyed sets of auxilary information, across all datasets. (h) Same figure as (g) with the correctness κ.

		ε			κ		PY parameters	
	c	Range	RMSE (PY)	RMSE (FL)	Range	RMSE (PY)	α^*	d^*
ADULT	40	0.01–0.79	0.008	0.310	0.02–0.87	0.011	7.6e0–6.4e3	0.28–0.88
HDV	71	0.01–0.94	0.008	0.283	0.04–0.97	0.017	1.4e1–2.6e4	0.00–0.92
MIDUS	230	0.01–0.94	0.013	0.176	0.01–0.97	0.021	2.6e0–2.6e4	0.00–0.92
WEB	32	0.01–0.90	0.052	0.347	0.02–0.93	0.059	1.9e1–2.6e4	0.62–0.97
WEB (GSM)	28	0.02–0.52	0.056	0.204	0.04–0.57	0.067	5.0e0–4.1e2	0.58–0.92
USA	40	0.01–0.96	0.040	0.452	0.04–0.98	0.046	9.7e3–6.6e6	0.09–0.97
Geometric	17	0.01–0.63	0.035	0.170	0.01–0.075	0.04	1.3e4–9.8e5	0.00–0.00
Poisson	8	0.01–0.12	0.034	0.047	0.01–0.14	0.04	2.3e4–2.2e5	0.00–0.00
Zipf	17	0.01–0.32	0.058	0.157	0.01–0.34	0.07	6.2e0–3.3e1	0.69–0.92

TABLE 6.1: Empirical validation of the PY model. We report the results of the PY model, as well as the uniform FL model, for all selected datasets, grouped by corpus, for the uniqueness ε and for the correctness κ. We also report the number of datasets c selected from each corpus.

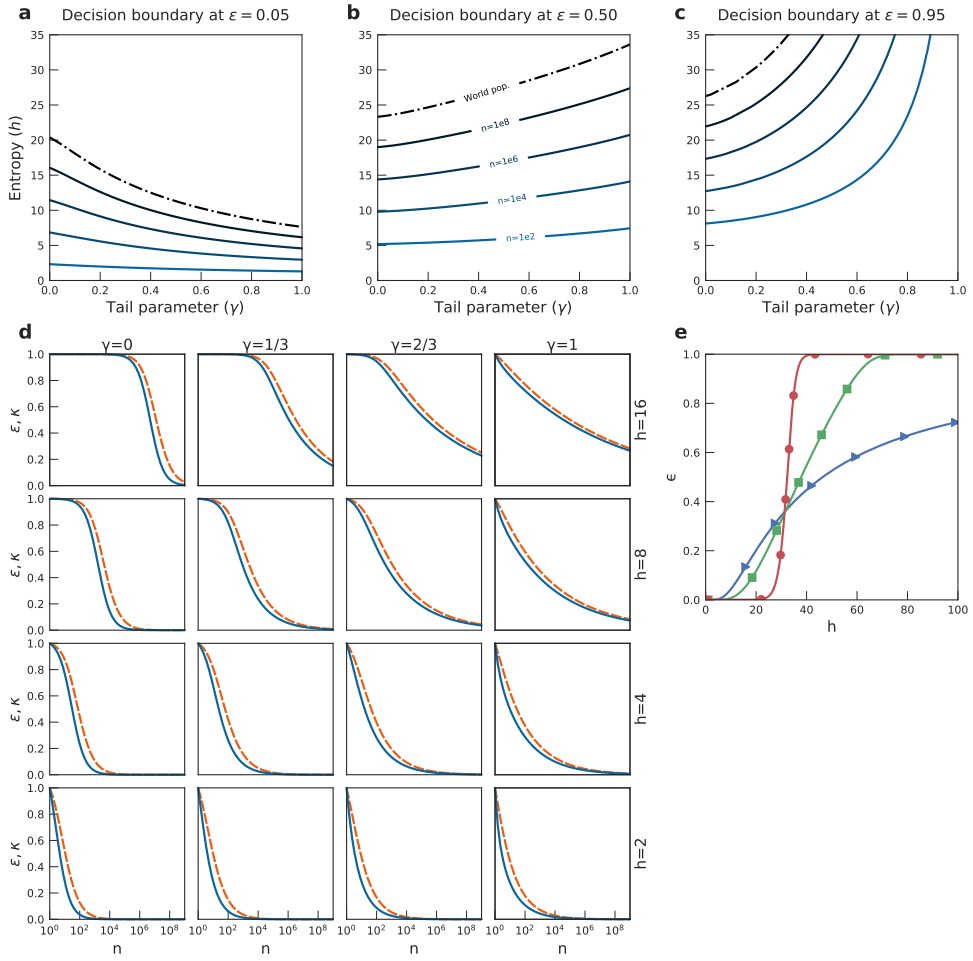


FIG. 6.2: Using the Pitman-Yor model to quantify the criticality and regimes of uniqueness and correctness. (a-c) Decision boundaries to obtain 5% (a), 50% (b), and 95% (c) of unique individuals, in populations ranging from 10^2 individuals (lowest blue line) to 7.5 10^9 individuals—the world population (semi-dashed line). (d) Functional forms of the expected uniqueness (blue solid line) and correctness (orange dashed line) for distributions with low to high entropy (y-axis) and geometric to power-law tails (x-axis). Each panel shows the uniqueness and correctness for a population ranging from 1 to 10^9 individuals (logarithmic scale). (e) Critical behaviors arise for exponential tails ($\gamma = 0$, red circle line) but not for power-law tails ($\gamma = 0.5$, green square; $\gamma = 1$, blue triangles).

second parameter, γ, γ capturing prior beliefs about tail behavior (independently of h):

$$h = \mathbb{E}\left[\mathbb{H}\left(X_A\right) \mid d, \alpha\right] = F(\alpha + 1) - F(1 - d)$$
$$\gamma = \frac{F(1) - F(1 - d)}{F(\alpha + 1) - F(1 - d)}$$

with F the digamma function.

The FLP expresses ε and κ in term of the entropy h alone, hence neglecting any effet of the shape the distribution. With the PY model, the tail parameter γ captures crucial variations of the uniqueness and correctness. For example, geometric ($\gamma \approx 0$) distributions yield a sustained high score (the higher the entropy, the longer the sustain) followed by a a fast decay, while distributions with power-law tails ($\gamma > 0$) show no sustained score but a slow decay (slower as the entropy increases). Intuitively, for the same entropy, draws from power-law distributions are more likely to come from the tail, often yielding new unique outcomes.

The PY model correctly predicts the subcritical and supercritical regimes of uniqueness. The frequency distributions X_A with an exponential tail ($\gamma \approx 0$) exhibit a critical uniqueness behavior, absent from heavy tails ($\gamma > 0$). A small increase of entropy yields a catastrophic increases of uniqueness, from almost 0 to almost 100%. This criticality does not occur with a power-law tail: a small increase of entropy yields a small increase of uniqueness. The criticality of re-identications further contradicts the common belief that entropy alone is a sufficient predictor of uniqueness or correctness.

Overall, the fundamental assumption of uniformity within the FLP yields poor estimates for the data collections studied here. Worst, this assumption of uniformity does not even provide a conservative bound on the uniqueness: the FLP can (incorrectly) predict the uniqueness to be above or below the actual empirical uniqueness.

6.2.3 *Anonymity bounds from summary statistics*

Having derived accurate laws that govern the risk of re-identification, we now propose a principled information theory framework to combine the PY model with few key statistics about a data collection, and bound the overall uniqueness.

For many data collections, the complete distribution of auxiliary information X_A is unknown, but few summary statistics can often be gathered regarding the marginal distributions or the association structure between them. To estimate the overall likelihood of re-identification, we propose to combine the developed PY model with information-based statistics.

Below, we propose three audit scenarios combining marginal entropies and pairwise association between variables to achieve coarse (with few statistics) to very accurate bounds (with a few more statistics) for ε and κ. Inspired by recent advances in linear optimization, we estimate the joint entropy h of the reparametrized model, using optimization on the entropic cone, a scalable method to integrate partial knowledge about marginals and association between them.

Recall that we consider a multivariate data collection D, with auxiliary information $X_A = (X_1, X_2, \ldots, X_d)$ without loss of generality. For every subset $\mathcal{S} \subseteq [d]$, let $X_\mathcal{S}$ be the random vector $(X_i)_{i\in\mathcal{S}}$ and the entropic vector $H(X_\mathcal{S})$ its associated Shannon entropy:

$$\mathbb{H}(X_\mathcal{S}) \equiv -\sum_{x_\mathcal{S}} p(x_\mathcal{S}) \log_2 p(x_\mathcal{S})$$

The entropic set $\mathcal{H}_d$ is the set of all possible entropic vectors [215, 216]:

$$\mathcal{H}_d \equiv \left\{ h \in \mathbb{R}^{2^d-1} \mid \exists Y_1, Y_2, \ldots Y_d, \forall \mathcal{S} \subseteq [d], h_\mathcal{S} = \mathbb{H}(Y_\mathcal{S}) \right\} \quad (6.3)$$

Its closure is known to be a convex cone, albeit its precise form is still not explicitly understood for $d \geq 4$. We use an outer approximation of the entropic cone, using a finite number of explicit Shannon-type inequalities (monotonicity and strong subadditivity of the entropy), to bound the total entropy of X_A (see section 6.4.3).

Based on any set of constraints over marginals of X_A or pairs of marginals, we solve two linear programs to estimate both the minimum and maximum (marginalized) entropy of X_A:

$$h_{\max} \equiv \max_C \ \mathbb{H}(X_1, X_2, \ldots, X_d)$$
$$h_{\min} \equiv \min_C \ \mathbb{H}(X_1, X_2, \ldots, X_d)$$

with C the combination of Shannon-type inequalities on the entropic cone and prior knowledge about the auxiliary information.

Combining the PY model and the optimization on the entropic cone, we obtain minimum and maximum bounds on the expected uniqueness and correctness with, e.g., $\varepsilon_{\max} = \mathbb{E}[\varepsilon \,|\, h_{\max}, \gamma]$ and $\varepsilon_{\min} = \mathbb{E}[\varepsilon \,|\, h_{\min}, \gamma]$.

In practice, we study three standard audit scenarii for prior knowledge (see Section 6.4.4), for which are known:

H 1 all attribute entropies $\mathbb{H}(X_i)$ and all mutual information scores $I(X_i; X_j)$, as well as the overall tail parameter γ;

H 2 all attribute entropies but only the average mutual information $\langle I(X_i; X_j)\rangle$, as well as the approximate tail parameter $\widetilde{\gamma}$ (low, medium, or high);

H 3 the sole number of symbols for each attribute (e.g., 2 potential outcomes for gender, 110 for age).

To audit the anonymization of a data collection, the scenario H3 relies on easy-to-guess information, while the scenario H1 requires further information that can be gathered from released statistics or previous measurements, yet possible without access to collected data.

The Table 6.2 reports the performance of H1, H2, and H3 to accurately bound ε and κ, with two measures: the gap $\Delta\varepsilon_{\min}$ (resp. $\Delta\varepsilon_{\max}$) between the upper (resp. lower) bound and the empirical uniqueness. An uninformation scenario would results in bounds $(\varepsilon_{\min}, \varepsilon_{\max}) = (0, 1)$ and gaps of approximately 50%. On the contrary, a data subject evaluating their own risks using, e.g., the scenario H2 will obtain close upper bounds on uniqueness, on average only 13.9 percentage points above the exact uniqueness.

Our results suggest that three statistics are therefore sufficient to achieve accurate minimum and maximum bounds for uniqueness: (i) the marginal entropies h_i for all variables, (ii) the average mutual information between them, and (iii) the approximate shape of the joint tail γ.

6.2.4 *Application to device fingerprints*

To illustrate the predictive power of the developed model, we apply it to test the anonymity of technical device identifiers, also called device or browser fingerprints. Web browsers share device-specific

	Marginals	Association	$\Delta\varepsilon_{\max}$	$\Delta\varepsilon_{\min}$	$S(\varepsilon)$	TP_{low}	TP_{high}
H1	Entropy	MI	0.100	0.238	95.3%	80.0%	85.9%
H2	Entropy	Average MI	0.139	0.263	98.8%	60.0%	49.3%
H3	NbSymbols	No Prior	0.394	0.475	100%	10.0%	0.0%
H$\varnothing$	No prior	No prior	0.512	0.487	100%	0.0%	0.0%

TABLE 6.2: Few summary statistics still correctly guess the overall uniqueness. For all surveyed datasets, we study three scenarios: informative and complex (H1), informative and simple (H2), uninformative and simple (H3), as well as a non informative prior H$\varnothing$ ($\varepsilon_{\min} = 0$ and $\varepsilon_{\max} = 1$). The metric $\Delta\varepsilon_{\max}$ (resp. $\Delta\varepsilon_{\min}$) measures the gap between the upper (resp. lower) bound and the empirical uniqueness. We report the agreement $S(\varepsilon)$ between the bounds and the empirical uniqueness as the fraction of datasets for which $\varepsilon_{\min} \leq \varepsilon \leq \varepsilon_{\max}$. TP_{low} measures the fraction of low uniqueness scores ($\varepsilon < 0.10$) correctly predicted as low ($\varepsilon_{\max} < 0.10$). TP_{high} measures the fraction of high uniqueness scores ($\varepsilon > 0.90$) correctly predicted as high ($\varepsilon_{\min} > 0.90$). Overall, the scenario H2 requires little information yet obtains accurate theoretical bounds, close to the performance of the scenario H1.

technical information with websites, such as identifiers about the browser, operating system, or even the timezone. Combined together, such identifiers form fingerprints that can potentially uniquely identify devices across websites and visits.

Due to the inherent difficulty of gathering the fingerprints of all devices worldwide, it is challenging to estimate what is the actual correctness of a fingerprint without a statistical model. Researchers have shown that, up to 2 million devices, device fingerprinting could uniquely identify from 33.6% to 80% of studied devices, for various dataset size and class of fingerprints [102, 217]. However, despite these studies, there is still no consensus on whether fingerprints can track more than 2 million devices and correctly identify them. This is also pinpointed by our findings: due to the potential critically of correctness, a high correctness for 2 million devices predicts little about the correctness of 2 billion devices.

Here, our model instead predicts that that simple fingerprints can actually uniquely identify almost all devices used online. First, using a sample of few thousand fingerprints, we estimate using the PY model, the overall uniqueness of billions of devices. Combining basic browser and device technical information could correctly identify 74.6% out of 4 billion users worldwide (Fig. 6.3). Even iOS devices, purportedly harder to fingerprint, yield high scores: out of 1 billion active devices, 27.76% could be correctly identified with the same set of attributes.

Overall, the PY model offers practitioners a robust framework to model the risk associated with re-identification, and audit the anonymization of data collections. Whether data controllers disclose enough summary statistics from the data (scenario H1) or auditors collect measurements from previous work and similar collections (scenario H2), the framework provide a rigorous way to test for anonymity.

6.2.5 *Application to set-valued data*

Finally, while we have until now focused on simple tabular data (few collected attributes), our model scales to and captures the uniqueness of individuals and correctness of re-identifications in modern high-dimensional set-valued data. Set-valued data collections contain, for each individual, a set of data points such

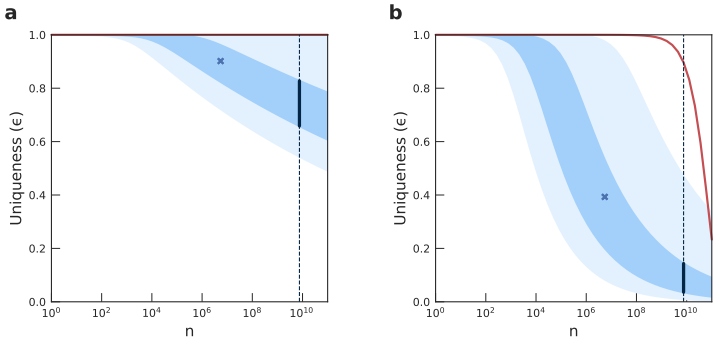


FIG. 6.3: Little prior knowledge sufficiently bounds the resulting uniqueness of device fingerprints. The light blue (resp. dark blue) range reports the predictions associated with the H2 (resp. H1) scenario. The panel (b) shows the uniqueness when collecting the HTTP accept, whether cookies and JavaScript is enabled, timezone, display size, and the panel (a) when the installed fonts and plugins are also collected. The red solid line depicts the uniqueness predicted by a uniform ATK model, based on marginal entropies.

as, e.g., installed apps on a smartphone, shopping patterns, movie ratings, or mobility traces [35–37, 41].

We use the PY model to quantify the likelihood, for an adversary knowing p points of auxiliary information (visited, installed, watched, *etc.*) matching an individual record x, that the individual can be uniquely identified or their record correctly matched. We denote by ε_p (resp. κ_p) the expected uniqueness (resp. correctness) knowing p points from an individual record:

$$\varepsilon_p \equiv \mathbb{E}_{x\sim X,|x|\geq p}\left[\binom{|x|}{p}^{-1}\sum_{\mathcal{S}\subseteq x,|\mathcal{S}|=p}\varepsilon[X_{\mathcal{S}}]\right]$$

and

$$\kappa_p \equiv \mathbb{E}_{x\sim X,|x|\geq p}\left[\binom{|x|}{p}^{-1}\sum_{\mathcal{S}\subseteq x,|\mathcal{S}|=p}\kappa[X_{\mathcal{S}}]\right]$$

with $\varepsilon[X_{\mathcal{S}}]$ and $\kappa[X_{\mathcal{S}}]$ the uniqueness and correctness restricted to the subset $\mathcal{S}$. These two quantities measure the mean uniqueness (resp. correctness) across all traces of p points from an individual record (across all records with at least p points). The expected uniqueness knowing p points is also called unicity. Introduced by

de Montjoye *et al.* to study mobile phone data [36], it has since been further applied to credit card transactions [37] and installed mobile apps [41].

Under two simplifying assumptions, both can be modeled by the expected uniqueness and correctness for the PYP prior

$$PY(h, \gamma_0) \text{ where } h = h_0 \frac{(\rho/2)^p - \rho/2}{\rho/2 - 1}$$

with h_0 (resp. γ_0) the entropy (resp. tail parameter) knowing a single data point and ρ the average normalized mutual information between data points [218]. First, we set a prior $PY(h_0, \gamma_0)$ on each marginalized data point T_i, with $X = T_1 \times T_2 \times \ldots \times T_d$ the joint distribution over all d points. Second, we assume that all pairs of points (T_j, T_k) are correlated with the same correlation coefficient $\rho(T_j, T_k) = \rho_0$. These strong assumptions are necessary to obtain a tractable close form expression of ε_p and κ_p, yet accurately model the data collections we study below.

Corpus	$p = 1$	$p = 2$	$p = 4$	$p = 8$	$p = 16$
APPS	0.045	0.038	0.026	0.011	0.010
CALLS [60min]	0.054	0.075	0.078	0.052	0.054
CALLS [10min]	0.061	0.078	0.041	0.009	0.009
SHOPS	0.032	0.038	0.005	0.000	

TABLE 6.3: RMSE of the PY-SV model to predict ε_p on the APPS, CALLS, and SHOPS corpus. The CALLS corpus is divided in two, the first with calls known with a 1h time window, the second with a 10min window. Overall, the PY-SV model obtains a low RMSE across number of known attributes (p) and corpora.

We test the resulting model, called PY-SV, on three additional datasets: installed Android apps (APPS), time of calls (CALLS), shopping patterns (SHOPS). From the data, we estimate parameters $\widehat{h_0}$, $\widehat{\rho}$, and $\widehat{\gamma_0}$ independently of p. Fig. 6.4 show that the model accurately predicts the correctness of re-identification from $p = 1$ to $p = 16$ on the APPS dataset. Table 6.3 further reports the low RMSE across the studied corpora.

The three parameters h_0, ρ, and γ_0, along with the poulation size n, explain 98% of the variance of correctness, with low RMSE overall

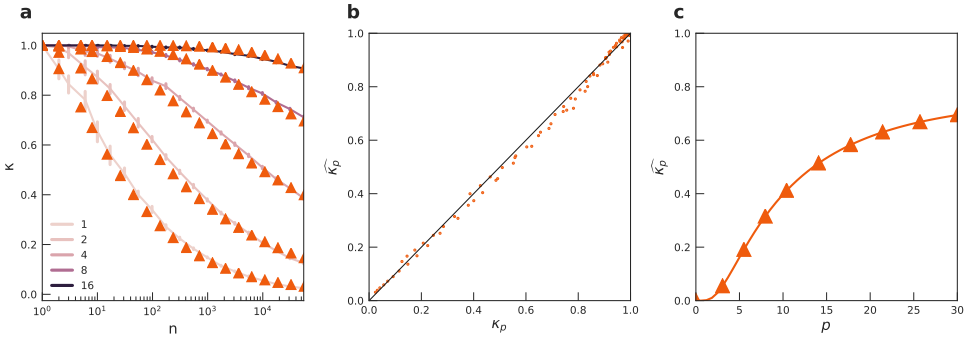


FIG. 6.4: Empirical validation of the PY-SV model on the APPS dataset. (a) Empirical correctness κ_p when knowing 1, 2, 4, 8, and 16 data points (solid lines with s.d. error bars), for a population size n ranging from 10^0 to 10^5 individuals. The triangles report the estimated correctness $\widehat{\kappa_p}$ of the PY-SV model. (b) Empirical vs. estimated correctness across all values of p. (c) Predicted correctness $\widehat{\kappa_p}$ as the number of available data points p increases, for 2 billion mobile devices.

(0.02 for APPS, 0.02 for SHOPS, 0.04 for CALLS). We perform a global sensitivity analysis by using the Sobol method, which accounts for individual effects and interactions, as well as nonlinear responses. We find that all parameters have a significant effect on the correctness κ_p. The population size ($\log_2 n$) accounts for 36.7% of the total effects, the entropy h_0 for 54.3%, the tail parameter γ_0 for 6.3%, the correlation ρ for 20.6%, and the number of points p for 22.6%. There are little interactions between parameters, the highest being between $log_2 n$ and h_0 (3.2%), suggesting low redundency between them.

6.3 SUMMARY

The proposed PY and PY-SV models provide, for the first time, a fundamental model for the uniqueness of individuals and correctness of re-identifications. In the field of re-identification science, previous approach to quantify anonymity assumed uniform individual effects, resulting in uniform distributions of auxiliary information. Studying a wide range of distributions from tabular census and survey data to high-dimensional set-valued data, we show that most of these data follow non-uniform frequency distributions, instead closely captured by Pitman-Yor processes. These processes provide a tractable model of uniqueness and

correctness, correctly predicting their critical behaviors. Using linear optimization on the entropic cone, we propose a new framework, combining easy-to-learn summary statistics about the collected information with the developed models, to obtain upper and lower bounds for the expected uniqueness and correctness.

Our approach can be directly used to audit current data collections, for which access to the data is often impossible to due legal and technical constraints. The model can further be used to approximately estimate the anonymity offered to individuals when participating in sensitive data collections.

6.4 ADDITIONAL METHODS

6.4.1 *Uniqueness and correctness for Pitman-Yor processes*

Pitman and Yor proved in 1997 the following proposition:

Proposition 1. *For $X \sim PY(d, \alpha)$, a countably-infinite distribution $(x_i)_{1 \leq i \leq +\infty}$, and f a continuous function on $[0, 1]$:*

$$\mathbb{E}_{(X|d,\alpha)}\left[\sum_{i=1}^{+\infty} f(x_i)\right] = \mathbb{E}_{(\tilde{\pi}_1|d,\alpha)}\left[\frac{f(\tilde{\pi}_1)}{\tilde{\pi}_1}\right]$$

with $\tilde{\pi}_1$ the first size-biased sample from X, $\tilde{\pi}_1 \sim Beta(1 - d, d + \alpha)$.

This allows us to obtain close-form expressions for the expected entropy, uniqueness, and correctness of X.

Proposition 2. *For $X \sim PY(d, \alpha)$, the expected entropy is:*

$$\mathbb{E}\left[\mathbb{H}(X) \,|\, d, \alpha\right] = \digamma(\alpha + 1) - \digamma(1 - d)$$

Proof. See the proof by Archer *et al.* [211]. □

Proposition 3. *For $X \sim PY(d, \alpha)$, the expected uniqueness is:*

$$\mathbb{E}\left[\varepsilon \,|\, d, \alpha\right] = \frac{\Gamma(\alpha + 1)}{\Gamma(d + \alpha)} \frac{\Gamma(n + d + \alpha - 1)}{\Gamma(n + \alpha)}$$

Proof. We use the Proposition 1 with $f(x) = x\,(1-x)^{n-1}$:

$$\begin{aligned}
\mathbb{E}\left[\varepsilon \,|\, d, \alpha\right] &= \frac{1}{B(1-d, d+\alpha)} \int_0^1 \frac{f(x)}{x} x^{-d} (1-x)^{d+\alpha-1}\,\mathrm{d}x \\
&= \frac{1}{B(1-d, d+\alpha)} \int_0^1 \frac{x\,(1-x)^{n-1}}{x} x^{-d} (1-x)^{d+\alpha-1}\,\mathrm{d}x \\
&= \frac{1}{B(1-d, d+\alpha)} \int_0^1 x^{-d} (1-x)^{n-1+d+\alpha-1}\,\mathrm{d}x \\
&= \frac{\Gamma(1+\alpha)}{\Gamma(1-d)\Gamma(d+\alpha)} \frac{\Gamma(1-d)\Gamma(n+\alpha+d-1)}{\Gamma(n+\alpha)} \\
&= \frac{\Gamma(1+\alpha)\Gamma(n+d+\alpha-1)}{\Gamma(d+\alpha)\Gamma(n+\alpha)}
\end{aligned}$$

□

Proposition 4. *For $X \sim PY(d, \alpha)$, the expected correctness is:*

$$\mathbb{E}\left[\kappa \,|\, d, \alpha\right] = \frac{1}{nd}\left(\frac{\Gamma(1+\alpha)}{\Gamma(d+\alpha)} \frac{\Gamma(n+d+\alpha)}{\Gamma(n+\alpha)} - \alpha\right)$$

Proof. We use the Proposition 1 with:

$$\begin{aligned}
f(x) &= x\,\frac{1}{n\,x}\left(1-(1-x)^n\right) \\
&= \frac{1}{n}\left(1-(1-x)^n\right)
\end{aligned}$$

so that:

$$\begin{aligned}
\mathbb{E}\left[\kappa \,|\, d, \alpha\right] &= \frac{1}{B(1-d, d+\alpha)} \int_0^1 \frac{f(x)}{x} x^{-d} (1-x)^{d+\alpha-1}\,\mathrm{d}x \\
&= \frac{1}{B(1-d, d+\alpha)} \int_0^1 \frac{1}{x}\frac{1}{n}\left(1-(1-x)^n\right) x^{-d} (1-x)^{d+\alpha-1}\,\mathrm{d}x \\
&= \frac{1}{n}\frac{1}{B(1-d, d+\alpha)} \int_0^1 x^{-d-1}(1-x)^{\alpha+d-1} - x^{-d-1}(1-x)^{n+\alpha+d-1}\,\mathrm{d} \\
&= \frac{1}{n}\frac{\Gamma(1+\alpha)}{\Gamma(1-d)\Gamma(d+\alpha)}\left(\frac{\Gamma(-d)\Gamma(d+\alpha)}{\Gamma(\alpha)} - \frac{\Gamma(-d)\Gamma(n+d+\alpha)}{\Gamma(n+\alpha)}\right) \\
&= \frac{1}{n}\frac{\Gamma(1+\alpha)}{(-d)\,\Gamma(d+\alpha)}\left(\frac{\Gamma(d+\alpha)}{\Gamma(\alpha)} - \frac{\Gamma(n+d+\alpha)}{\Gamma(n+\alpha)}\right) \\
&= \frac{1}{nd}\left(\frac{\Gamma(1+\alpha)}{\Gamma(d+\alpha)} \frac{\Gamma(n+d+\alpha)}{\Gamma(n+\alpha)} - \alpha\right)
\end{aligned}$$

□

6.4.2 *Entropy, uniqueness, and correctness for uniform distributions*

Let F be a uniform distribution over d distinct outcomes, with $\mathbb{P}(F = i) = 1/d$. We have:

$$\begin{aligned}\mathbb{H}(F) &= -\sum_{i=1}^{d} \mathbb{P}(F = i) \log_2(\mathbb{P}(F = i)) \\ &= \log_2 d\end{aligned}$$

In a population of n individuals, the expected correctness is

$$\begin{aligned}\mathbb{E}[\kappa] &= \sum_{i=1}^{d} \mathbb{P}(F = i) \frac{1}{n\ \mathbb{P}(F = i)} \left(1 - (1 - \mathbb{P}(F = i))^n\right) \\ &= \frac{d}{n}\left(1 - \left(1 - \frac{1}{d}\right)^n\right) \\ &= \frac{2^h}{n}\left(1 - \left(1 - 2^{-h}\right)^n\right)\end{aligned}$$

and the expected uniqueness

$$\begin{aligned}\mathbb{E}[\varepsilon] &= \sum_{i=1}^{d} \mathbb{P}(F = i)\, (1 - \mathbb{P}(F = i))^{n-1} \\ &= \sum_{i=1}^{d} \frac{1}{d}\left(1 - \frac{1}{d}\right)^{n-1} \\ &= \left(1 - \frac{1}{d}\right)^{n-1}\end{aligned}$$

6.4.3 *Maximum a-posteriori (MAP) estimates from Pitman-Yor priors*

Give a sample $x = (x_i)_{1 \le i \le N}$ from an unknown distribution $X \sim PY(d, \alpha)$, we have:

$$p(x \mid d, \alpha) = \frac{\prod_{l=1}^{k-1}(\alpha + ld)\ \prod_{i=1}^{K} \Gamma(n_i - d)\ \Gamma(1 + \alpha)}{\Gamma(1 - d)^K\, \Gamma(\alpha + n)} \tag{6.4}$$

with $n = (n_i)_{1 \le i \le K}$ the observed frequency counts in x, and K the total number of frequency classes.

Numerically, computations for $\prod_{l=1}^{k-1}(\alpha + ld)$ can be accelerated by noting that:

$$\prod_{l=1}^{k-1}(\alpha + ld) = d^{K-1}\prod_{l=1}^{k-1}(\alpha/d + l)$$
$$= d^{K-1}\frac{\Gamma(\alpha/d + K - 1)}{\Gamma(\alpha/d + 1)}$$

Recall that a PYP is equally parametrized by (d, α) and (h, γ). We place a uniform prior on h and set an exponential prior $q(\gamma)$ on γ:

$$q(\gamma) = \exp(-\frac{10}{1-\gamma})$$

. We refer to Archer et al. [211] for a detailed study of the priors for PYP distributions.

OPTIMIZATION ON THE ENTROPIC CONE The entropic set $\mathcal{H}_d$ is bounded by a combination of inequalities, the Shannon-type inequalities [219]. We denote by Γ_n the set of vectors that satisfies them. For random variables $X_1, X_2, \ldots, X_d$, their entropies satisfy the monotonicity inequalities:

$$\mathbb{H}(X_\varnothing) = 0$$

$$\forall I \subseteq J \subseteq [d],\ \mathbb{H}(X_I) \leq H(X_J)$$

as well as submodularity or strong subadditivity inequalities:

$$\forall I, J \subseteq [d],\ \mathbb{H}(X_I) + \mathbb{H}(X_J) \geq \mathbb{H}(X_{I\cup J}) + \mathbb{H}(X_{I\cap J}) \tag{6.5}$$

To bound the overall entropy $\mathbb{H}(X_{[d]})$, we encode these inequalities as a linear problem:

$$\begin{aligned} Q(c) = \text{maximize} \quad & c^T h \\ \text{s.t.} \quad & \mathrm{A}h \leq a \\ & h \geq 0 \end{aligned}$$

with $h \in (\mathbb{R}^+)^{2^d-1}$ a candidate entropic vector and $c \in \mathbb{R}^{2^d-1}$ the objective.

When given additional (in)equalities on the entropic set (e.g., $\mathbb{H}(X_i) = z$ or $I(X_i; X_j) = z$), we further incorporate them as

additional constraints $\mathrm{B}h \leq b$. Overall, we solve the following problem:

$$\begin{aligned} Q(c, \mathrm{B}, b) = \text{maximize} \quad & c^T h \\ \text{s.t.} \quad & \mathrm{A}h \leq a \\ & \mathrm{B}h \leq b \\ & h \geq 0 \end{aligned}$$

We obtain an upper bound $h_{\max} = Q(c_{\max}, \mathrm{B}, b)$ on $\mathbb{H}(X_{[d]})$ with $c_{\max} = (0, \ldots, 0, 1)$. We obtain a lower bound $h_{\min} = -Q(c_{\min}, \mathrm{B}, b)$ on $\mathbb{H}(X_{[d]})$ with $c_{\min} = (0, \ldots, 0, -1)$.

6.4.4 *Audit scenarios*

We study three audit scenarios H1, H2, H3.

Under the informative scenario H1, the marginal entropy of each attribute, the mutual information between each pair of attributes, and the overall tail parameter γ are known:

$$C[H_1] = \{\mathbb{H}(X_i) = h_i \,|\, i \in [d]\} \cup \{I(X_i; X_j) = i_{ij} \,|\, i \in [d], j \in [d]\}$$

and $\gamma[H_1] = \gamma$

Under the approximative scenario H2, the marginal entropy of each attribute, the average mutual information between pair of attributes, and the approximative tail parameter γ are known:

$$C[H_2] = \{\mathbb{H}(X_i) = h_i \,|\, i \in [d]\} \cup \{\langle I(X_i; X_j)\rangle = i\}$$

and $\gamma[H_2] \in [0, 1/3[$ if $\gamma \in [0, 1/3[$ (resp. $[1/3, 2/3[$ and $[2/3, 1]$)

Under the least informative scenario H3, only the alphabet size of each attribute are known:

$$C[H_3] = \{|X_i| = n_i \,|\, i \in [d]\}$$

and $\gamma[H_3] \in [0, 1]$. From $C[H_3]$, we obtain constraints on the marginal entropy of each attribute:

$$C'[H_3] = \{0 \leq \mathbb{H}(X_i) \leq \log_2 n_i \,|\, i \in [d]\}$$

Part III

ENGINEERING PRIVACY

7

MODERN PRIVACY-PRESERVING PLATFORMS

Parts of this chapter originate from a joint research article on the privacy-conscientious use of mobile phone data [48].

The balance between using personal data and preserving people's privacy has historically relied, both practically and legally, on the concept of data anonymization (see Chapter 4).

While de-identification algorithms are widely used in industry and academia to transform and release anonymous datasets, a large body of research has shown that these practices are not resistant to a wide range of re-identification attacks [18, 35–37, 168, 174]. Exposure of the limits of de-identification have led policy makers to call for stricter regulations: for instance, the [US] President's Council of Advisors on Science and Technology concluded that data anonymization "*is not robust against near-term future re-identification methods. PCAST do not see it as being a useful basis for policy*" [220].

In Chapters 5 and 6, we showed that de-identification is further limited by easier re-identification that once thought, and called for better technical solutions to disseminate data.

In this final part, we review potential privacy-conscientious solutions for data controllers to collect and disclose personal data moving forward, highlighting the research directions requiring further work before being implemented. In particular, we cover alternatives including releasing aggregated, summary statistics from personal data. In this approach, the data processor computes and releases answers—ahead of time or in real-time—to queries on the raw dataset, while the original dataset is never published.

Two use cases are finally discussed in this part. In Chapter 8, we demonstrate some of the concerns associated with such modern privacy-preserving platforms, using Diffix, a privacy-preserving database, as an example. In Chapter 9, we show how (aggregated) statistical insights can be computed and released at scale from sensitive mobile phone metadata.

7.1 FOUR MODELS TO RELEASE DATA

The failure of de-identification and the lack of a modern and agreed-upon framework to share data are strongly hindering the privacy-conscientious use of personal data, a valuable asset for researchers, policy makers, humanitarian workers, and companies.

We recently proposed four different models to share personal data for the big data era [48]. While focusing on the use of mobile phone metadata, our work can also be generalized to various data collections. Our suggestions, summarized below and in Fig. 7.1, form a layered approach to privacy, combining security practices with data reduction and access control.

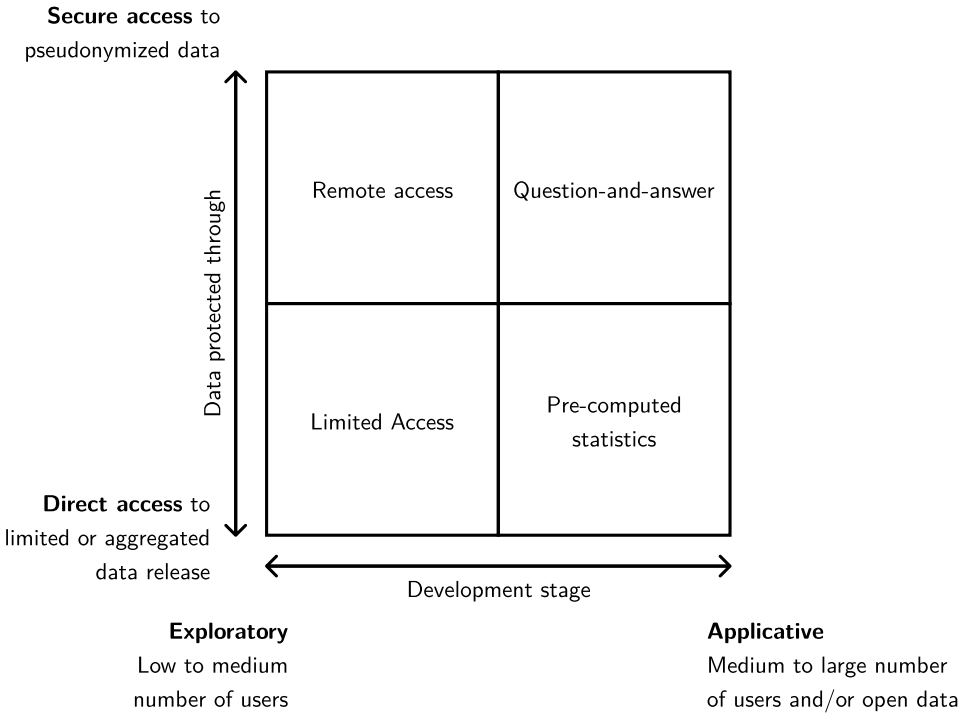


FIG. 7.1: Matrix of the four models for the privacy-conscientious use of personal data.

7.1.1 *Limited Release*

This first model builds upon the traditional de-identification framework. Personal data are transformed by the data collector and a copy is given to third parties under a legal agreement. Transformations can include data coarsening or sampling, with minimal changes to the data in order to maximize statistical significance. As discussed in Chapter 4, these pseudonymization techniques are not necessarily enough to anonymize the underlying data, often requiring explicit consent from the data subjects for their data to be shared. Appropriate non-disclosure and data use agreements (DUA) are therefore required.

This is the model used by national statistics to release census microdata at large sampling fractions or by the IRS to release sensitive individual-level tax records to trusted researchers [221]. Upon signature of a legal agreement, researchers get access to accurate pseudonymized data. This model is therefore slow to implement and difficult to scale to more than a few third parties.

The Limited Release model share the same membership and inference vulnerabilities we discussed earlier. We consider re-identification using auxiliary information to be the main anonymity threat in this model: re-identification would allow an attacker to link the released data about one participants to their identities. This model should therefore be limited to exploratory analysis, inside a company or with trusted parties.

7.1.2 *Precomputed Statistics*

Instead of releasing raw data, aggregating or deriving synthetic data aims to disclose only anonymous data, thus not requiring the explicit consent of participants. Statistics can be computed at the individual level (*e.g.*, number of calls or radius of gyration for mobile phone metadata, average rating for car sharing) or aggregated across individuals (*e.g.*, age distribution, origin-destination matrices). Static aggregation is a standard procedure used to disclose insights from sensitive data. It is used, *e.g.*, by the US Census Bureau to release vast amounts of accurate demographic and socio-economic statistics, from simple counts to multiway contingency tables. Because individual indicators or aggregated statistics do not relate

directly to personal data, they often offer a higher level of privacy compared to pseudonymous or de-identified data.

Both approaches have several advantages over individual-level anonymization. Firstly, aggregate statistics are often enough to perform a large array of analyses, including advanced machine learning use cases [222, 223]. The individual-level data are not shuffled, coarsened, infused with noise, or subsampled, which helps preserve the utility of any statistical aggregation method. This is particularly true for high-dimensional data—*e.g.*, mobility data—for which a robust anonymization method will strongly damage the utility [36, 134]. Finally, anonymity can be assumed to be better protected, as any individual data is putatively hidden in the aggregate summary (as with k-anonymity for instance). A malicious analyst can still combine aggregated statistics to gain enough level to break the purposely anonymity. Recent work has shown that both aggregated and synthetic data can still contain enough information to perform membership or inference attacks. Researchers have shown that releasing too many, overly accurate, summary statistics can breach participants' privacy [150, 151]. These results show that, without theoretical guarantees, it is increasingly difficult to draw the line between sensitive and non-sensitive aggregated or synthetic data [123].

Secondly, synthetic data can fasten research, by letting researchers quickly access data in the same format as sensitive data without the associated risks of de-identifying data. Synthetic approaches aim to provide better privacy guarantees since no individual data is released; sufficient care need to be taken to provide enough utility, by balancing the complexity of the underlying data generation model with the privacy guarantee offered. While synthetic data is an appealing approach to mitigate the risk of disclosing sensitive information, it lacks theoretical guarantees and membership privacy may still be breached by using too complex models that encode too much information [224].

The Precomputed Statistics model should therefore be reserved to release well-established metrics of interests, such as demographics indicators for census and survey data, or density and mobility maps from human behavior data.

7.1.3 *Remote Access*

When data are too sensitive to be released, and when analysts must run a large variety of analyses, analysts can directly access the data within the premise of the data controller. The controller supervises who access the data, how the data is being used, and ensures that no personal data leaves the remote server (legally, with non-disclosure agreements, or technically, by monitoring the executed commands and the returned outputs).

The Remote Access model requires, similarly to the Limited Access, trust in the parties accessing the data, and signature of legal agreements. It also often includes a physical infrastructure (*e.g.*, physical 2-factor access tokens, dedicated authentication hardware using fingerprints) and a secure network environment (VPN for remote access, DMZ, *etc.*).

Due to the higher operational costs, this model is often limited to trusted and experimented researchers or statisticians who need to perform ad-hoc experiments on the raw data. For instance, in France, the Secure Access Data Center (CASD) grants researchers access to sensitive health data through a remote access interface, under a case by case access policy[1].

7.1.4 *Interactive and decentralized computations*

Finally, other techniques can be used to analyze data remotely without accessing the raw data itself but by providing instructions or algorithms to be run on the data. Interactive aggregation, query-based, or question-and-answer (QA) systems let analysts send queries and potentially run their own code within the premise of the data controller. They offer data analysts a (remote) platform to ask questions that return data aggregated on-the-fly from several, potentially many, records.

Compared to the Remote Access model, interactive aggregation platforms do not grant access to the underlying data, but only to statistics computed on the data. Compared to the Precomputed Statistics model, interactive platforms let third parties adjust query parameters or run their own code on the protected infrastructure

1 Centre d'accès sécurisé aux données (CASD), https://www.casd.eu/.

holding the data. QA platforms are therefore often sought by data controllers since they require less infrastructure and operational costs, yet aim to disclose complex statistics. Interactive aggregation is, *e.g.*, used by Uber to release origin-destination traffic patterns via their Uber Movement platform[2].

In a typical Question-and-Answers platform, the individual-level (often pseudonymous) data is stored on the data controller's server. Users access the data exclusively by sending queries that only return information aggregated from several records. Interactive aggregation can be based either on trusted third parties running the computations or on formal methods to execute computations in a decentralized and privacy-preserving way (*e.g.*, secure multi-party computation, homomorphic encryption, functional encryption, trusted computing platforms such as Intel SGX) while keeping the input data private. The latter provide additional privacy guarantees, especially during commmunication, but are still prone to side channel attacks or privacy leakage attacks.

Depending on the setting, an adversary can for instance easily submit a series of seemingly innocuous queries whose outputs, when combined, will allow them to infer private information about participants in the dataset. Interactive aggregation indeed often offer a more expressive syntax than already computed aggregated data. This approach shares similar privacy risks with the Precomputed Statistics model: too many, too precise answers can reveal personal data. Most aggregation platforms therefore add a random amount of noise to each answer, either independant [225] or correlated between answers [226]. While these safeguards may prevent traditional re-identification attacks on anonymized data [18, 35–37, 168, 174], a large range of attacks on such query-based systems have been developed since the late 70's [227, 228].

Most of these attacks show how to circumvent privacy safeguards (for instance, query set size restriction and noise addition) in specific setups. In 2003, Dinur *et al.* proposed the first example of an attack that works on a large class of aggregation scenarios [229]. In what they called a reconstruction attack, they showed that if the noise added to every query is at most $o(\sqrt{n})$, where n is the size of the dataset, then an attacker can reconstruct almost the entire dataset using only polynomially many queries. Sankararaman

2 Uber Movement, https://movement.uber.com/.

et al. performed the first formal study of membership attacks (see Section 4.2.1), introducing a theoretical attack model based on hypothesis testing [144] to infer whether the data about a certain individual is present in the protected dataset. Numerous reconstruction and membership attacks have been proposed in the literature. A recent survey from Dwork *et al.* gives a detailed overview of attacks on interactive aggregation systems [149].

The privacy offered by such systems therefore highly depends on the implementation. Few platforms rely entirely on provable theoretical guarantees (*e.g.*, the recent collaboration between Facebook and Social Science One aiming to release differentially privacy statistics results to researchers[3]) but such guarantees are often difficult to develop and implement correctly. Without provable guarantees, increased security and practical defenses can hinder re-identification attempts and are often used instead.

7.2 PRIVACY MOVING FORWARD

Releasing aggregated instead of individual data has emerged as the major alternative to de-identification. But its implementation, as we discussed above, lacks precise privacy guarantees. Too many aggregated statistics, combined, start revealing personal data. This is increasingly true as powerful machine learning approaches are being developed to perform reconstruction or membership attacks [150], requiring less and less aggregated data to perform re-identification attacks.

Two distinct defense avenues could help enforce privacy regulations: provable guarantees and practical security defenses. Provable guarantees, based on theoretical mathematical grounds, provide protections against a wide range of attack, although they can be difficult to implement while keeping a high utility. Security-based defenses for interactive platforms, based on practical secure design principles, instead do not provide theoretical guarantees but rely on an adversarial engineering approach. As no physical system is perfect, they rely on a *penetrate-and-patch* cycle along with transparent and open designs.

3 Social Science One, "Our Facebook Partnership", https://socialscience.one/our-facebook-partnership.

7.2.1 *Provable guarantees and heuristics*

To protect aggregation methods, privacy research has been increasingly focused on providing provable privacy guarantees to defend interactive systems against membership and inference attacks. Differential privacy is the main privacy guarantee considered by researchers so far [123, 230].

DIFFERENTIAL PRIVACY Differential privacy, now a fast-growing field in computer security, essentially provides provable guarantees when releasing data, with respect to past and future successive data releases. It is so far the only framework to provide such guarantees [123]. Intuitively, a randomized algorithm is differentially private if its output does not depend on any single individual's record in the dataset. Adding or removing a single individual data point should not significantly modify the output, hence protecting participants from any membership and inference attack. This privacy guarantee can be enforced by Precomputed Statistics and Question-and-Answer platforms, and has been mathematically proven to be robust against a very large class of attacks [231] when used with an appropriate privacy *budget* ε. However, research has shown that attacks on *implementations* of differentially private systems exist [232–234]: (a) while the theory might be secure, implementations can always result in software bugs, (b) adversaries can leverage side-channel attacks and privacy leakages not covered by the original threat model.

Furthermore, efficient differential privacy mechanisms are generally developed for a specific use case and, even if a large range of differentially private mechanisms have been proposed, there is still no practical widely deployed differential privacy solution for general-purpose data analytics [235]. Platforms based on differential privacy have been proposed, the main ones being PINQ [236], wPINQ [237], Airavat [238], and GUPT [239]. All of these systems however present limitations, *e.g.*, simplicity of use and support for various operators that join different datasets [235]. In 2017, Johnson *et al.* [235] proposed a new framework for general-purpose analytics, called FLEX, developed in collaboration with Uber. FLEX enforces differential privacy for SQL queries without requiring any knowledge about differential privacy from the analyst. However, the

actual utility achieved by the current implementation of FLEX has been questioned [240].

To achieve strong privacy guarantees, practitioners must often sharply limit the data utility by adding large amounts of noise and restrict the total number of requests that the system is allowed to answer [241]. Moreover, while differential privacy is a strong guarantee, the privacy risks of data breaches associated to implementation issues remain, exposing systems to attacks that differential privacy should in theory rule out [242, 243]. Overall, with some rare exceptions, the complexity of correctly implementing differential privacy and choosing the right privacy budget often prevents practitioners from using it.

DATA-DEPENDENT NOISE Instead of relying on theoretical mechanisms (*e.g.*, adding noise to enforce differential privacy), researchers have suggested to use heuristics to improve the privacy-utility balance of differential privacy. For example, the Diffix algorithm, which we study in Chapter 8, adds dynamic noise to every answer depending on the query set (*i.e.* the set of users selected by the query), and hence on the data. Data-dependent noise, also called instance-based noise, has been shown to provide significantly better accuracy than data-independent noise [244]. However, naive implementations of data-dependent noise can leak information about the data, a result Nissim *et al.* theorized as a potential way to attack these systems [244].

7.2.2 *Practical defenses*

In this section, we briefly outline some of the approaches that may be used to mitigate the effects of vulnerabilities against privacy-preserving (interactive) platforms. Overall, it is our belief that practical secure design principles apply here just as they do in many other contents. Specifically, privacy-preserving platforms (regardless of whether they have provable guarantees or not) would benefit from a defense-in-depth approach, *e.g.*, by monitoring the query stream for queries that are likely to lead to exploitation or auditing the computations and the accesses.

INTRUSION DETECTION The set of queries generated by a specific attack vector often follows a determined template. Learning this pattern may help prevent privacy leakage attacks, as well as potentially related attacks. A more sophisticated attacker might however vary the shape of the queries and interleave them with other more natural-looking queries, including over long periods of time.

AUDITABILITY If all users of a platform are authenticated, then a suspicious-looking query stream can lead to temporary account suspension and further investigations of the suspicious activity, including after happening, as new attacks are being uncovered.

INCREASED FRICTION Another strategy involves imposing time delays, rate limiting, or financial charges on queries, for instance by charging by the number of queries, instead of using a subscription-based model. This strategy can be refined to, for instance, charge more or create longer delays for suspicious queries. This would make it more difficult to automate the inference process at scale.

LIMITED EXPRESSIVENESS Instead of a rich syntax, an aggregation mechanism could allow only for a small set of conditions that are easier to validate. This could also include a limit to the number of conditions per query, or to the total number of conditions that may be used by the authenticated system user during a specific interval of time. This involves a compromise between rough data summarization and fine-grained queries, and limits the utility of the system in practice.

8

CASE STUDY: EXPLOITING DIFFIX'S STICKY NOISE

This chapter originates from a joint research article co-leaded by myself, Andrea Gadotti, and Florimond Houssiau [49]. The research was conducted during my stay at Imperial College London in 2018.

Diffix, a patented commercial solution that acts as an SQL[11] proxy between an analyst and a protected database [226, 245], has recently been proposed by researchers affiliated with Aircloak and the Max Planck Institute for Software Systems as a practical alternative to differential privacy, based on the [EU] Article 29 Working Party (Art. 29 WP)'s definition of anonymous data. It defines a dataset as anonymous if the anonymization mechanism used protects against *singling out* (identify one user), *linkability* (match users across datasets), and *inference* (learn participants' records) attacks [45, 226]. Diffix relies on a novel noise addition framework called "sticky noise", which aims to give analysts a rich query syntax, minimal noise addition, and an infinite number of queries, all while satisfying the WP29 definition of anonymous.

The authors claim that data produced by Diffix (*i*) falls outside of the scope of the new European GDPR regulation; (*ii*) has been determined by the French National Commission on Informatics and Liberty (CNIL) to offer "*GDPR-level anonymization*" for all cases; and (*iii*) has been certified by TÜViT as fulfilling "*all requirements for data collection and anonymized reporting*" [246, 247]. Diffix is used in production and Aircloak reports working with partners such as Telefonica, DZ Bank and Cisco.

[11] *SQL (Structured Query Language) is a standard programming language designed for managing and interrogating structured relational databases, invented in 1974.*

EXPLOITING DIFFIX'S NOISE In this chapter, we present a new class of attacks called noise-exploitation attacks. The attacks work in three parts: (*i*) canceling out part of the sticky noise using multiple queries, (*ii*) exploiting the noise Diffix adds to one query in order to learn information about the query set associated to this query, and (*iii*) using logical equivalence between multiple queries to obtain independent noise samples for the same query.

We develop two noise-exploitation attacks that take advantage of the structure of Diffix's sticky noise to infer private (also called secret) attributes of individuals in the dataset, violating

the inference requirement from the Article 29 WP definition of anonymization [45].

Our first attack, the differential attack, uses samples obtained by the difference between two queries' outputs, to discriminate between two distributions, depending on the value of the private attribute. We show that, under specific conditions, this attack potentially allows an attacker to infer private information of unique users with up to 96.8 % accuracy knowing only 5 pieces of auxiliary information we call attributes.

Our second noise-exploitation attack, the cloning attack, uses dummy conditions that affect the output of queries conditionally to the value of the private attribute. This attack relies on weaker assumptions and automatically validates them with high accuracy, without the need for an oracle. It proceeds in two steps: (*i*) a *validation* step, searching for subsets of known attributes to use for the attack, that will satisfy the assumptions required for its success, and (*ii*) an *inference* step that uses the attributes found to predict users' private attribute's value. We perform the attack against four real-world datasets, and show that it can infer the private attributes of between 87.0 % and 97.0 % of all records across datasets. We then present an optimized cloning attack that targets 55.4% of the users and achieves the same accuracy using as little as 32 queries. This demonstrates that introducing rate limiting (reducing the number of allowed queries) would not protect against our attacks.

8.1 THE DIFFIX FRAMEWORK

Here we summarize the Diffix framework as described by Francis *et al.* [226] and introduce notations for our attack. Diffix acts as an SQL proxy between an analyst and a *protected database D* where each row is an individual record and each column one attribute (Fig. 8.1). The analyst can send SQL queries to Diffix, which will process the queries and then output a noisy answer.

We denote with $\mathcal{A}_D$ the set of attributes in the database D. For instance, $\mathcal{A}_D$ could contain 4 attributes $\mathcal{A}_D = \{gender, age, zip, HIV\}$ with HIV a binary attribute (0 or 1). A record x is a row of D with values for the attributes in $\mathcal{A}_D$. For example, with $\mathcal{A}_D$ as above, we could have

$x = (M, 27, 55416, 1)$. We assume, for simplicity, that there is one and only one record for every user in D.

While Diffix can process a large part of the SQL syntax, we here focus on simple count queries:

```
SELECT count(*)
FROM table
WHERE condition1 AND condition2 [AND ...]
```

where every condition is an expression of the form "*attribute* $\square$ *value*" with $\square$ being $=$, $\neq$, $\leq$, $<$, $\geq$, or $>$. Here, this query collects all records from *table*, selects only those that validates all given boolean conditions, and returns the total number of matching records.
For simplicity, we use a shorter notation for queries using "$\wedge$" for the logical AND:

$$Q \equiv \mathsf{count}(condition_1 \wedge condition_2 \wedge \ldots).$$

The following query would, for example, count the number of individuals in the database who are male, 37 years old, and live in the area with ZIP code 48828:

$$Q \equiv \mathsf{count}(gender = M \wedge age = 37 \wedge zip = 48828)$$

DIFFIX'S PRIVACY-PRESERVING MECHANISM Diffix protects privacy through *sticky noise* addition (static and dynamic noise) and bucket suppression (see Fig. 8.1). Let $Q \equiv \mathsf{count}(C_1 \wedge \ldots \wedge C_h)$, and denote by $Q(D)$ the true result of Q evaluated on D (without noise). Diffix's output for Q on D (without bucket suppression, see below) is:

$$\widetilde{Q}(D) = Q(D) + \sum_{i=1}^{h} \mathsf{static}[C_i] + \sum_{i=1}^{h} \mathsf{dynamic}_Q[C_i] \qquad (8.1)$$

with $\mathsf{static}[C_i]$ the *static noise* for condition C_i, $\mathsf{dynamic}_Q[C_i]$ the *dynamic noise* for condition C_i in Q.

STATIC NOISE Let C be a condition, for example $age = 34$. The *static noise* $\mathsf{static}[C]$ associated to C is a random number drawn

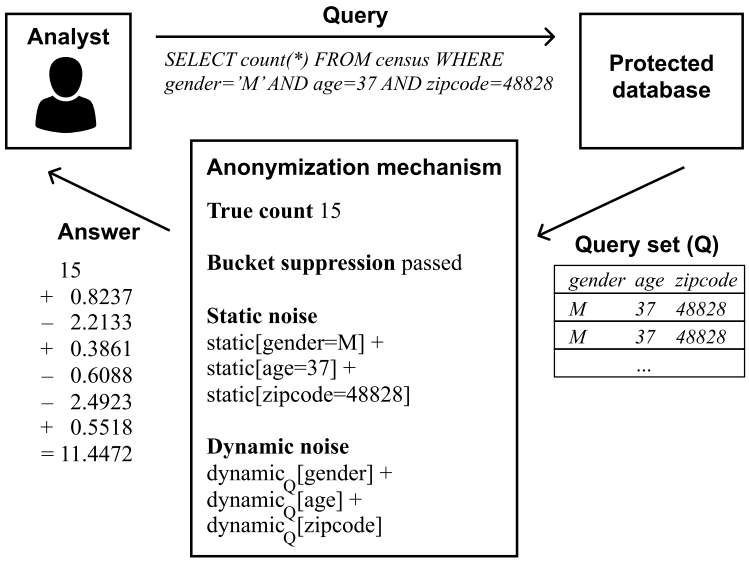


FIG. 8.1: Diffix privacy-preserving architecture.

from a normal distribution $\mathcal{N}(0,1)$. The value is generated by a pseudo-random number generator (PRNG), whose seed is a salted hash of the string literal C:

$$static\_seed_C = \mathsf{XOR}(\mathsf{hash}(C), salt)$$

This ensures that the static noise associated with C is always the same independently of the query where C appears. The noise is "sticky" thereby preventing an attacker to send the same query multiple times, average out the results, and obtain a precise estimate of the private value (*averaging attack*) [226].

DYNAMIC NOISE In the Diffix framework, every record in D is associated with a user ID, a unique string for that user. These pseudonyms are used to compute the dynamic noise. Let $Q \equiv \mathsf{count}(C_1 \wedge \ldots \wedge C_h)$ be a query and C any condition C_i. The dynamic noise depends not only on C, but also on the *query set* of Q in the dataset D, i.e. the set of users which satisfies *all* conditions $C_1, \ldots, C_h$. More precisely, if the query set for Q on D is $S = \{uid_1, uid_2, \ldots, uid_m\}$, the dynamic noise for C ($\mathsf{dynamic}_Q[C]$) is

generated from a normal distribution $\mathcal{N}(0,1)$ by the PRNG seeded with:

$$\begin{aligned} \mathit{dynamic\_seed} = \mathsf{XOR}(&\mathit{static\_seed}_C, \\ &\mathsf{hash}(\mathit{uid}_1),\ \ldots,\ \mathsf{hash}(\mathit{uid}_m)) \end{aligned}$$

Note that we don't include D in the notation $\mathsf{dynamic}_Q[C]$, as the dataset is usually fixed and clear from the context.

The output $\widetilde{Q}(D)$ is therefore the realization of a random variable distributed as a normal distribution $\mathcal{N}(\mu, \sigma^2)$, with mean $\mu = Q(D)$ and variance $\sigma^2 = 2h$.

EXAMPLE Consider again the query

$$Q \equiv \mathsf{count}(\mathit{gender} = M \wedge \mathit{age} = 37 \wedge \mathit{zip} = 48828).$$

Diffix's output for Q on the database D is

$$\begin{aligned} \widetilde{Q}(D) = \quad & Q(D) \\ & + \mathsf{static}[\mathit{gender} = M] + \mathsf{dynamic}_Q[\mathit{gender} = M] \\ & + \mathsf{static}[\mathit{age} = 37] + \mathsf{dynamic}_Q[\mathit{age} = 37] \\ & + \mathsf{static}[\mathit{zip} = 48828] + \mathsf{dynamic}_Q[\mathit{zip} = 48828] \end{aligned}$$

where $\widetilde{Q}(D)$ is a random value drawn from a normal distribution $\mathcal{N}(Q(D), 6)$.

Static and dynamic noise layers are both needed to prevent *intersection attacks* [226, 245], a class of attacks that combine multiple queries to infer private attributes of records.

BUCKET SUPPRESSION In addition to static and dynamic noise, Diffix implements another security measure called *bucket suppression*, similar to the classic query set size restriction. If the size of the query set is smaller than a certain threshold, the bucket suppression rejects the query. Previous research has shown that a fixed threshold constitutes a risk for privacy [248]. Diffix addresses this issue by using a *noisy* (and sticky to the query set) *threshold*. Specifically, suppose Diffix processes a query $Q \equiv \mathsf{count}(C_1 \wedge \ldots \wedge C_h)$. If $Q(D) \leq 2$, then the query gets suppressed. If $Q(D) > 1$, then Diffix computes a noisy threshold $T \sim \mathcal{N}(4, 1/2)$, using the seed:

$$\mathit{threshold\_seed} = \mathsf{XOR}(\mathit{salt}, \mathsf{hash}(\mathit{uid}_1), \ldots, \mathsf{hash}(\mathit{uid}_m))$$

If $Q(D) < T$, the query is suppressed; otherwise, the noisy output $\widetilde{Q}(D)$ is computed and sent to the analyst. In the original Diffix mechanism [245], the queries are said to be "silently suppressed" when censored by bucket suppression. This could mean that

1. a value of 0 is returned as result,
2. a random value is returned, or
3. Diffix displays an error message.

Here, we assume that a bucket-suppressed query will return a value of zero. This gives less information to a potential attacker than an error message, as a value of zero can be the result of either noise addition or bucket suppression. We consider that returning a random result affects utility too significantly to be applied in practice.

8.1.1 *Noise-exploitation attacks*

Our noise-exploitation attacks, which we call *differential* and *cloning*, are both based on three observations. First, since the noise is sticky, it is possible to *cancel out* part of it using multiple queries. Second, since the noise depends on the query set, the noise itself leaks information about the query set. Third, exploiting logical equivalence between some queries, it is possible to circumvent the "stickiness" of the noise by repeating (almost) the same query and consequently obtain independent noise samples. Our differential attack uses samples in a likelihood ratio test to discriminate between two probability distributions depending on the value of the private (also called secret) attribute. Our cloning attack relies on dummy conditions that conditionally strongly affect the output of the query depending on the value of the secret attribute.

8.1.2 *Differential noise-exploitation attack*

We first define further notations: with $A \subseteq \mathcal{A}_D$ a set of attributes, $x^{(A)}$ is the *restriction* of the record x to A, i.e. the vector one obtains after removing from x every entry for attributes that are not in A. For example, if $\mathcal{A}_D = \{gender, age, zip, HIV\}$, $x = (M, 27, 55416, 1)$ and

$A = \{gender, age, HIV\}$, then $x^{(A)} = (M, 27, 1)$. If A contains a single attribute a, we simply write $x^{(a)}$. So, for example, $x^{(gender)} = M$.

For this attack, we make the following assumptions:

H 1 The attacker wants to find out some information about Bob, the victim. The attacker knows that Bob's record is in the dataset. We denote Bob's record by x. The attacker has access to the protected dataset only through Diffix.

H 2 The attacker knows all of Bob's attributes for some set of attributes A. Our *background knowledge* (also called auxiliary information in the literature) is the restricted record $x^{(A)}$. This is a standard assumption in the literature on re-identification attacks [125, 249].

H 3 The attacker wants to infer a secret attribute s about Bob, the victim. That is, she wants to infer the value of $x^{(s)}$, where $s \notin A$. For simplicity of notation, s is a binary attribute, i.e. $x^{(s)} \in \{0, 1\}$. This means that the attack can be seen as a classifier, with the output of the attack being negative if the algorithm returns $x^{(s)} = 0$ and positive if it returns $x^{(s)} = 1$.
While we here focus on the binary case, our results fairly easily extend to non-binary cases.

H 4 There exists an oracle Unique that takes as input any restricted record $z^{(R)}$ and outputs whether $z^{(R)}$ is *unique*. Therefore $\mathsf{Unique}(z^{(R)})$ is True if and only if there is no other record y in D such that $z^{(R)} = y^{(R)}$.

For the attack to succeed, we first need to ensure that the background knowledge $x^{(A)}$ uniquely identifies Bob. This is given to us by the oracle $\mathsf{Unique}(x^{(A)})$. In this attack, we only attempt to infer secret attributes of records that are unique in the dataset. The cloning attack, presented later, extends this by requiring a weaker notion of uniqueness, which we call value-uniqueness.

While the existence of such an oracle is a strong assumption, it is weaker than Diffix's claims to protect against an "*analyst [that] has full access to the inference dataset*" [226], where the *inference dataset* is the original dataset with only the secret attribute $x^{(s)}$ removed. If the attacker has access to every record, she can easily verify that no other record shares the same restricted record $x^{(A)}$.

Firstly, the attack needs to bypass bucket suppression. For example, an attacker could ask how many records have both the background knowledge $x^{(A)}$ and the private attribute $s = 0$:

$$Q \equiv \mathsf{count}(a_1 = x_1 \wedge \ldots \wedge a_k = x_k \wedge s = 0). \tag{8.2}$$

with $A = \{a_1, \ldots, a_k\}$ and $x^{(A)} = (x_1, \ldots, x_k)$.

While an accurate answer to Q would immediately disclose the value of $x^{(s)}$, since $Q(D)$ can be either 0 (if $x^{(s)} = 1$) or 1 (if $x^{(s)} = 0$), this query will always be blocked by bucket suppression since the query set is either empty or $\{\text{Bob}\}$.

Intersection attacks have been proposed in the literature to circumvent similar kinds of restrictions [228]. Picking an attribute, *e.g.*, a_1, we can define the following queries:

$$Q_1 \equiv \mathsf{count}(a_2 = x_2 \wedge \ldots \wedge a_k = x_k \wedge s = 0) \tag{8.3}$$

$$Q_1' \equiv \mathsf{count}(a_1 \neq x_1 \wedge a_2 = x_2 \wedge \ldots \wedge \ldots \wedge a_k = x_k \wedge s = 0) \tag{8.4}$$

As the record x is unique, by assumption, it is the only record that can differ between Q_1 and Q_1'. This allows us to directly compute $Q(D)$:

$$Q(D) = Q_1(D) - Q_1'(D) = \begin{cases} 0 & \text{if } x^{(s)} = 1 \\ 1 & \text{if } x^{(s)} = 0 \end{cases} \tag{8.5}$$

[12] *For the intersection attack to be successful, both Q_1 and Q_1' need to be large enough to not trigger the bucket suppression. We discuss this assumption later on.*

To prevent this vulnerability[12], Diffix adds static and dynamic noises:

$$\begin{aligned} \widetilde{Q}_1(D) = \quad & Q_1(D) \\ & + \sum_{i=2}^{k} \mathsf{static}[a_i = x_i] + \mathsf{static}[s = 0] \\ & + \sum_{i=2}^{k} \mathsf{dynamic}_{Q_1}[a_i = x_i] + \mathsf{dynamic}_{Q_1}[s = 0] \end{aligned}$$

and

$$\begin{aligned}\widetilde{Q}'_1(D) = \quad & Q'_1(D) \\ & + \mathsf{static}[a_1 \neq x_1] + \mathsf{dynamic}_{Q'_1}[a_1 \neq x_1] \\ & + \sum_{i=2}^{k} \mathsf{static}[a_i = x_i] + \mathsf{static}[s = 0] \\ & + \sum_{i=2}^{k} \mathsf{dynamic}_{Q'_1}[a_i = x_i] + \mathsf{dynamic}_{Q'_1}[s = 0].\end{aligned}$$

The first part of our attack relies on noticing that $k-1$ static noise layers cancel out. Let

$$q_1 = \widetilde{Q}_1(D) - \widetilde{Q}'_1(D).$$

Then:

$$\begin{aligned}q_1 = \quad & Q_1(D) - Q'_1(D) \\ & - \mathsf{static}[a_1 \neq x_1] - \mathsf{dynamic}_{Q'_1}[a_1 \neq x_1] && (\textit{fixed}) \\ & + \sum_{i=2}^{k} \mathsf{dynamic}_{Q_1}[a_i = x_i] + \mathsf{dynamic}_{Q_1}[s = 0] && (\textit{dynamic}_{Q_1}) \\ & - \sum_{i=2}^{k} \mathsf{dynamic}_{Q'_1}[a_i = x_i] - \mathsf{dynamic}_{Q'_1}[s = 0] && (\textit{dynamic}_{Q'_1})\end{aligned}$$

leaving us with $2k+2$ noise layers: $\textit{fixed} \sim \mathcal{N}(0,2)$, $\textit{dynamic}_{Q_1} \sim \mathcal{N}(0,k)$, and $\textit{dynamic}_{Q'_1} \sim \mathcal{N}(0,k)$.

The second part of the attack relies on the fact that both $\mathsf{dynamic}_{Q_1}[a_i = x_i]$ and $\mathsf{dynamic}_{Q'_1}[a_i = x_i]$ (resp. $\mathsf{dynamic}_{Q_1}[s = 0]$ and $\mathsf{dynamic}_{Q'_1}[s = 0]$) relate to the same condition "$a_i = x_i$" (resp. "$s = 0$"). This means that the noise values for Q_1 and Q'_1 will cancel out if Q_1 and Q'_1 have the same query set. Therefore either $Q_1(D) - Q'_1(D) = 0$, and the $2k$ dynamic noise layers cancel out, or $Q_1(D) - Q'_1(D) = 1$, and no dynamic noise layer is canceled out:

$$q_1 \sim \begin{cases}\mathcal{N}(0,2) & \text{if } x^{(s)} = 1 \\ \mathcal{N}(1,2k+2) & \text{if } x^{(s)} = 0\end{cases} \tag{8.6}$$

Using this result, an attacker can run a likelihood ratio test (see Appendix 8.5.1) to estimate whether q_1 is distributed as $\mathcal{N}(0,2)$ or $\mathcal{N}(1,2k+2)$ and predict the value of $x^{(s)}$. The larger k is, the easier it

becomes to discriminate between the two distributions. This alone already allows the attacker to infer Bob's secret, $x^{(s)}$, with better than random accuracy.

The third part of the attack allows us to strongly improve the accuracy of our inference. While the stickiness of the noise prevents us from running the same query again to collect more sample, we circumvent it by using different pairs of queries for which equation (8.6) is still true. Specifically, instead of removing (resp. negating) the condition $a_1 = x_1$, we remove (resp. negate) other conditions $a_j = x_j$ for $j \leq k$, obtaining queries Q_j (resp. Q'_j) for $j \leq k$. In our notation:

$$Q_j \equiv \text{count}\left(\bigwedge_{i=1\neq j}^{k} a_i = x_i \wedge s = 0\right) \tag{8.7}$$

$$Q'_j \equiv \text{count}\left(\bigwedge_{i=1\neq j}^{k} a_i = x_i \wedge a_j \neq x_j \wedge s = 0\right) \tag{8.8}$$

Running all queries $\{(Q_j, Q'_j)\}_{j\leq k}$, the attacker collects a vector of realizations $\{q_j\}_{j\leq k}$ where $q_j = \widetilde{Q}_j(D) - \widetilde{Q}'_j(D)$. All q_j values are computed from different queries, which generate different noises. Hence, the noises all have different values (with probability 1). Nevertheless, the equation (8.6) is still true for each q_j, so we can combine them as k different samples from the same distribution, and estimate the value of $Q(D)$.

Finally, replacing the "$s = 0$" condition with "$s = 1$" in every pair (Q_j, Q'_j) defines k new pairs of queries $\{(R_j, R'_j)\}_{j\leq k}$. This allows us to obtain k more samples $\{r_j\}_{j\leq k}$ (with different noises and inverted results in equation (8.6)). This gives us a total of $2k$ samples before bucket suppression (see Appendix 8.5.1), generated by issuing $4k$ queries.

On a technical note, observe that in principle we cannot be certain that the q_j's (resp. r_j's) are all independent samples, because two different queries Q_j and Q_l (resp. R_j and R_l) with $j \neq l$ might have the same query set. For example, the queries count $(sex = F \wedge diagnosis = endometriosis)$ and count $(diagnosis = endometriosis)$ have the same query set despite having different conditions. In that case, the dynamic noise layers associated to the same conditions would have the same values, and hence the total dynamic noises of the two queries would be

heavily correlated. While this affects the accuracy of the likelihood ratio test, the impact is negligible in practice. Hereafter, we always assume that the samples are independent. Note that this potential dependence can add uncertainty to the results of an attack but also be used, *e.g.*, for a background knowledge attack.

EXAMPLE We present an example of how the differential attack would work in practice. We consider a dataset D containing the attributes *age*, *zip*, and *HIV*, and assume that the attacker knows Bob's restricted record $x \setminus x^{(HIV)}$ for which $x^{(age)} = 37$ and $x^{(zip)} = 48828$. As before, we assume that Bob is the only user with both *age* = 37 and *zip* = 48828. Consider the query:

$$Q \equiv \mathsf{count}(age = 37 \wedge zip = 48828 \wedge HIV = 0).$$

which has true value 0 if Bob has *HIV*, and 1 if Bob does not have HIV. To circumvent the bucket suppression, the attacker defines two queries for the intersection attack. The first pair of queries removes (resp. negates) the *age* condition:

$$\begin{aligned} Q_1 &\equiv \mathsf{count}(zip = 48828 \wedge HIV = 0) \\ Q_1' &\equiv \mathsf{count}(age \neq 37 \wedge zip = 48828 \wedge HIV = 0). \end{aligned}$$

Based on the previous findings, we know that $q_1 = \widetilde{Q}_1(D) - \widetilde{Q}_1'(D)$ satisfies:

$$q_1 \sim \begin{cases} \mathcal{N}(0,2) & \text{if Bob has HIV} \\ \mathcal{N}(1,4) & \text{if Bob does not have HIV.} \end{cases} \tag{8.9}$$

The realization q_1 is a first sample for our likelihood ratio test. Repeating the same procedure with the *zip* attribute gives us:

$$\begin{aligned} Q_2 &\equiv \mathsf{count}(age = 37 \wedge HIV = 0) \\ Q_2' &\equiv \mathsf{count}(age = 37 \wedge zip \neq 48828 \wedge HIV = 0), \end{aligned}$$

and $q_2 = \widetilde{Q}_2(D) - \widetilde{Q}_2'(D)$ is again distributed as $\mathcal{N}(0,2)$ if Bob has HIV, and as $\mathcal{N}(1,4)$ otherwise. This is the second sample. Repeating the procedure defining queries R_1 and R_1', as well as R_2 and R_2', and replacing $HIV = 0$ with $HIV = 1$ (and inverting the results), the attacker obtains two more samples.

FULL DIFFERENTIAL ATTACK For larger values of k, the queries used by the differential attack contain many conditions, and hence potentially select a low number of records. Depending on the dataset, this might result in a large fraction of queries being bucket suppressed, leaving the attacker with few or no samples for the likelihood ratio test. To counteract this effect, we integrate the attack with a subset exploration step to obtain a *full differential attack*. Assume that the attacker knows a set A^* of the correct attributes for the victim with $|A^*| = k^*$, i.e. the background knowledge is $x^{(A^*)}$. The full attack proceeds as follows. The algorithm selects random subsets of A^* until it finds a subset $A \subseteq A^*$ such that $\mathsf{Unique}(x^{(A)})$ returns True. It then performs the differential attack using $x^{(A)}$ as background knowledge. For the likelihood ratio test, the attack considers only query pairs $(\widetilde{Q}_j(D), \widetilde{Q}'_j(D))$ or $(\widetilde{R}_j(D), \widetilde{R}'_j(D))$ which have outputs larger than zero in both entries. If no such pair exists, the algorithm samples a new subset A and iterates the procedure. If no feasible subset is found, the algorithm outputs NonAttackable. The subsets of A^* are sampled by decreasing size, as bucket suppression is less likely for lower values of k (but on the other hand the likelihood ratio test is less accurate).

The procedure FullDifferentialAttack presents in detail the algorithm outlined above.

8.1.3 *Cloning noise-exploitation attack*

In this section, we present an extension of the differential noise-exploitation attack, which we call *cloning attack*. This attack adds dummy conditions, that don't affect the query set, to queries, in such a way that several queries with different dummy conditions will have either identical or very different results conditional to the secret attribute's value.

We first introduce some new notations and definitions. Denoting by x the victim's entire record, the attacker's background information is now $x^{(A)} = (x^{(A')}, x^{(u)})$ with $A = A' \cup \{u\}$ and $|A| = k$. We use the shorthand (A', u) for $A' \cup \{u\}$. We also define a restricted record $z^{(A)}$ to be *value-unique* for the attribute s in D if $y^{(A)} = z^{(A)}$ implies $y^{(s)} = z^{(s)}$. That is, every record that shares the same attributes of the restricted record $z^{(A)}$ holds the same value for the secret attribute s. To simplify the notation, we write $A' = x^{(A')}$ to indicate

Procedure DifferentialAttack($A, x^{(A)}, s$)

Input: known attributes (names A, values $x^{(A)}$), secret s
Output: True if $x^{(s)} = 1$, False if $x^{(s)} = 0$, NoSamples if $x^{(A)}$ does not yield any positive sample

$k \leftarrow |A|,\ \mathcal{Q} \leftarrow \varnothing,\ \mathcal{R} \leftarrow \varnothing$
for $j \leftarrow 1$ **to** k **do**
  $I \leftarrow \{i \in [1,k] \mid i \neq j\}$
  $\widetilde{Q} \leftarrow \mathsf{count}\left(\bigwedge_{i\in I} a_i = x_i \wedge s = 0\right)$
  $\widetilde{Q}' \leftarrow \mathsf{count}\left(\bigwedge_{i\in I} a_i = x_i \wedge a_j \neq x_j \wedge s = 0\right)$
  if $\widetilde{Q} > 0$ and $\widetilde{Q}' > 0$ **then**
    $q_j \leftarrow \widetilde{Q} - \widetilde{Q}'$
    $\mathcal{Q} \leftarrow \mathcal{Q} \cup \{q_j\}$
  end if
end for
for $j \leftarrow 1$ **to** k **do**
  $I \leftarrow \{i \in [1,k] \mid i \neq j\}$
  $\widetilde{R} \leftarrow \mathsf{count}\left(\bigwedge_{i\in I} a_i = x_i \wedge s = 1\right)$
  $\widetilde{R}' \leftarrow \mathsf{count}\left(\bigwedge_{i\in I} a_i = x_i \wedge a_j \neq x_j \wedge s = 1\right)$
  if $\widetilde{R} > 0$ and $\widetilde{R}' > 0$ **then**
    $r_j \leftarrow \widetilde{R} - \widetilde{R}'$
    $\mathcal{R} \leftarrow \mathcal{R} \cup \{r_j\}$
  end if
end for
if $\mathcal{Q} = \varnothing$ and $\mathcal{R} = \varnothing$ **then**
  return NoSamples
end if
$f \leftarrow$ PDF of $\mathcal{N}(0,2), \quad g \leftarrow$ PDF of $\mathcal{N}(1, 2k+2)$
$L \leftarrow \prod_{q\in\mathcal{Q}} \frac{f(q)}{g(q)} \prod_{r\in\mathcal{R}} \frac{g(r)}{f(r)}$
return $L \geq 1$

Procedure FullDifferentialAttack($A^*, x^{(A^*)}, s$)

Input: known attributes (names A^* and values $x^{(A^*)}$), secret s

Output: True if $x^{(s)} = 1$, False if $x^{(s)} = 0$, NonAttackable if $x^{(A^*)}$ is not attackable

```
for k ← |A*| to 1 do
    for iter ← 1 to 100 do
        A ← RandomSubsetOfSize(A*, k)
        if Unique(x^(A)) and DifferentialAttack(A, x^(A), s) ≠
          NoSamples then
            return DifferentialAttack(A, x^(A), s)
        end if
    end for
end for
return NonAttackable
```

the condition that all attributes in A' must match the ones in $x^{(A')}$, i.e. $\wedge_{a \in A'} a = x^{(a)}$

The attack addresses several limitations of the differential attack, making it much stronger in practice.

First, the cloning attack does not require an oracle to confirm that the background information uniquely identifies a user. Instead, it replaces the oracle with a heuristic to automatically validate the assumptions.

Second, the differential attack has to ignore any pair $(\widetilde{Q}_j(D), \widetilde{Q}'_j(D))$ where at least one of the entries is not positive, as it cannot tell whether the null output comes from noise addition or from bucket suppression. This significantly reduces the total number of samples used. Our cloning attack instead uses "dummy conditions" that do not impact the user set. As queries now only differ in the dummy conditions, the corresponding query sets will always be identical. This allows us to rule out bucket suppression in case at least one output is greater than zero.

Third, while the differential attack requires records to be unique, the cloning attack only requires records to be value-unique. This is a

weaker condition and makes the cloning attacks effective on a larger set of users.

Fourth, the cloning attack only requires that the set of users who share all attributes in $x^{(A')}$ is "large enough" to not be bucket suppressed. This is a weaker assumption than for the differential attack, where this needed to hold for a large number of subsets of A with one attribute removed. Furthermore, the cloning attack validates this automatically (and thus prevents bucket suppression) with high confidence.

While much stronger, the attack relies on the attacker being able to produce a set of distinct *dummy conditions* $\Delta = \{\Delta_j\}_{1 \le j \le |\Delta|}$, where each Δ_j is an SQL statement such that the set of users matching $A' = x^{(A')}$ is the same as the set of users matching $A' = x^{(A')} \wedge \Delta_j$. In Section 8.3, we discuss how dummy conditions are easy to obtain, slow to detect, and how automatically filtering them might introduce new vulnerabilities.

DESCRIPTION OF THE ATTACK For each dummy condition Δ_j, we define the two queries:

$$Q_j \equiv \text{count}\left(A' = x^{(A')} \wedge \Delta_j \wedge s = 0\right) \tag{8.10}$$

$$Q'_j \equiv \text{count}\left(A' = x^{(A')} \wedge \Delta_j \wedge u \neq x^{(u)} \wedge s = 0\right) \tag{8.11}$$

With $q_j = \widetilde{Q}_j(D) - \widetilde{Q}'_j(D)$, we have:

$$\begin{aligned} q_j = \quad & Q_j(D) - Q'_j(D) \\ & - \text{static}[u \neq x^{(u)}] \; - \text{dynamic}_{Q_j}[u \neq x^{(u)}] \\ & + \sum_{i \in A'} \text{dynamic}_{Q_j}[a^{(i)} = x^{(i)}] + \text{dynamic}_{Q_j}[s = 0] \\ & - \sum_{i \in A'} \text{dynamic}_{Q'_j}[a^{(i)} = x^{(i)}] - \text{dynamic}_{Q'_j}[s = 0] \\ & + \left(\text{dynamic}_{Q_j}[\Delta_j] - \text{dynamic}_{Q'_j}[\Delta_j]\right) \end{aligned}$$

By the same argument we presented for the differential attack, if $x^{(s)} = 1$ then $Q_j(D) = Q'_j(D)$ and most dynamic and static noises cancel out, giving:

$$q_j = -\,\text{static}[u \neq x^{(u)}] - \text{dynamic}_{Q_j}[u \neq x^{(u)}] \tag{8.12}$$

As this value does not depend on the dummy condition used, we have that $q_1 = q_2 = \cdots = q_{|\Delta|}$.

On the contrary, if $x^{(s)} = 0$, then the noise layers do not cancel out with probability 1. As the noise values given by $\mathsf{dynamic}_{Q_j}[\Delta_j]$ and $\mathsf{dynamic}_{Q'_j}[\Delta_j]$ depend on Δ_j, the probability that all (or any) q_j are equal is zero.

We can therefore complete the attack by inferring that $x^{(s)} = 1$ if $q_1 = \cdots = q_{|\Delta|}$, and $x^{(s)} = 0$ otherwise. Under the current assumptions, the attack always infers the correct value with 100% confidence (up to pseudo-random collisions in Diffix's noise addition mechanism).

ROBUSTNESS AGAINST ROUNDING In the previous section, we follow the Diffix papers [226, 245] and assume that $\widetilde{Q}(D)$ is returned directly without any rounding, admitting also negative values. We now consider the case where results are rounded to the nearest nonnegative integer, and propose a simple modification of our attack that accounts for this.

When the results of the queries Q_j and Q'_j are rounded, the corresponding q_j might not be identical if $x^{(s)} = 0$. However, the q_js will vary less if $x^{(s)} = 1$ than if $x^{(s)} = 0$. Hence, instead of checking if $q_1 = \ldots = q_{|\Delta|}$, we check if the q_j values are "similar" to one another. While for high values of k this is easy to detect, the total variance of the noise for low values of k is small, making it harder to distinguish between the two hypotheses (i.e. whether the q_j values are "similar" or not). To overcome this issue, we "amplify" the noise for each query: instead of adding a single dummy condition Δ_j to the queries for Q_j and Q'_j, we add the conjunction $\wedge_{l \neq j} \Delta_l$:

$$Q_j \equiv \mathsf{count}\left(A' = x^{(A')} \wedge \bigwedge_{l \neq j} \Delta_l \wedge s = 0\right) \tag{8.13}$$

$$Q'_j \equiv \mathsf{count}\left(A' = x^{(A')} \wedge \bigwedge_{l \neq j} \Delta_l \wedge u \neq x^{(u)} \wedge s = 0\right) \tag{8.14}$$

This increases the total variance of the noise in q_j in the $x^{(s)} = 0$ case, making it easy to distinguish between the two hypotheses: all the q_j values will be very similar if $x^{(s)} = 1$ and fluctuate heavily if $x^{(s)} = 0$. Measuring the sample variance S^2 of $\{q_j\}_{1 \leq j \leq |\Delta|}$, we infer that $x^{(s)} = 1$ if $S^2 \leq \sigma^*$, and $x^{(s)} = 0$ otherwise with a

cutoff threshold σ^* chosen by the attacker (see Appendix 8.5.2 for an empirical analysis of σ^*). The cloning attack is described in detail in the procedure CloningAttack.

Procedure CloningAttack($A', u, x^{(A',u)}, \Delta, s, v$)

Input: known attributes (names A', u and values $x^{(A',u)}$), dummy conditions Δ, secret s and target value v

Output: True if $x^{(s)} = v$, False if $x^{(s)} \neq v$

for $j \leftarrow 1$ **to** $|\Delta|$ **do**
  $\phi \leftarrow A' = x^{(A')} \wedge \bigwedge_{l \neq j} \Delta_l$
  $\widetilde{Q} \leftarrow \mathsf{count}\,(\phi \wedge s \neq v)$
  $\widetilde{Q}' \leftarrow \mathsf{count}\left(\phi \wedge u \neq x^{(u)} \wedge s \neq v\right)$
  $q_j \leftarrow \widetilde{Q} - \widetilde{Q}'$
end for
$\bar{r} \leftarrow \frac{1}{|\Delta|} \sum_{j=1}^{|\Delta|} q_j, \quad S^2 \leftarrow \frac{1}{|\Delta|-1} \sum_{j=1}^{|\Delta|} (q_j - \bar{r})^2$
return $S^2 \leq \sigma^*$

AUTOMATED ASSUMPTION VALIDATION The cloning attack relies on two assumptions on the attacker's background knowledge $x^{(A',u)}$:

1. The queries $\{Q_j\}_{1 \leq j \leq |\Delta|}$ and $\{Q'_j\}_{1 \leq j \leq |\Delta|}$ in equations (8.13) and (8.14) are not bucket suppressed.

2. The user is value-unique in the dataset according to (A', u) for the secret attribute s.

We here propose procedures for an attacker to determine whether (A', u) satisfies the two assumptions with high probability.

Validating the first assumption can be done easily by submitting queries $\{Q_j\}_{1 \leq j \leq |\Delta|}$ and $\{Q'_j\}_{1 \leq j \leq |\Delta|}$ to Diffix. Recall that the threshold for bucket suppression for a query depends only on the corresponding query set. All the queries in $\{Q_j\}_{1 \leq j \leq |\Delta|}$ have the same query set, and the same applies for $\{Q'_j\}_{1 \leq j \leq |\Delta|}$. Hence, if *any* query Q_j is bucket suppressed (i.e. has output zero), then *all* queries in $\{Q_j\}_{1 \leq j \leq |\Delta|}$ must have output zero, and similarly for $\{Q'_j\}_{1 \leq j \leq |\Delta|}$. Thus, if *any* query Q_j and *any* query Q'_j have output higher than zero, we are sure that no query was bucket suppressed, and hence

all q_j's are valid samples. The test is considered passed in this case, and failed otherwise. See the algorithm NoBucketSuppression for an implementation example.

Procedure NoBucketSuppression($A', u, x^{(A',u)}, \Delta, s, v$)

Input: known attributes (names A', u and values $x^{(A',u)}$), dummy conditions Δ, secret s and target value v

Output: True if (A', u) passes the tests and is deemed to satisfy assumption 1, False otherwise

$ok_Q \leftarrow 0$, $ok_{Q'} \leftarrow 0$
for $j \leftarrow 1$ **to** $|\Delta|$ **do**
  $\phi \leftarrow A' = x^{(A')} \wedge \bigwedge_{l \neq j} \Delta_l$
  $\widetilde{Q} \leftarrow \mathsf{count}\left(\phi \wedge s \neq v\right)$
  $\widetilde{Q}' \leftarrow \mathsf{count}\left(\phi \wedge u \neq x^{(u)} \wedge s \neq v\right)$
  if $\widetilde{Q} > 0$ **then**
    $ok_Q \leftarrow 1$
  end if
  if $\widetilde{Q}' > 0$ **then**
    $ok_{Q'} \leftarrow 1$
  end if
end for
return $ok_Q = 1$ & $ok_{Q'} = 1$

Validating the second assumption relies on a heuristic. We run the query:

$$\mathsf{count}\left(A' = x^{(A')} \wedge u = x^{(u)}\right)$$

and consider the assumption validated if the output is zero, and not otherwise. The idea is that if the output is larger than zero, then the query was not bucket suppressed, and many users are likely share the same attributes $x^{(A',u)}$, meaning that $x^{(A',u)}$ is unlikely to be value-unique. Experiments in Section 8.2 show that this heuristic works very well on real-world datasets.

The procedures CloningAttack and NoBucketSuppression both issue (the same) $2|\Delta|$ queries, while ValueUnique uses only one query. Validating the assumptions and performing the attack thus requires only $2|\Delta| + 1$ queries, for a given set of attributes (A', u). We empirically obtain accuracy above 93.3% with $|\Delta|$ as low as 10 (see Section 8.2 and Appendix 8.5.5).

Procedure ValueUnique(A', u, $x^{(A',u)}$)

Input: known attributes (names A', u and values $x^{(A',u)}$)
Output: True if (A', u) passes the tests and is deemed to satisfy assumption 2, False otherwise

1 $\widetilde{Q} \leftarrow \mathsf{count}(A' = x^{(A')} \wedge u = x^{(u)})$
2 **return** $\widetilde{Q} = 0$

FULL CLONING ATTACK Combining procedures CloningAttack, NoBucketSuppression and ValueUnique, we design a fully fledged procedure FullCloningAttack that performs the entire attack under the following assumptions:

H 1 The attacker knows that the victim's record x is in the dataset.

H 2 The attacker knows a set A^* of the correct attributes for the victim with $|A^*| = k^*$, i.e. the background knowledge is $x^{(A^*)}$.

H 3 The secret attribute $x^{(s)}$ is a binary attribute.

The full cloning attack includes a subset exploration step similar to the one used in the full differential attack. The algorithm selects random subsets A' of A^* (and an element u from $A^* \setminus A'$ at random) by decreasing size until it finds a subset that passes both tests, upon which it then performs the attack using $x^{(A',u)}$ as background knowledge. If no feasible subset is found, the algorithm outputs NonAttackable.

REDUCING THE NUMBER OF QUERIES While Diffix allows each analyst to send arbitrarily many queries, we study how many queries are required to perform the cloning attack in practice. In Appendix 8.5.5 we present a heuristic that reduces the median number of queries by a factor of 100. Using this heuristic, the attack targets 55.4% of the users in the dataset, achieving 91.7% accuracy with a maximum of 32 queries per user in our experiments.

8.2 EXPERIMENTS

In order to assess the effectiveness of our attacks, we implemented Diffix's mechanism for counting queries as described in the

Procedure FullCloningAttack($A^*, x^{(A^*)}, \Delta, s, v$)

Input: known attributes (names A^* and values $x^{(A^*)}$), dummy conditions Δ, secret s and target value v
Output: True if $x^{(s)} = v$, False if $x^{(s)} \neq v$

```
for k ← |A*| to 1 do
    for iter ← 1 to 100 do
        A' ← RandomSubsetOfSize(A*, k − 1)
        u ← RandomElement(A* ∖ A')
        if NoBucketSuppression(A', u, x^(A), Δ, s, v) and
          ValueUnique(A', u, x^(A*)) then
            return CloningAttack(A', u, x^(A), Δ, s, v)
        end if
    end for
end for
return NonAttackable
```

original paper [226]. The implementation outputs zero when queries are bucket-suppressed and results are rounded to the nearest nonnegative integer. We apply our attacks to four datasets and an additional synthetic dataset on which the assumptions of the differential attack are always validated.

8.2.1 *Description of the datasets*

In our experiments, we use the following datasets:

1. ADULT[1]: a U.S. Census dataset with 30,162 records and 11 attributes, incl. salary class as secret attribute [250].
2. CREDIT[2]: a credit card application dataset with 690 records and 16 attributes, incl. accepted credit as secret attribute [250].
3. CENSUS[3]: a U.S. Census dataset with 199,523 records and 42 attributes, incl. total personal income (digitized, null income as negative condition) as secret attribute [250].

1 https://archive.ics.uci.edu/ml/datasets/adult
2 https://archive.ics.uci.edu/ml/datasets/Credit+Approval
3 https://archive.ics.uci.edu/ml/datasets/Census-Income+(KDD)

4. CDR: a synthetic collection of phone metadata with 2,000,000 records generated using real-world data for human behaviour and the geography of the UK for the location of antennas. Every user is a record of 11,674,870 binary attributes (an attribute being whether a user was geographically present at a certain place and time, and placed a call or received a text message). As the vast majority of the attributes in a record are null, the distribution of values for a random attribute is heavily skewed towards zero. To obtain a balanced experiment, for 50% of runs we select as the secret attribute a pair $(location, time)$ where the user was present, and for the other 50% we select a pair where the user was absent.

8.2.2 *Differential noise-exploitation attack*

SYNTHETIC VALIDATION We first test the differential attack on a synthetic dataset where all users satisfy the uniqueness assumption.

In the $\mathsf{Complete}_k$ dataset, every user is unique according to k attributes (excluding the secret attribute), whereas $k-1$ attributes always identify a *larger* set of users. This ensures that (*i*) every user is vulnerable to the attack and (*ii*) bucket suppression is unlikely to be triggered by the attack queries. To create the dataset, we fix an integer B and generate every possible k–tuple whose values are in $\{1, \ldots, B\}$. We then append to each tuple a random value of either 0 or 1 for the secret attribute. $\mathsf{Complete}_k$ contains B^k records, one for each combination of k attributes. For our experiments, we set $B = 12$, to ensure that close to no bucket suppression occurs. For computational reasons, as the size of the dataset in memory grows as B^k, the maximum k we can use is limited to 6.

Fig. 8.2 compares the accuracy $\mathrm{acc}(k)$ of the attack, knowing k attributes, on $\mathsf{Complete}_k$ with the theoretical accuracy. The procedure we use here is DifferentialAttack, which does *not* include subset exploration and has no access to the oracle. Hence, this experiment simulates a realistic attacker. We report the empirical fraction of users whose secret attribute is correctly predicted, estimated by performing the differential attack on a sample of 1000 users. We also report the theoretical distribution of accuracy, (*i*) without rounding (closed-form expression, see Appendix 8.5.1) and

(*ii*) with rounding (numerical simulation, see Appendix 8.5.1). For the Complete$_k$ dataset, the accuracy reaches 92.6% for 5 points. Even knowing only $k = 2$ attributes, the accuracy is above 66% both theoretically and empirically. While rounding has close to no effect on the theoretical accuracy of the attack, comparing the Complete$_k$ with the Theoretical (rounding) curves shows that bucket suppression and potential correlations between the samples in empirical experiments noticeably decrease the accuracy.

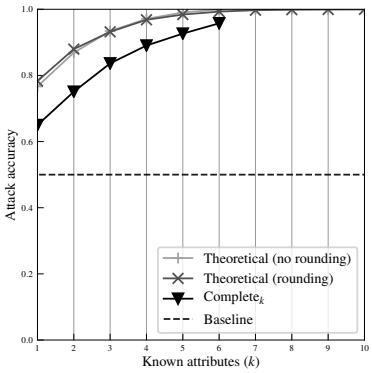


FIG. 8.2: Accuracy of the differential attack when the uniqueness assumption is always validated and with balanced truth values. The baseline accuracy is 0.5 and represents the expected success rate when randomly predicting the secret attribute using a uniform prior.

EMPIRICAL VALIDATION We now evaluate the accuracy of the differential attack on 1,000 users selected at random in each of the four datasets. Contrary to the synthetic experiment, bucket suppression is more prevalent on real-world datasets. Therefore, we run the FullDifferentialAttack algorithm, knowing k^* attributes A^* that are selected at random for each record. If the attack outputs NonAttackable, the secret attribute is predicted at random (uniformly).

Fig. 8.3 shows, for each dataset and knowing k^* attributes, the percentage of unique individuals and the percentage of individuals, in each dataset, for which the secret is correctly inferred. The latter divided by the former gives us the accuracy of our attack. Our attack realizes an accuracy of 68.4% for ADULT with $k^* = 10$, 64.0% for

CREDIT with $k^* = 15$, 68.8% for CENSUS with $k^* = 40$, and 68.8% for CDR with $k^* = 6$.

Observe that the fraction of correctly inferred attributes plateaus with larger k^* for the CREDIT, CENSUS and CDR datasets. The reason is that, on these datasets, most users are unique for larger values of k^*. As explained in section 8.1.1, this makes bucket suppression more prevalent and reduces the total number of samples for the likelihood ratio test, which is a limitation of the differential attack.

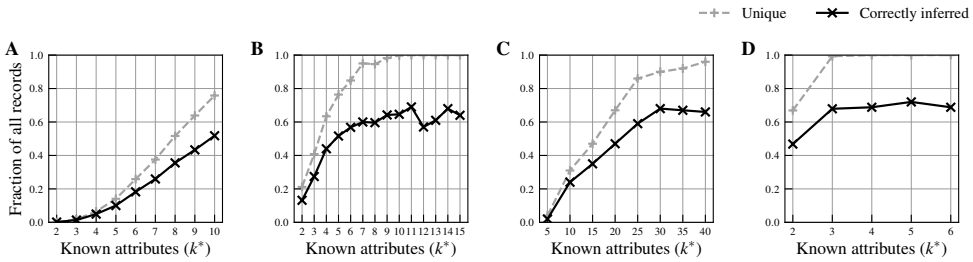


FIG. 8.3: Results of the differential attack for the (A) ADULT, (B) CREDIT, (C) CENSUS, and (D) CDR datasets.

8.2.3 *Cloning noise-exploitation attack*

We implement the attack as described by algorithm FullCloningAttack. As before, for each value of k^* we select 1,000 users at random, and for each user a random subset A^* of their attributes of size k^*. A^* represents the total number of attributes known to the attacker about the victim.

We set a threshold $\sigma^* = 0.7$ for the variance cutoff (see Appendix 8.5.2). We generate $|\Delta| = 10$ dummy conditions for $x_1 = a_1$ of the form $x_1 \neq b_j$ for $j \leq |\Delta|$, with b_j being some plausible values for x_1 different from a_1. We present the results when the attacker knows enough attributes (k^*) to identify every user, or up to all available attributes in the dataset.

Table 8.1 shows the proportion of records that are value-unique (third column) and fraction of the value-unique records that are predicted as attackable by procedures NoBucketSuppression and ValueUnique (fourth column). We then perform the cloning attack on all records that are predicted as attackable and report $\text{accuracy}_{\text{pa}}$,

the fraction of predicted attackable records whose secret attribute was successfully inferred (fifth column). For completeness, we also report $\text{accuracy}_{\text{all}}$, the fraction of all records in the datasets (including the ones deemed NonAttackable) whose secret attribute was successfully inferred (last column).

Table 8.1 shows that the cloning attack—including the assumption validation step—performs really well on all datasets considered, between 87.0 and 97.0 % of secret attributes and 97.0 % on the CREDIT dataset when knowing 15 attributes.

Dataset	Attributes (k^*)	Value-unique	Predicted ättackable	$\text{accuracy}_{\text{pa}}$	$\text{accuracy}_{\text{all}}$
ADULT	10	93.0%	96.8%	93.3%	87.0%
CREDIT	15	100.0%	100.0%	97.0%	97.0%
CENSUS	40	99.7%	94.6%	97.1%	91.6%
CDR	6	100.0%	100.0%	91.3%	91.3%

TABLE 8.1: Empirical results of the cloning attack on four real-world datasets.

Fig. 8.4 shows, knowing k^* attributes, the fraction of all records that are value-unique, predicted attackable, and correctly inferred. The curves for value-unique users and for predicted attackable users are always very close, suggesting that the assumption validation step is effective. Out of all records predicted as attackable, most of them are correctly inferred, demonstrating that the attack works on targeted records across all k. For the CREDIT dataset, with only six attributes, the attack reaches the inference step for 95% of the users in the dataset, and correctly infers the secret attribute for 93% of the total records.

8.3 DISCUSSION

8.3.1 *Attribute predictability*

Value-uniqueness plays an important role in the cloning attack. As Fig. 8.4 shows, it is a valid assumption for real-world datasets.

Value-uniqueness means that a group of people who share the same attributes also share the same secret attribute. If this group were to be large enough, the noise added by Diffix might not be enough to hide the secret attribute, which could then be revealed by using a simple count query. While this might be true for some datasets, it is not the case for any of the datasets we considered. For instance, the average size of the value-unique class (i.e. the set of value-unique users sharing the same restricted record) in the ADULT dataset is 1.44, with no class containing more than 4 users and similar numbers for the other datasets. This means that, most of the times, value-unique users are simply *unique*.

If secret attributes are predictable from the other attributes, a trained machine learning classifier could predict them with potentially high accuracy[13]. Despite our datasets coming from the machine learning literature, our attack does not rely at all on the predictability of secret attributes and performs equally well if no correlation at all exists between attributes and the secret attribute. In Appendix 8.5.3, we run our attack on a modified version of the ADULT dataset where sensitive attributes have been randomly sampled, thereby theoretically destroying any correlation. Our attacks perform as well on this modified dataset as on the original dataset.

[13] *Whether this would constitute a privacy attack is debated [251].*

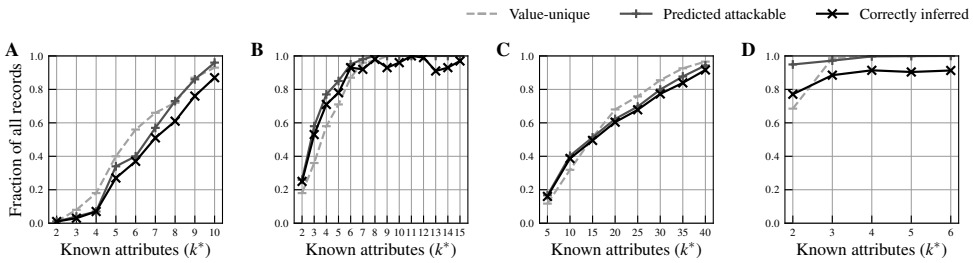


FIG. 8.4: Results of the cloning attack for the (A) ADULT, (B) CREDIT, (C) CENSUS, and (D) CDR datasets.

8.3.2 *Producing and detecting dummy conditions*

The cloning attack requires the attacker to provide a set of dummy conditions that affect the noise addition without affecting the query set. These conditions can be syntactic (*e.g.*, *age* ≥ 15 for the query *age* = 23), semantic (*e.g.*, *status* ≠ *retired* for *age* = 23), or pragmatic (*e.g.*, *age* ≠ 15 against a database containing only adult individuals).

When the language is rich enough, detecting redundant clauses is not a trivial task. At the same time, the richer the syntax is, the more utility an analyst gets out of the system. Diffix offers a fairly rich syntax including boolean expression, GROUP BY, JOIN, seven aggregation functions, set membership, fourteen string functions, and ten maths functions. In this context, automatically detecting dummy conditions would likely require iterating recursively through every condition and evaluating the query with and without them, a costly operation. Moreover, if dummy conditions are detected by evaluating them on the dataset, filtering them might not be safe. Removing semantic and pragmatic dummy conditions from a query would indeed reduce the total variance of the added noise and leak information about the dataset itself (*e.g.*, only adult individuals are present if the condition *age* ≠ 15 is always removed).

8.3.3 *Improving the attacks*

The cloning attack can be modified to run a double NoBucketSuppression test: once with $v = 0$ and once with $v = 1$. If both pass and the ValueUnique test is passed as well, the attack proceeds with the actual inference procedure CloningAttack for both $v = 0$ and $v = 1$. The attack then makes a guess only if both inferences return the same value, and continues with the subset exploration otherwise (deeming the user NonAttackable if no working set of attributes is found).

We found that this modified attack improves the $\text{accuracy}_{\text{pa}}$ figure on all datasets (*e.g.*, from 93.3% to 97.3% for ADULT, all results available in Appendix 8.5.4). However, double tests are more likely to fail, meaning that less users are predicted as attackable (from 96.8% to 87% in ADULT). Because of bucket suppression, this effect is particularly strong for datasets where the overall distribution of the secret attribute is very skewed, such as the CDR dataset where

the number of predicted attackable users goes from 100% to 8.5%. Depending on the aims of the attacker (precision versus coverage), she might prefer the original or the double version of the cloning attack.

OTHER IMPROVEMENTS To properly quantify the strength of our attacks, none of them use prior knowledge on the distribution of the secret attribute. In practice, an attacker might want to use this information, *e.g.*, obtaining it by querying Diffix. We discuss this in Appendix 8.5.4. We also discuss how to generate more samples for the differential attack and outline how to generalize both attacks to infer non-binary attributes.

8.4 RELATED WORK

We published the first version of our paper on ArXiv.org in April 2018, describing the differential attack. We updated it with a cloning attack in July 2018. Two months later, in October 2018, two other attacks on Diffix were disclosed. A membership attack by Pyrgelis *et al.* [152], based on a previous paper[150], and a reconstruction attack by Cohen and Nissim [252], based on previous work by Dinur *et al.* [229] and Dwork *et al.* [253]. These attacks are very different from ours and require a large number of queries in a typical setting, while our cloning attack can work with only 32 queries (see Appendix 8.5.5). These are, to the best of our knowledge, the only three attacks specifically targeting Diffix.

MEMBERSHIP ATTACK ON LOCATION DATA The attack by Pyrgelis *et al.* [152] is as follows: the attacker trains a machine learning algorithm on a *linkability* dataset (the attacker's background knowledge) to infer the presence of a user in a *protected* dataset (accessible only through queries on Diffix). Both datasets contain the full trajectories of users and half of the users are present in both datasets. The classifier is trained on the linkability dataset and queries on Diffix that count the number of people transiting in a certain area at a given time. The experimental results focus on the top 100 users with the highest number of reported locations in the linkability dataset. Out of 62 users present in both datasets, the classifier correctly infers the presence for 50 of them.

This attack presents three limitations. First, it is a membership attack and only allows an attacker to infer whether a person is in the protected dataset or not. Second, it assumes a strong adversary who has access to the full trajectory of a user exactly as it exists in the protected dataset, for a large number of users. Third, the attack requires about 32,000 queries to assess the presence of a user. Membership attacks are however very useful when combined with inference attacks like ours, allowing an adversary to effectively verify our assumption that the victim is in the dataset.

RECONSTRUCTION ATTACK The attack by Cohen *et al.* [252] focuses on reconstructing the dataset. In its simplest form, the attack assumes that the dataset contains n records, and each user $i \in [n]$ has a binary attribute s_i. The attack then selects random subsets of users and, for each subset $I \subseteq [n]$, queries Diffix for the result of $\sum_{i \in I} s_i$. This allows the attacker to produce a noisy linear system that can be solved using linear programming techniques to reconstruct the entire set of secret attributes $\{s_1, \ldots, s_n\}$ with perfect accuracy in polynomial time.

While this attack can successfully reconstruct the entire dataset, it presents two limitations compared to our attack.

First, it requires that the system allows queries of the type $\sum_{i \in I} s_i$, i.e. queries that select any analyst-defined set of users $I \subseteq [n]$, the "row-naming problem". The authors here exploit SQL functions supported by Diffix to define hash functions which they then use to select "random enough" sets of users. Following the disclosure, Aircloak restricted the available SQL functions to prevent the attack [254].

Second, to target a specific user, the attack would require a number of queries proportional to the number of records. Since the attacker does not know *which* name $i \in [n]$ corresponds to the victim's record, it is necessary to fully reconstruct at least a few columns entirely. The attacker would then perform a uniqueness attack on the reconstructed dataset to infer the secret attribute of the victim.

On the contrary, the number of queries used by our attacks is independent of the number of records in the dataset.

8.5 APPENDICES

8.5.1 *Likelihood ratio test*

Let X and Y be independent random variables. Suppose that we have two hypotheses about the distributions of X and:

$$\begin{aligned} H_0: &\quad X \sim \mathcal{N}(\mu_0, \sigma_0^2) \quad \text{and} \quad Y \sim \mathcal{N}(\mu_1, \sigma_1^2) \\ H_1: &\quad X \sim \mathcal{N}(\mu_1, \sigma_1^2) \quad \text{and} \quad Y \sim \mathcal{N}(\mu_0, \sigma_0^2) \end{aligned}$$

where $\mu_0, \mu_1, \sigma_0, \sigma_1$ are known and fixed values such that $\mu_0 < \mu_1$ and $\sigma_0^2 < \sigma_1^2$. $\mathcal{N}(\mu, \sigma^2)$ denotes the the normal distribution with mean μ and variance σ^2.

Suppose we have a vector of n realizations $\vec{x} = (x_1, \ldots, x_n)$ of X and a vector of n realizations $\vec{y} = (y_1, \ldots, y_n)$ of Y. We assume that all the $2n$ realizations are mutually independent. The standard frequentist way to accept the preferred hypothesis H_0 or refute it (in favor of H_1) would use a likelihood ratio test with a pre-defined confidence level from which to derive critical regions [255]. In our case we do not have a preferred hypothesis, and hence we define a slightly different test.

Let f and g denote the probability density functions of $\mathcal{N}(\mu_0, \sigma_0^2)$ and $\mathcal{N}(\mu_1, \sigma_1^2)$ respectively. We define the likelihood ratio function Λ as follows:

$$\Lambda(\vec{x}, \vec{y}) = \prod_{j=1}^{n} \frac{f(x_j)}{g(x_j)} \prod_{j=1}^{n} \frac{g(y_j)}{f(y_j)}.$$

We accept H_0 if $\Lambda(\vec{x}, \vec{y}) \geq 1$, and we accept H_1 if $\Lambda(\vec{x}, \vec{y}) < 1$.

THEORETICAL ACCURACY OF THE TEST The test will sometimes yield the wrong result. It is possible to determine what is the probability that this happens. Such probability depends on mean and variance of the two specific normal distributions.

Lemma 3. *Let* $p_{err_0} = \Pr[\Lambda(\vec{x}, \vec{y}) < 1 \mid H_0]$ *and* $p_{err_1} = \Pr[\Lambda(\vec{x}, \vec{y}) \geq 1 \mid H_1]$. *Then*

$$p_{err_0} = p_{err_1} = \Pr[\alpha_0 Z_0 - \alpha_1 Z_1 < 0]$$

where, for $i = 0, 1$, Z_i *is a noncentral chi-squared distribution with n degrees of freedom and noncentrality parameter*

$$\lambda_i = n \left(\frac{\mu_i}{\sigma_i} + \frac{\mu_0 \sigma_1^2 - \mu_1 \sigma_0^2}{\sigma_i (\sigma_0^2 - \sigma_1^2)} \right)^2$$

and

$$\alpha_i = \frac{\sigma_0^2 - \sigma_1^2}{2\sigma_{1-i}^2}.$$

To prove the lemma, one considers the log-likelihood ratio function $\log \Lambda(\vec{x}, \vec{y})$ and applies elementary algebra to the obtained expression to derive a linear combination of noncentral chi-squared distributions. We omit the details.

Since $p_{\text{err}_0} = p_{\text{err}_1}$, we refer to this quantity simply as p_{err}. The accuracy of the test is $\text{acc} = 1 - p_{\text{err}}$.

We now show how to apply the lemma to our differential noise-exploitation attack. For simplicity, we suppose that Diffix's outputs are not rounded to the nearest nonnegative integer and bucket suppression is never triggered for the queries in the attack, so that every pair of queries $(\widetilde{Q}_j, \widetilde{Q}'_j)$ and $(\widetilde{R}_j, \widetilde{R}'_j)$ yields a valid sample. Thus, for k known attributes, we have two vectors of samples $\vec{q} = (q_1, \ldots, q_k)$ and $\vec{r} = (r_1, \ldots, r_k)$ and for every $j \leq k$:

$$q_j \sim \begin{cases} \mathcal{N}(0, 2) & \text{if } x^{(s)} = 1 \\ \mathcal{N}(1, 2k+2) & \text{if } x^{(s)} = 0 \end{cases}$$

$$r_j \sim \begin{cases} \mathcal{N}(1, 2k+2) & \text{if } x^{(s)} = 1 \\ \mathcal{N}(0, 2) & \text{if } x^{(s)} = 0 \end{cases}$$

We assume that the $2k$ samples in $\vec{q}$ and $\vec{r}$ are mutually independent. As discussed in section 8.1.1, this is not always guaranteed to be true, but it has close to no effect on the actual accuracy of the test. Let

$$\begin{aligned} H_0 &: \quad x^{(s)} = 1 \\ H_1 &: \quad x^{(s)} = 0. \end{aligned}$$

Let f and g denote the probability density functions of $\mathcal{N}(0, 2)$ and $\mathcal{N}(1, 2k+2)$ respectively. Observe that H_0 holds if and only if every $q_j \sim f$ and every $r_j \sim g$. Similarly, H_1 holds if and only if every $q_j \sim g$ and every $r_j \sim f$. Then we can apply the test defined above. Define

$$\Lambda(\vec{q}, \vec{r}) = \prod_{j=1}^{k} \frac{f(q_j)}{g(q_j)} \prod_{j=1}^{k} \frac{g(r_j)}{f(r_j)}.$$

Our test concludes that $x^{(s)} = 1$ if $\Lambda(\vec{q}, \vec{r}) \geq 1$, and $x^{(s)} = 0$ if $\Lambda(\vec{q}, \vec{r}) < 1$.

To measure the theoretical accuracy of the attack for k known attributes, we can apply the lemma to $\Lambda(\vec{q}, \vec{r})$ with $\mu_0 = 0, \sigma_0^2 = 2, \mu_1 = 1, \sigma_1^2 = 2k + 2$ and $n = k$, and finally find $\mathrm{acc}(k) = 1 - p_{\mathrm{err}}(k)$.

Fig. 8.2 shows the values of $\mathrm{acc}(k)$ for increasing values of k. Computing the value of p_{err} requires an approximation of the cumulative distribution function of a linear combination of noncentral chi-squared distributions, for which an exact closed-form expression is not known [256]. We compute these values using the R package sadists[4] version 0.2.3.

NUMERICAL SIMULATION WITH ROUNDING If we suppose that Diffix's outputs are rounded to the nearest nonnegative integer, no simple expression can be determined for the error rate. To estimate the accuracy in this case, we numerically simulate the values of $\widetilde{Q}_j(D)$ and $\widetilde{Q}'_j(D)$ that would result from querying Diffix (without bucket suppression), for different values of the secret attribute $x^{(s)}$. We then obtain each sample as the difference of the rounded results:

$$q_j = \mathsf{round}(\widetilde{Q}_j(D)) - \mathsf{round}(\widetilde{Q}'_j(D)).$$

Finally, we perform the likelihood ratio test as for the continuous case (considering also null outputs) and check whether the result is correct. We use balanced truth values for $x^{(s)}$, and perform 1000 experiments (on different queries) for each value of $x^{(s)}$. The results are shown in Fig. 8.2.

8.5.2 *Cutoff threshold for the cloning attack*

The cloning attack tests whether the empirical variance of $|\Delta|$ samples is higher than a cutoff threshold σ^*. In order to choose this threshold σ^*, we simulate the noises and samples that would result from Diffix queries. We use the expressions resulting from equations (8.13) and (8.14), introducing the rounding operator $\mathsf{round}(\cdot)$ that rounds to the nearest nonnegative integer.

4 sadists, https://github.com/shabbychef/sadists.

For $j = 1, \ldots, |\Delta|$, we have:

$$\begin{aligned}\widetilde{Q}_j(D) = \mathsf{round}\Big(&Q_j(D)\\ &+ \sum_{i=1,i\neq j} \mathsf{static}[\Delta_i] + \sum_{i=1,i\neq j} \mathsf{dynamic}[\Delta_i]\\ &+ \mathsf{static}[A' = x^{(A')} \wedge s = 0]\\ &+ \mathsf{dynamic}[A' = x^{(A')} \wedge s = 0]\Big)\end{aligned}$$

and, with $q = Q_j(D)$ if $x^{(s)} = 1$ or $q = Q'_j(D)$ otherwise:

$$\begin{aligned}\widetilde{Q}'_j(D) = \mathsf{round}\Big(&q\\ &+ \sum_{i=1,i\neq j} \mathsf{static}[\Delta_i] + \sum_{i=1,i\neq j} \mathsf{dynamic}[\Delta_i]\\ &+ \mathsf{static}[A' = x^{(A')} \wedge s = 0]\\ &+ \mathsf{dynamic}[A' = x^{(A')} \wedge s = 0]\\ &+ \mathsf{static}[u \neq x^{(u)}] + \mathsf{dynamic}[u \neq x^{(u)}]\Big)\end{aligned}$$

We perform a balanced experiment, simulating 1000 set of answers for $x^{(s)} = 1$ and 1000 set of answers for $x^{(s)} = 0$. For each sample, we generate the different static and dynamic noises once by sampling each from $\mathcal{N}(0, 1)$, and use these values to compute the pairs $(\widetilde{Q}_j, \widetilde{Q}'_j)$. Without loss of generality, since we only measure the variance, we fix $Q_j(D) = Q'_j(D) + 1$. We then compute the sample variance of $\left(\widetilde{Q}_1 - \widetilde{Q}'_1, \ldots, \widetilde{Q}_{|\Delta|} - \widetilde{Q}'_{|\Delta|}\right)$.

Fig. 8.5 shows the accuracy at correctly predicting whether $x^{(s)} = 1$ or $x^{(s)} = 0$ from the variance of the $|\Delta|$ pairs. We find that, for $|\Delta| = 10$ (10 dummy queries) and a query set of size 10, selecting $\sigma^* = 0.7$ leads to a 98.2% true positive rate for a 98.4% true negative rate despite rounding. This choice is arbitrary, and an attacker could choose a different value for different true and false positive rates. We however argue that the attack is robust to the choice of σ^*: while 0.7 seems to overall maximize accuracy, values of the threshold between 0.4 and 1 have comparable accuracy.

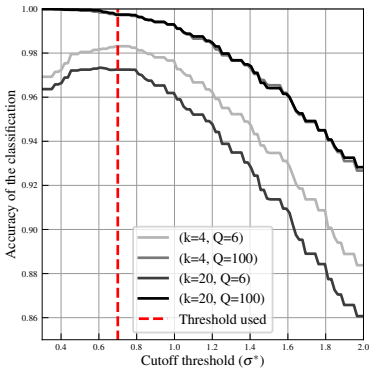


FIG. 8.5: Influence of the variance cutoff σ^* on the accuracy of the classification for the cloning attack with $|\Delta| = 10$. The samples are obtained by simulating Diffix results (Q_j, Q'_j) from their distribution, using different values of the number of attributes k and the true user set size Q.

8.5.3 *Size of value-unique classes and cloning attack on randomized dataset*

SIZE OF VALUE-UNIQUE CLASSES Fig. 8.6 shows the distribution of the size of value-unique classes of users predicted attackable by the cloning attack (using all attributes). In most cases, the restricted record used by the cloning attack uniquely identifies the victim and the victim's value-unique class never contains more than 5 users.

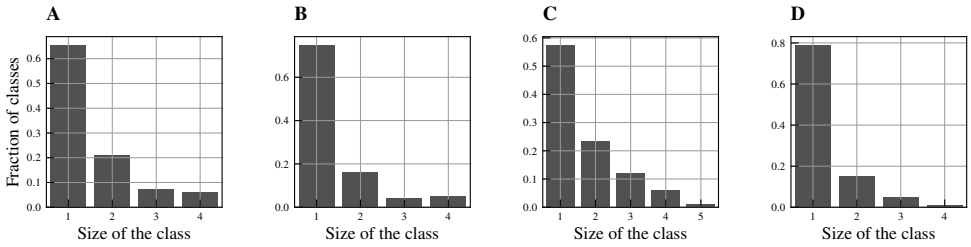


FIG. 8.6: Size of value-unique classes for users predicted attackable by the cloning attack (using all attributes) on the (A) ADULT, (B) CREDIT, (C) CENSUS, and (D) CDR datasets.

RANDOMIZED DATASET Our attacks do not rely at all on the predictability of secret attributes and perform equally well if

no correlation exists between attributes and the secret attribute. We create a modified version of the ADULT dataset where the secret attributes have been randomly sampled (0 or 1 with equal probability), which we call ADULT-randomized. Fig. 8.7 shows that our cloning attack achieves roughly the same accuracy on both datasets, and similarly for the differential attack (Fig. 8.8).

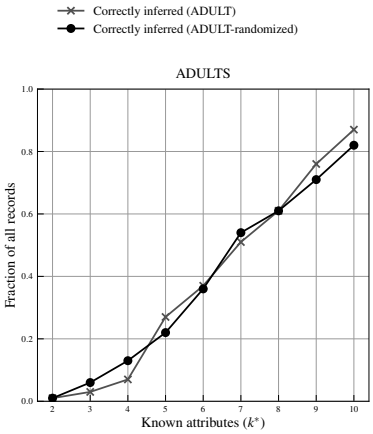


FIG. 8.7: Results of the cloning attack for the ADULT and the ADULT-randomized datasets.

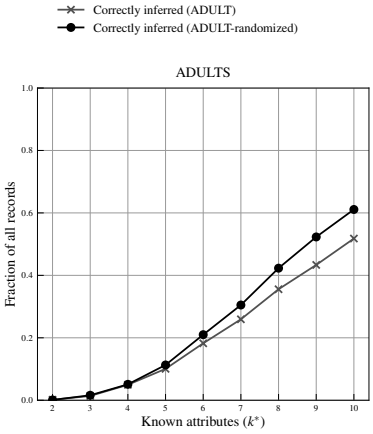


FIG. 8.8: Results of the differential attack for the ADULT and the ADULT-randomized datasets.

Dataset	Attributes (k^*)	Value-unique	Predicted attackable	$\text{accuracy}_{\text{pa}}$	$\text{accuracy}_{\text{all}}$
ADULT	10	87.0%	86.2%	97.3%	74.0%
CREDIT	11	100.0%	100.0%	99.0%	99.0%
CENSUS	45	100.0%	87.0%	98.9%	86.0%
CDR	6	100.0%	8.5%	94.1%	8.0%

TABLE 8.2: Empirical results of the double cloning attack, using all attributes.

8.5.4 *Improving the accuracy and generalizing to non-binary attributes*

DOUBLE CLONING ATTACK The cloning attack can be modified using a double test, as explained in section 8.3.3. We implement this attack and perform it on all four datasets. The results are reported in Table 8.2. Comparing these results with the ones for the single cloning attack (Table 8.1), we observe that $\text{accuracy}_{\text{pa}}$ is slightly higher on all datasets, but less users are predicted attackable on all but the CREDIT dataset. The issue is particularly evident on the CDR dataset, where the figure drops from 100% to 8.5%. The reason is that the distribution of the secret attribute in the CDR dataset is extremely skewed towards 0. This means that the part of the attack using $v = 1$ is likely to fail, due to many queries being bucket suppressed.

PRIOR KNOWLEDGE The differential attack and the cloning attack do not assume any prior knowledge about the distribution of the attribute s in the population (*i.e.* in the dataset). In some cases, one outcome of s could be much more likely than the other. For example, if the secret attribute is *HIV* and the dataset is randomly sampled from the population, it is more likely to observe $HIV = 0$ rather than $HIV = 1$. This prior knowledge about the population can be obtained empirically by querying the dataset itself, using $\text{count}(HIV = 0)$. This test can be easily integrated in the likelihood ratio test using Bayes' rule. For simplicity, we do not employ this improvement and we rather use an uninformative (uniform) prior.

OBTAINING MORE SAMPLES FOR THE DIFFERENTIAL ATTACK The differential attack relies on a likelihood ratio test with $2k$ samples, where k is the number of uniquely identifying attributes. There exist several ways to generate even more samples and thus increase the attack accuracy, useful when k is small or too many queries are bucket suppressed.

Consider again equations (8.7) and (8.8) for queries Q_j and Q'_j. The main property of Q_j and Q'_j is $Q_j(D) - Q'_j(D) = Q(D)$, where $Q(D)$ is the value we want to infer. There are many other pairs of queries that satisfy the same requirement and will be processed independently. For example, one could define $P_j(D)$ by adding the condition $a_j \leq x_j$ to Q_j and $P'_j(D)$ by replacing $a_j \neq x_j$ with $a_j < x_j$ in Q'_j. The record x is still the only user that can differ from one query to the other, and thus $P_j(D) - P'_j(D) = Q(D)$. By an argument similar to the one for Q_1 and Q'_1, we obtain:

$$\widetilde{P}_j(D) - \widetilde{P}'_j(D) \sim \begin{cases} \mathcal{N}(0,4) & \text{if } x^{(s)} = 1 \\ \mathcal{N}(1, 2k+2) & \text{if } x^{(s)} = 0. \end{cases}$$

Repeating this for $j = 1, \ldots, k$ yields $2k$ additional samples. Note that, as we add one condition, this creates two independent noises which will not cancel out. We can furthermore repeat the same procedure, inverting the inequalities to obtain $2k$ additional samples, leading to a total of $6k$ samples.

In general, $2k$ samples can be obtained from any pair of queries that define a partitioning attack for Q; different mathematical operators could also be employed to construct such partitions. Such samples will be independent most of the time. One could further exploit Diffix's rich SQL syntax by writing conditions with different syntax but identical meaning. For example, we could replace every condition "$a_i = x_i$" with "a_i `IN (`x_i`)`", and similarly "$a_i \neq x_i$" with "a_i `NOT IN (`x_i`)`". Since both static and dynamic noise layers depend on the string that defines the condition, changing the SQL expression produces independent noise values.

EXTENDING THE ATTACKS TO NON-BINARY ATTRIBUTES The differential and cloning attacks, as presented, assume that the secret attribute is binary. This is a common assumption in the literature [149, 229], as it is generally possible to extend the attacks to non-binary attributes.

We now outline, using a bisection search, how to adapt our noise-exploitation attacks for secret attributes that take values in the set of real numbers (or any set with a defined order). First, we modify our noise-exploitation attacks to test whether the secret attribute is smaller or larger than a certain starting value t_0. To do this, it is sufficient to change the condition $s = 0$ to $s \leq t_0$ (or $s \geq t_0$) in each attack query (8.7), (8.8), (8.13), (8.14). We then find the correct value (up to the desired accuracy) using a bisection search that iterates the attack with the right condition $s \leq t_i$ or $s \geq t_i$. Since the attack does not achieve perfect accuracy, each step is associated to a probability of success. This stochastic root-finding search can be solved with a probabilistic bisection algorithm (see, *e.g.*, Frazier *et al.* [257]).

8.5.5 *Reducing the number of queries*

One of the main features of Diffix is that it allows analysts to send an unlimited amount of queries. Many privacy attacks work by issuing a relatively large number of queries (see also 8.4). Limiting the number of queries allowed by Diffix would thwart or significantly affect these attacks. While our actual attack procedures require a small number of queries ($2|\Delta| + 1$ for the cloning attack), the subset exploration step can sometimes explore many sets of attributes before finding an exploitable one. To minimize the number of queries, we replace the iterative exploration with a greedy heuristic that selects only one subset which is likely to work. We focus only on the cloning attack, as it does not require an oracle and achieves much better accuracy.

The cloning attack requires a set of attributes (A', u), where the restricted record $x^{(A',u)}$ uniquely identifies the victim, but the vector $x^{(A')}$ is shared across a larger population (to avoid bucket suppression). The FullCloningAttack starts with a larger set of attributes A^* and iteratively explores subsets of A^* to find a candidate (A', u). We replace this iterative process with a single deterministic step.

Intuitively, we want u to be as discriminative as possible, while for the attributes in A' to select as many users as possible. This is what the procedure GreedySelectSubset does. First, it computes the (approximate) fraction of users that share the same value $x^{(a)}$, for each attribute $a \in A^*$. Then it selects as u the attribute associated with the lowest fraction. Now suppose that N is the estimated total

number of users in the dataset. The set A' is selected as the smallest set of attributes associated with the highest fraction, additionally requiring that the product of all the fractions for (A', u) is smaller than $1/N$. This ensures that, with high probability, the victim is uniquely identified by $x^{(A',u)}$.

Procedure GreedySelectSubset($A^*, x^{(A^*)}, s, v$)

Input: known attributes (names A^* and values $x^{(A^*)}$), secret s and target value v

Output: a set of attributes $(A', u) \subseteq A^*$

$N \leftarrow$ count() `// approx. tot. number of users`
foreach $a \in A^*$ **do**
  $C_a \leftarrow \text{count}(a = x^{(a)} \wedge s \neq v)$
  $\rho_a \leftarrow \frac{C_a}{N}$
end foreach
$\{\rho_1, \ldots, \rho_{|A^*|}\} \leftarrow \text{SortDescendingOrder}(\{\rho_a\}_{a \in A^*})$
$u \leftarrow a_{|A^*|}$
$i \leftarrow 1, \quad A' \leftarrow \varnothing$
while $\rho_u \prod_{a_i \in A'} \rho_{a_i} > \frac{1}{N}$ **do**
  $A' \leftarrow A' \cup \{a_i\}$
  $i \leftarrow i + 1$
end while
return (A', u)

We can modify the FullDifferentialAttack replacing the subset exploration with this heuristic. The modified full attack is described in the GreedyFullCloningAttack procedure.

Observe that the GreedySelectSubset procedure issues exactly $|A^*| + 1$ queries. The differential attack with the assumption validation step issues at most $2|\Delta| + 1$ queries. So, the GreedyFullCloningAttack algorithm requires at most $|A^*| + 2|\Delta| + 2$ queries.

We compared the performances of the FullCloningAttack and GreedyFullCloningAttack on the ADULT dataset, with the salary class as secret attribute and the other 10 attributes as A^*. As in section 8.2, we used $|\Delta| = 10$ dummy conditions and ran the attack on 1000 random users. The results are summarized in Table 8.3.

Procedure GreedyFullCloningAttack($A^*, x^{(A^*)}, \Delta, s, v$)

Input: known attributes (names A^* and values $x^{(A^*)}$),
dummy conditions Δ, secret s and target value v
Output: True if $x^{(s)} = v$, False if $x^{(s)} \neq v$

```
(A', u) ← GreedySelectSubset(A*, x^(A*), s, v)
if NoBucketSuppression(A', u, x^(A), Δ, s, v) and
  ValueUnique(A', u, x^(A*)) then
    return CloningAttack(A', u, x^(A), Δ, s, v)
end if
return NonAttackable
```

Subset selection	Median n. queries	Max n. queries	Predicted attackable	$\text{accuracy}_{\text{pa}}$
Iterative	304	5310	96.8%	93.3%
Heuristic	32	32	55.4%	91.7%

TABLE 8.3: Empirical results of the cloning attack with iterative subset exploration and with the heuristic subset selection.

The maximum (and median) number of queries used by GreedyFullCloningAttack for a single user is $10 + 2 \times 10 + 2 = 32$. The median number of queries used by GreedyFullCloningAttack is about 10 times higher, and the maximum is 100 times higher.

GreedyFullCloningAttack effectively attacks more than half of the users, as opposed to 96.8% of the users for the FullCloningAttack. This is due to the fact that the first attack tries a single subset of attributes per user. However, this figure is still remarkably high, given the huge reduction of required queries. Finally, the accuracy of the inference for the attacked users is almost the same.

We believe that these results give additional evidence of the power, extendability and practicability of our noise-exploitation attacks. Introducing additional optimizations, the accuracy could be improved and the number of queries could be further reduced (see Appendix 8.5.4).

8.5.6 *Attacks on implementations of differential privacy*

While differential privacy provably protects against a very large class of privacy attacks [231], the actual implementation of differentially private mechanisms can open new vulnerabilities. Haeberlen et al. [243] exploited covert channels [258] in two popular implementations of differential privacy, PINQ [236] and Airavat [238], to design state- and timing-based attacks, showing that implementation choices and bugs can introduce vulnerability, effectively disrupting the theoretical guarantees that differentially private systems are supposed to offer. Mironov showed that finite precision and rounding effects of floating-point operations can undermine the actual guarantees of differentially private mechanisms [242]. Finally, Kifer et al. argued that even perfect implementations of differential privacy are vulnerable to some attacks if the attacker has very powerful background knowledge [259], although this is debated [260].

Attacks on implementations of differential privacy suggest that no practical system is perfectly secure against any attack. For this reason, we believe that differential privacy-based systems could benefit from defence-in-depth measures such as the ones discussed in section 8.3.

In general, any query-based system, whether based on differential privacy or not, stores sensitive data on the curator's servers. This constitutes a risk, as an attacker could try to break into the server by exploiting vulnerabilities unrelated to the query interface. Once an attacker obtains access to the raw, potentially pseudonymized, individual-data, it is likely very easy for them to re-identify users.

8.6 SUMMARY

The Diffix mechanism has recently been proposed as an alternative to data anonymization methods and differential privacy, and is currently used in production. The mechanism is claimed to allow an analyst to submit an unbounded number of queries, while thwarting inference attacks, as defined by EU's Art. 29 WP. In this chapter, we show that Diffix's anonymization mechanism is vulnerable to a new class of attacks, which we call noise-exploitation attacks. Our attacks leverage design flaws in Diffix's data-dependent noise to infer

private attributes of an individual in the dataset, solely from prior knowledge about other attributes of this individual. In our opinion, Diffix alone and in its present state likely fails to satisfy the EU's Art. 29 WP requirements for data anonymization. Furthermore, our results show that naive data-dependent noise leads to highly vulnerable systems.

Our differential noise-exploitation attack, given little auxiliary information about the victim, combines specific queries and estimates how the noise is distributed to infer the value of the private attribute. In a synthetic best-case dataset, the attacker can predict with 92.6 % accuracy private attributes, using only 5 attributes.

Our cloning noise-exploitation attack extends the first one by adding "dummy" conditions that do not change the selected query set. It relies on weaker assumptions, that are automatically validated with high accuracy by our algorithm. We evaluate its performances on four real-world datasets and find that it infers private attributes of between 87.0 % and 97.0 % of all records across datasets.

Aircloak proposed a fix to our original attack in an article on their blog. Based on the high-level description in the article (no technical document is available yet), it seems that this is a mitigation that may prevent the differential and cloning attacks. However, it potentially also opens up new vulnerabilities, as it does not directly address the risk of data-dependent noise, and instead introduces a new data-dependent measure. Diffix's approach–the use of sticky noise and a security-based solutions to protect personal data–is very interesting and promising. We agree with their pragmatic approach to enable practical data analysis while minimizing the risk of individual data leakage. However, we disagree with its characterization as a "silver bullet" that, alone, is sufficient to rule out practical attacks.

Without formal privacy guarantees, it is very often impossible to prevent attacks such as the one we proposed here. Moving forward, we suggest for such query-based systems to use a layered approach that combines anonymization mechanisms with defense-in-depth measures such as access control, intrusion detection, and above all auditability.

9

CASE STUDY: AGGREGATED MOBILE PHONE METADATA

This chapter originates from a joint research article on the bandicoot toolbox [46], for which I co-leaded the development. The research was conducted during my stay at the Massachussets Institute of Technology in 2016.

The metadata generated at large scale by mobile phones and collected by every carrier around the world have the potential to fundamentally transform the way we fight diseases and design transportation systems. Scientists have compared the recent availability of these large-scale behavioral data sets to the invention of the microscope [261] and their business value is considerable. In machine learning, mobile phone metadata have been used to predict people's gender [262], age [262], personality [263], literacy rates [264]; as well as their likelihood to repay loans [265], subscribe to services [266], and commit crimes [267]. In disaster relief operations for instance, knowing the demographics of a person along with his or her mobility data is extremely valuable [268]. As most phones in low and middle income countries are prepaid, we often lack these information about the users. Being able to predict million of people's demographic information with a small training set is therefore tremendously valuable and cost-efficient.

Despite a great potential for impact and close to 10 years of academic and industry research in using mobile phone metadata [269], there were so far no open-source software to process and extract robust features from them. All prediction works were consequently based on a limited number of custom indicators. This has prevented research from progressing: features had to be redeveloped every time and numerous implementation choices (*e.g.*, reconciling data, choices of thresholds and time periods, edge cases) were lost from one paper to another. This made it hard to replicate results, quantify the impact of new methods, and transfer learnings.

In this final chapter, we present a toolbox, *bandicoot*[1], released in 2016, solving this problem by providing researchers and practitioners with an efficient open-source Python feature extractor for mobile phone metadata [46]. It belongs to the Precomputed

1 bandicoot, https://bandicoot.mit.edu.

Statistics model from Chapter 7: instead of releasing individual-level raw mobility and interaction traces, this toolbox only compute statistics, making re-identification harder (although not impossible).

bandicoot is a complete, easy-to-use and extensible environment: with a few lines of code, a user can load mobile phone data, extract features, and export them. bandicoot's modular structure makes it easy for users to add new features and leverage existing pre- and post-processing functions.

9.1 USING THE BANDICOOT TOOLBOX

bandicoot provides users with more than 1442 individual, spatial, and social network features:

INDIVIDUAL FEATURES (*e.g.*, percent of nocturnal interactions, time it takes someone to answer text message, inter-event time between two phones calls) describe an individual's phone usage and interactions with his or her contacts.

SPATIAL FEATURES (*e.g.*, entropy of visited antennas, radius of gyration) describe an individual's mobility patterns. Note that to avoid sampling biases, bandicoot groups location data per 30 min slots.

SOCIAL NETWORK (*e.g.*, clustering coefficient, assortativity) describe an individual's social network and compare his or her behavior with the one of their contacts.

bandicoot computes these indicators or, when these are distributions, their mean and standard deviation, on a weekly basis. It then returns the weekly mean and standard deviation to the user (see Fig. 9.1). The user can also compute indicators only on call/texts, weekdays/weekends, or days/nights and an extended set of weekly summary statistics (median, min, max, kurtosis, and skewness).

bandicoot also standardizes from the mobile phone research literature [269] the definition of conversations between individuals, the conversion of directed to undirected matrices, the assortativity of attributes within ego-networks, as well as a the binning scheme to avoid sampling biases in location data.

9.2 PROJECT FOCUS

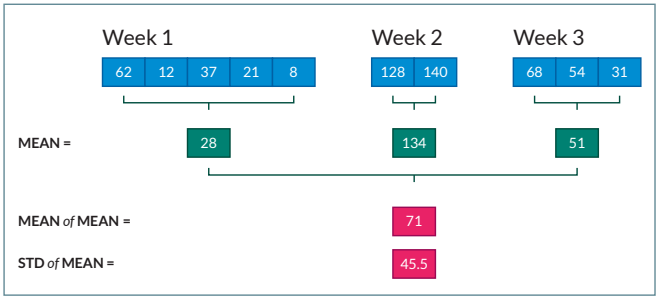
CODE QUALITY Correctness and consistency is absolutely crucial for us. The metrics implemented need to be correct and the values returned stable over time. To ensure this, we implemented functions to generate synthetic mobile phone data, regression tests, and more than 50 function-level unit tests covering 88% of the source code.

COMMUNITY-DRIVEN DEVELOPMENT The choice of the Python language, as well as a specific focus on readability, helps contributors understand and modify bandicoot. We are actively building a community of users around bandicoot to foster changes and improvements in the source code, develop new features, and to report and help correct faulty behaviors. bandicoot is hosted on GitHub and has already been developed by 10 contributors over the last years with many others helping with bug reports, features requests, and discussions.

DOCUMENTATION An extensive documentation has been written and is maintained. This includes an exhaustive description of the input and output of functions, a quick start guide, and a demonstration notebook. A specific part of the documentation is written to help users test and contribute new features.

EASY-TO-INSTALL AND WITHOUT DEPENDENCIES bandicoot runs on Python 2 and 3; we support (and test bandicoot with) GNU/Linux, Mac OS, and Windows. bandicoot is meant to (and already is) used in highly secured and heterogeneous telco environments where packages have to be approved and verified. To make its use (including in Hadoop environments) and installation easier, we developed bandicoot to be free of any dependencies such as pandas or compilers.

DETECTING DATA ISSUES Numerous data issues can happen in mobile phone data sets: incorrect locations, missing incoming records, duplicates, etc. We have built 38 reporting variables that are automatically added when exporting features. These include details about the underlying data (start and end date, number of records,



bandicoot exports all the indicators:

```
{
  "call_duration_mean_mean": 71.0,
  "call_duration_std_mean": 45.526549030940906,
  "call_duration_mean_max": 90.0,
  "call_duration_std_max": 35.440090293338699,
  ...
}
```

FIG. 9.1: bandicoot's behavioral indicators can be aggregated along and across weeks to generate further statistics.

percentage of antenna missing location information, etc.) and about the data processing (bandicoot version number, type of aggregation used, records that have been removed and why, whether a home location has been detected, etc.).

9.3 VISUALIZING PATTERNS IN USER DATA

bandicoot includes an interactive tool to visualize the data of a specific user. The visualization allows researchers to inspect the data, detect issues, and spot potentially important patterns. Fig. 9.2 shows the visualization tool, with options to display only specific weeks and to plot indicators (or their weekly mean/sum) over the time period considered. It can be generated and served on-the-fly over HTTP, or exported to disk and shared.

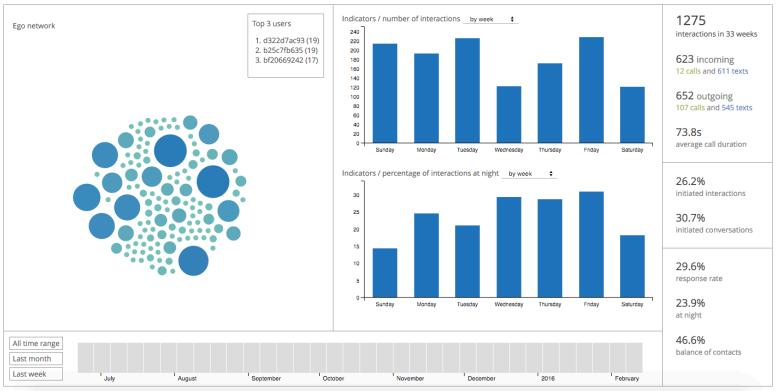


FIG. 9.2: bandicoot includes an interactive visualization tool for an individual's data.

9.4 PERFORMANCE FOR LARGE-SCALE DATASETS

While our priority developing bandicoot is to produce correct and verifiable code that can be easily extended, a certain number of implementation choices and optimizations were made to ensure that a standard large-scale dataset can be processed overnight. For instance, bandicoot caches groups of records by week, weekend, call or text to limit both the computing time and the memory footprint.

All together these optimizations have sped up computation by a factor 3 compared to a naive implementation.

We tested bandicoot on a computer with an Intel i7 CPU (2.6GHz) and 8GB of memory for users with an average of 20 records per days over 3 months. It takes on average 250ms to compute all 1442 behavioral and mobility indicators using the standard Python interpreter, CPython, and 160ms using pypy, a fast just-in-time compiler. For all indicators, including network features, the total time is 1.11s using CPython (740ms with pypy). Network indicators take significantly more time as data for all the neighbors of one node have to be loaded and their behavioral indicators computed. All indicators run in linear time with the number of records (Fig. 9.3) and a standard large-scale data set of one million mobile phone users is processed in less than 9 hours (resp. 39h for the network indicators) using the multiprocessing code we provide.

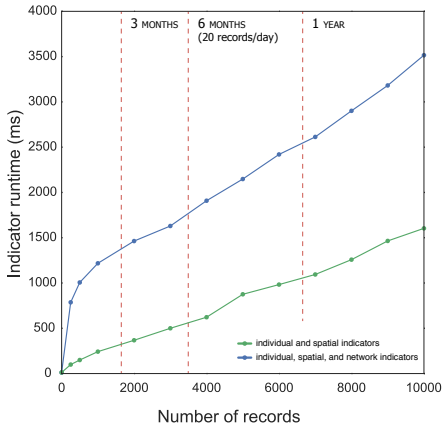


FIG. 9.3: bandicoot's runtime as a function of the number of records (single core) with and without network indicators.

9.5 CONCLUSION

The application of machine learning algorithms to mobile phone metadata has a great potential for good and businesses. Until now, there were no toolbox to efficiently process and extract robust features from them. This was a major issue for both researchers and practitioners. bandicoot implements a full data pipeline for mobile

phone data including more than 1442 features. It has been used by academic researchers as well as by carriers (*e.g.*, Orange, Telenor) and international organizations (*e.g.*, WFP).

10

CONCLUSION

With our increasing ability to collect, store and process personal data, privacy breaches and misuse of sensitive information are on the rise. To find a balance between data availability and protecting the privacy of individuals, our data are often used anonymously, *e.g.*, in research, for the training of machine learning algorithms or for the production of national statistics. Anonymization is, however, highly challenging and difficult to enforce.

This thesis reviewed how data anonymization practices are currently evaluated, spanning from disclosed medical and demographic microdata to high-dimensional datasets, as well as methods to release aggregated statistics instead of the raw data.

Firstly, it provided a critical response to the current framework used to de-identify and anonymize personal data. While anonymizing datasets through de-identification and sampling before sharing them has been the main tool used to share sensitive socio-demographic, medical, or financial data, we showed that these techniques are unlikely to satisfy the modern standards for anonymization set forth by GDPR, for example. We proposed a generative copula-based method that can accurately estimate the likelihood of a specific person to be correctly re-identified, even in a heavily incomplete dataset. On 210 populations, our method obtains AUC scores for predicting individual uniqueness ranging from 0.84 to 0.97, with a low false-discovery rate. Using our model, we found that 99.98% of Americans would be correctly re-identified in any dataset using 15 demographic attributes.

The limitations of traditional de-identiication has led to the development of modern privacy-preserving platforms. In this thesis, we secondly described two privacy breaches in Diffix, a recent and novel query-based mechanism to release aggregated statistics based on the concept of "sticky noise". According to its authors, Diffix adds less noise to answers than solutions based on differential privacy while allowing for an unlimited number of queries. Our findings

instead demonstrate that adding data-dependent noise, as done by Diffix, is not sufficient to prevent inference of private attributes.

Overall, this thesis exposes the limitations of current techniques used to widely use and share personal data. To comply with modern data protection regulations and reinstate trust in the digital economy, companies and organisations collecting and using personal data must change how they process our data.

Both in academia and industry, privacy-preserving techniques (PETs) range from practical empirical and practical guarantees (*e.g.*, statistical disclosure control, synthetic data, Diffix) to formal guarantees (*e.g.*, secure multi-party computation, differential privacy, homomorphic encryption). We should, however, not forget that no system is perfect. Digital privacy shares many similarities with computer security: protocol or implementation errors, side-channel attacks, *etc.* no system will be 100% secure forever.

For the reader eager to restore a balance between data availability and individual privacy rights, I suggest two actionable directions:

1. *transparency*, for methods and software used to design and implement PETs. Modern cryptography relies on open peer-reviewed academic research. Software based upon open free software licenses can be analyzed and vulnerabilities tested. It is time for the privacy industry to foster open transparent methods as well.

2. *defense-in-depth* and *auditability* for platforms, databases, and interactive systems using personal data. Practical secure design principles apply in digital privacy just as they do in computer security for instance. Specifically, privacy-preserving platforms (regardless of whether they have provable guarantees or not) would benefit from a defense-in-depth approach, *e.g.*, by monitoring the query stream for queries that are likely to lead to exploitation or auditing the computations and the accesses.

BIBLIOGRAPHY

[1] European Commission. Digital Single Market Strategy for Europe, *COM* **192** (2015).

[2] Jacob Poushter. Smartphone ownership and internet usage continues to climb in emerging economies, *Pew Research Center* **22** (2016).

[3] N Yang and E Hing. *National Electronic Health Records Survey*. Tech. rep. CDC (Cent. Dis. Control Prev.), 2015.

[4] Travis B Murdoch and Allan S Detsky. The Inevitable Application of Big Data to Health Care, *JAMA* **309**, n. 13 (Apr. 2013), pp. 1351–1352.

[5] Rosemary Wyber, Samuel Vaillancourt, William Perry, Priya Mannava, Temitope Folaranmi, and Leo Anthony Celi. Big data in global health: improving health in low- and middle-income countries, *Bull. World Health Organ.* **93**, n. 3 (Mar. 2015), pp. 203–208.

[6] David Lazer, Alex Sandy Pentland, Lada Adamic, Sinan Aral, Albert Laszlo Barabasi, Devon Brewer, Nicholas Christakis, Noshir Contractor, James Fowler, Myron Gutmann, et al. Life in the network: the coming age of computational social science, *Science* **323**, n. 5915 (2009), p. 721.

[7] A Halevy, P Norvig, and F Pereira. The Unreasonable Effectiveness of Data, *IEEE Intell. Syst.* (2009).

[8] Rob Kitchin. The real-time city? Big data and smart urbanism, *GeoJournal* **79**, n. 1 (Feb. 2014), pp. 1–14.

[9] Andrew McAfee, Erik Brynjolfsson, Thomas H Davenport, D J Patil, and Dominic Barton. Big Data: The Management Revolution, *Harv. Bus. Rev.* **90**, n. 10 (Oct. 2012), pp. 60–68.

[10] Hal Hodson. Revealed: Google AI has access to huge haul of NHS patient data, *New Sci.* **230(3072)** (Apr. 2016).

[11] Carole Cadwalladr and Emma Graham-Harrison. Revealed: 50 million Facebook profiles harvested for Cambridge Analytica in major data breach, *The Guardian* **17** (2018).

[12] Timothy Morey, Theodore Forbath, and Allison Schoop. Customer data: Designing for transparency and trust, *Harv. Bus. Rev.* **93**, n. 5 (May 2015), pp. 96–105.

[13] Mary Madden, Lee Rainie, Kathryn Zickuhr, Maeve Duggan, and Aaron Smith. Public perceptions of privacy and security in the post-Snowden era, *Pew Research Center* **12** (2014).

[14] TNS Opinion & Social. *Special Eurobarometer 431: Data Protection*. Tech. rep. 431. European Commission, June 2015.

[15] Paul M Schwartz and Daniel J Solove. Reconciling Personal Information in the United States and European Union, (Sept. 2013).

[16] Daniel J Solove and Danielle Keats Citron. Risk and Anxiety: A Theory of Data Breach Harms, *Tex. Law Rev.* **96**, n. 4 (Dec. 2016).

[17] P M Schwartz and D J Solove. The PII problem: Privacy and a new concept of personally identifiable information, *NYUL rev.* (2011).

[18] P Ohm. Broken promises of privacy: Responding to the surprising failure of anonymization, *UCLA Law Review* **57** (2010), p. 1701.

[19] José Luis Fernández-Alemán, Inmaculada Carrión Señor, Pedro Ángel Oliver Lozoya, and Ambrosio Toval. Security and privacy in electronic health records: a systematic literature review, *J. Biomed. Inform.* **46**, n. 3 (June 2013), pp. 541–562.

[20] M Rushanan, A D Rubin, D F Kune, and C M Swanson. "SoK: Security and Privacy in Implantable Medical Devices and Body Area Networks", *2014 IEEE Symposium on Security and Privacy*. ieeexplore.ieee.org, May 2014, pp. 524–539.

[21] N Papernot, P McDaniel, A Sinha, and M P Wellman. "SoK: Security and Privacy in Machine Learning", *2018 IEEE European Symposium on Security and Privacy (EuroS P)*. ieeexplore.ieee.org, Apr. 2018, pp. 399–414.

[22] S D Warren and L D Brandeis. Right to privacy, *Harv. Law Rev.* (1890).

[23] *Olmstead v. United States*. 1928.

[24] *Katz v. United States*. 1967.

[25] OHCHR. *The right to privacy in the digital age: Report of the United Nations High Commissioner for Human Rights*. Tech. rep. A/HRC/39/29. HRC, Mar. 2018.

[26] *Health Insurance Portability and Accountability Act*. 1996.

[27] OJ L. *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*. May 2016.

[28] Issie Lapowsky, Fred Vogelstein, Klint Finley, Tom Simonite, and Alex Baker-Whitcomb. The Man Who Saw the Dangers of Cambridge Analytica Years Ago, *Wired* (June 2018).

[29] Alex Hern. 'Anonymous' browsing data can be easily exposed, researchers reveal, *The Guardian* (Aug. 2017).

[30] Thomas Fox-Brewster. 120 Million American Households Exposed In 'Massive' ConsumerView Database Leak, *Forbes Magazine* (Dec. 2017).

[31] A Narayanan and V Shmatikov. Myths and fallacies of personally identifiable information, *Commun. ACM* (2010).

[32] P Golle. "Revisiting the uniqueness of simple demographics in the US population", *Proceedings of the 5th ACM workshop on Privacy in Electronic Society*. ACM, 2006.

[33] Latanya Sweeney. Simple demographics often identify people uniquely, *Health* **671** (2000), pp. 1–34.

[34] Latanya Sweeney. Matching Known Patients to Health Records in Washington State Data, (July 2013). arXiv: 1307.1370 [cs.CY].

[35] A Narayanan and V Shmatikov. "Robust De-anonymization of Large Sparse Datasets", *2008 IEEE Symposium on Security and Privacy*. IEEE, May 2008, pp. 111–125.

[36] Yves-Alexandre de Montjoye, César A Hidalgo, Michel Verleysen, and Vincent D Blondel. Unique in the Crowd: The privacy bounds of human mobility, *Sci. Rep.* **3** (2013), p. 1376.

[37] Yves-Alexandre de Montjoye, Laura Radaelli, Vivek Kumar Singh, and Alex "sandy" Pentland. Unique in the shopping mall: On the reidentifiability of credit card metadata, *Science* **347**, n. 6221 (Jan. 2015), pp. 536–539.

[38] Nils Homer, Szabolcs Szelinger, Margot Redman, David Duggan, Waibhav Tembe, Jill Muehling, John V Pearson, Dietrich A Stephan, Stanley F Nelson, and David W Craig. Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays, *PLoS Genet.* **4**, n. 8 (Aug. 2008), e1000167.

[39] Kevin B Jacobs, Meredith Yeager, Sholom Wacholder, David Craig, Peter Kraft, David J Hunter, Justin Paschal, Teri A Manolio, Margaret Tucker, Robert N Hoover, Gilles D Thomas, Stephen J Chanock, and Nilanjan Chatterjee. A new statistic and its power to infer membership in a genome-wide association study using genotype frequencies, *Nat. Genet.* **41**, n. 11 (Nov. 2009), pp. 1253–1257.

[40] Melissa Gymrek, Amy L McGuire, David Golan, Eran Halperin, and Yaniv Erlich. Identifying personal genomes by surname inference, *Science* **339**, n. 6117 (Jan. 2013), pp. 321–324.

[41] Jagdish Prasad Achara, Gergely Acs, and Claude Castelluccia. "On the Unicity of Smartphone Applications", *Proceedings of the 14th ACM Workshop on Privacy in the*

Electronic Society. WPES '15. Denver, Colorado, USA: ACM, 2015, pp. 27–36.

[42] Guang Chen and Sallie Keller-McNulty. Estimation of identification disclosure risk in microdata, *J. Off. Stat.* **14**, n. 1 (1998), p. 79.

[43] Nobuaki Hoshino. Applying Pitman's sampling formula to microdata disclosure risk assessment, *J. Off. Stat.* **17**, n. 4 (2001), p. 499.

[44] Fida Kamal Dankar, Khaled El Emam, Angelica Neisa, and Tyson Roffey. Estimating the re-identification risk of clinical data sets, *BMC Med. Inform. Decis. Mak.* **12** (July 2012), p. 66.

[45] Article 29 Data Protection Working Party. Opinion 05/2014 on Anonymisation Techniques, (Apr. 2014).

[46] Yves-Alexandre de Montjoye, Luc Rocher, and Alex Sandy Pentland. bandicoot: a Python Toolbox for Mobile Phone Metadata, *J. Mach. Learn. Res.* **17**, n. 175 (2016), pp. 1–5.

[47] Yves-Alexandre de Montjoye, Ali Farzanehfar, Julien M Hendrickx, and Luc Rocher. Solving artificial intelligence's privacy problem, *Field Actions Science Reports* n. 17 (2017), pp. 80–83.

[48] Yves-Alexandre de Montjoye, Sébastien Gambs, Vincent Blondel, Geoffrey Canright, Nicolas de Cordes, Sébastien Deletaille, Kenth Engø-Monsen, Manuel Garcia-Herranz, Jake Kendall, Cameron Kerry, Gautier Krings, Emmanuel Letouzé, Miguel Luengo-Oroz, Nuria Oliver, Luc Rocher, Alex Rutherford, Zbigniew Smoreda, Jessica Steele, Erik Wetter, Alex Sandy Pentland, and Linus Bengtsson. On the privacy-conscientious use of mobile phone data, *Sci Data* **5** (Dec. 2018), p. 180286.

[49] Andrea Gadotti, Florimond Houssiau, Luc Rocher, Benjamin Livshits, and Yves-Alexandre de Montjoye. "When the Signal is in the Noise: Exploiting Diffix's Sticky Noise", *USENIX Security Conference Proceedings*. 2019.

[50] Luc Rocher, Julien M Hendrickx, and Yves-Alexandre de Montjoye. Estimating the success of re-identifications in incomplete datasets using generative models, *Nat. Commun.* **10**, n. 3069 (July 2019).

[51] Shoshana Zuboff. Big other: surveillance capitalism and the prospects of an information civilization, *J. Inf. Technol. Impact* **30**, n. 1 (Mar. 2015), pp. 75–89.

[52] David J Seipp. *The right to privacy in American history*. Harvard University, Program on Information Resources Policy, 1978.

[53] Daniel J Solove. *The Digital Person: Technology and Privacy in the Information Age*. NYU Press, Dec. 2004.

[54] Natasha Singer. Acxiom, the Quiet Giant of Consumer Database Marketing, *The New York Times* (June 2012).

[55] Experian. *ConsumerViews*. 2017.

[56] Edith Ramirez, Julie Brill, Maureen K Ohlhausen, Joshua D Wright, and Terrell McSweeny. Data brokers: A call for transparency and accountability, *Fed. Trade Com.* (May 2014).

[57] Sree Vidya Bhaktavatsalam and Patrick Clark. Marriott Hit by Starwood Hack That Ranks Among Biggest Ever, *Bloomberg News* (Nov. 2018).

[58] *Data Breaches Compromised 4.5 Billion Records in First Half of 2018*. https://www.gemalto.com/press/pages/data-breaches-compromised-4-5-billion-records-in-first-half-of-2018.aspx. Accessed: NA-NA-NA. Oct. 2018.

[59] UpGuard. *Losing Face: Two More Cases of Third-Party Facebook App Data Exposure*. https://www.upguard.com/breaches/facebook-user-data-leak. Accessed: 2019-5-30.

[60] Michael Barbaro, Tom Zeller, and Saul Hansell. A face is exposed for AOL searcher no. 4417749, *NY Times* **9**, n. 2008 (2006), 8For.

[61] Gunther Eysenbach. Infodemiology: tracking flu-related searches on the web for syndromic surveillance, *AMIA Annu. Symp. Proc.* (2006), pp. 244–248.

[62] Rosie Jones, Ravi Kumar, Bo Pang, Andrew Tomkins, Andrew Tomkins, and Andrew Tomkins. "I know what you did last summer: query logs and user privacy", *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*. 2007, pp. 909–914.

[63] Roger Dingledine, Nick Mathewson, and Paul Syverson. *Tor: The Second-Generation Onion Router*. Tech. rep. 2004.

[64] E Balsa, C Troncoso, and C Diaz. "OB-PWS: Obfuscation-Based Private Web Search", *2012 IEEE Symposium on Security and Privacy*. May 2012, pp. 491–505.

[65] Daniel C Howe and Helen Nissenbaum. TrackMeNot: Resisting surveillance in web search, *Lessons from the Identity trail: Anonymity, privacy, and identity in a networked society* **23** (2009), pp. 417–436.

[66] Sai Teja Peddinti and Nitesh Saxena. Web search query privacy: Evaluating query obfuscation and anonymizing networks, *Int. J. Inf. Comput. Secur.* **22**, n. 1 (2014), pp. 155–199.

[67] Ralf Bendrath and Milton Mueller. The end of the net as we know it? Deep packet inspection and internet governance, *New Media & Society* **13**, n. 7 (Nov. 2011), pp. 1142–1160.

[68] George Dean Bissias, Marc Liberatore, David Jensen, and Brian Neil Levine. "Privacy Vulnerabilities in Encrypted HTTP Streams", *Privacy Enhancing Technologies*. Springer Berlin Heidelberg, 2006, pp. 1–11.

[69] Andrew Hintz. "Fingerprinting Websites Using Traffic Analysis", *Privacy Enhancing Technologies*. Springer Berlin Heidelberg, 2003, pp. 171–178.

[70] D R Simon, W Russell, and V N Padmanabhan. "Statistical identification of encrypted Web browsing traffic", *Proceedings 2002 IEEE Symposium on Security and Privacy*. May 2002, pp. 19–30.

[71] K P Dyer, S E Coull, T Ristenpart, and T Shrimpton. "Peek-a-Boo, I Still See You: Why Efficient Traffic Analysis Countermeasures Fail", *2012 IEEE Symposium on Security and Privacy*. May 2012, pp. 332–346.

[72] S Chen, R Wang, X Wang, and K Zhang. "Side-Channel Leaks in Web Applications: A Reality Today, a Challenge Tomorrow", *2010 IEEE Symposium on Security and Privacy*. May 2010, pp. 191–206.

[73] Thai Duong and Juliano Rizzo. "The CRIME attack", *Presentation at ekoparty Security Conference*. Vol. 110. 2012.

[74] Yoel Gluck, Neal Harris, and Angelo Prado. BREACH: reviving the CRIME attack, *Unpublished manuscript* (2013).

[75] John Kelsey. "Compression and Information Leakage of Plaintext", *Fast Software Encryption*. Springer Berlin Heidelberg, 2002, pp. 263–276.

[76] C V Wright, L Ballard, S E Coull, F Monrose, and G M Masson. "Spot Me if You Can: Uncovering Spoken Phrases in Encrypted VoIP Conversations", *2008 IEEE Symposium on Security and Privacy (sp 2008)*. May 2008, pp. 35–49.

[77] G Danezis, R Dingledine, and N Mathewson. "Mixminion: design of a type III anonymous remailer protocol", *2003 Symposium on Security and Privacy, 2003.* May 2003, pp. 2–15.

[78] Stefanos Gritzalis. Enhancing Web privacy and anonymity in the digital era, *Inf. Manage. Comput. Secur.* **12**, n. 3 (2004), pp. 255–287.

[79] Philipp Winter, Roya Ensafi, Karsten Loesing, and Nick Feamster. "Identifying and characterizing Sybils in the Tor network", *25th {USENIX} Security Symposium ({USENIX} Security 16)*. usenix.org, 2016, pp. 1169–1185.

[80] E Chan-Tin, J Shin, and J Yu. "Revisiting Circuit Clogging Attacks on Tor", *2013 International Conference on Availability, Reliability and Security*. Sept. 2013, pp. 131–140.

[81] Nicholas Hopper, Eugene Y Vasserman, and Eric Chan-TIN. How Much Anonymity Does Network Latency Leak?, *ACM Trans. Inf. Syst. Secur.* **13**, n. 2 (Mar. 2010), 13:1–13:28.

[82] S J Murdoch and G Danezis. "Low-cost traffic analysis of Tor", *2005 IEEE Symposium on Security and Privacy (S P'05)*. May 2005, pp. 183–195.

[83] Sambuddho Chakravarty, Marco V Barbera, Georgios Portokalidis, Michalis Polychronakis, and Angelos D Keromytis. "On the Effectiveness of Traffic Analysis against Anonymity Networks Using Flow Records", *Passive and Active Measurement*. Springer International Publishing, 2014, pp. 247–257.

[84] Aaron Johnson, Chris Wacek, Rob Jansen, Micah Sherr, and Paul Syverson. "Users Get Routed: Traffic Correlation on Tor by Realistic Adversaries", *Proceedings of the 2013 ACM SIGSAC Conference on Computer & Communications Security*. CCS '13. Berlin, Germany: ACM, 2013, pp. 337–348.

[85] Steven J Murdoch and Piotr Zieliński. "Sampled Traffic Analysis by Internet-Exchange-Level Adversaries", *Privacy Enhancing Technologies*. Springer Berlin Heidelberg, 2007, pp. 167–183.

[86] Benjamin Greschbach, Tobias Pulls, Laura M Roberts, Philipp Winter, and Nick Feamster. The Effect of DNS on Tor's Anonymity, (Sept. 2016). arXiv: 1609.08187 [cs.CR].

[87] Christian M Schneider, Vitaly Belik, Thomas Couronné, Zbigniew Smoreda, and Marta C González. Unravelling daily human mobility motifs, *J. R. Soc. Interface* **10**, n. 84 (July 2013), p. 20130246.

[88] Steven Goldfeder, Harry Kalodner, Dillon Reisman, and Arvind Narayanan. *When the cookie meets the blockchain: Privacy risks of web payments via cryptocurrencies*. 2018.

[89] Michael Fleder, Michael S Kester, and Sudeep Pillai. Bitcoin Transaction Graph Analysis, (Feb. 2015). arXiv: 1502.01657 [cs.CR].

[90] Elli Androulaki, Ghassan O Karame, Marc Roeschlin, Tobias Scherer, and Srdjan Capkun. "Evaluating User Privacy in Bitcoin", *Financial Cryptography and Data Security*. Springer Berlin Heidelberg, 2013, pp. 34–51.

[91] Sarah Meiklejohn, Marjori Pomarole, Grant Jordan, Kirill Levchenko, Damon McCoy, Geoffrey M Voelker, and Stefan Savage. "A fistful of bitcoins: characterizing payments among men with no names", *Proceedings of the 2013 conference on Internet measurement conference*. ACM, Oct. 2013, pp. 127–140.

[92] Alex Biryukov, Dmitry Khovratovich, and Ivan Pustogarov. "Deanonymisation of Clients in Bitcoin P2P Network", *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security*. ACM, Nov. 2014, pp. 15–29.

[93] S Goldwasser, S Micali, and C Rackoff. The Knowledge Complexity of Interactive Proof Systems, *SIAM J. Comput.* **18**, n. 1 (Feb. 1989), pp. 186–208.

[94] Malte Möser, Kyle Soska, Ethan Heilman, Kevin Lee, Henry Heffan, Shashvat Srivastava, Kyle Hogan, Jason Hennessey, Andrew Miller, Arvind Narayanan, and Nicolas Christin. "An Empirical Analysis of Traceability in the Monero Blockchain". Apr. 2018.

[95] Antoine Vastel, Pierre Laperdrix, Walter Rudametkin, and Romain Rouvoy. "FP-scanner: the privacy implications of browser fingerprint inconsistencies", *27th {USENIX} Security Symposium ({USENIX} Security 18)*. 2018, pp. 135–150.

[96] Gunes Acar, Christian Eubank, Steven Englehardt, Marc Juarez, Arvind Narayanan, and Claudia Diaz. "The Web Never Forgets: Persistent Tracking Mechanisms in the Wild", *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security*. CCS '14. Scottsdale, Arizona, USA: ACM, 2014, pp. 674–689.

[97] Károly Boda, Ádám Máté Földes, Gábor György Gulyás, and Sándor Imre. "User Tracking on the Web via Cross-Browser Fingerprinting", *Information Security Technology for Applications*. Springer Berlin Heidelberg, 2012, pp. 31–46.

[98] Peter Eckersley. A primer on information theory and privacy, *Electronic Frontier Foundation* (2010).

[99] P Laperdrix, W Rudametkin, and B Baudry. "Beauty and the Beast: Diverting Modern Web Browsers to Build Unique Browser Fingerprints", *2016 IEEE Symposium on Security and Privacy (SP)*. May 2016, pp. 878–894.

[100] Amin Faiz Khademi, Mohammad Zulkernine, and Komminist Weldemariam. "FPGuard: Detection and Prevention of Browser Fingerprinting", *Data and Applications Security and Privacy XXIX*. San Antonio, TX, USA: Springer International Publishing, 2015, pp. 293–308.

[101] Christof Ferreira Torres, Hugo Jonker, and Sjouke Mauw. "FP-Block: Usable Web Privacy by Controlling Browser Fingerprinting", *Computer Security – ESORICS 2015*. Springer International Publishing, 2015, pp. 3–19.

[102] Alejandro Gómez-Boix, Pierre Laperdrix, and Benoit Baudry. "Hiding in the Crowd: an Analysis of the Effectiveness of Browser Fingerprinting at Large Scale", *WWW 2018: The 2018 Web Conference*. 2018.

[103] Hui Zang and Jean Bolot. "Anonymization of Location Data Does Not Work: A Large-scale Measurement Study", *Proceedings of the 17th Annual International Conference on Mobile Computing and Networking*. MobiCom '11. Las Vegas, Nevada, USA: ACM, 2011, pp. 145–156.

[104] Gábor György Gulyás, Gergely Acs, and Claude Castelluccia. Near-Optimal Fingerprinting with Constraints, *Proceedings on Privacy Enhancing Technologies* **2016**, n. 4 (Oct. 2016), pp. 470–487.

[105] Mathy Vanhoef, Célestin Matte, Mathieu Cunche, Leonardo S Cardoso, and Frank Piessens. "Why MAC Address Randomization is Not Enough: An Analysis of Wi-Fi Network Discovery Mechanisms", *Proceedings of the 11th ACM on Asia Conference on Computer and Communications Security*. ASIA CCS '16. Xi'an, China: ACM, 2016, pp. 413–424.

[106] Piotr Sapiezynski, Arkadiusz Stopczynski, Radu Gatej, and Sune Lehmann. Tracking Human Mobility Using WiFi Signals, *PLoS One* **10**, n. 7 (July 2015), e0130824.

[107] Miro Enev, Alex Takakuwa, Karl Koscher, and Tadayoshi Kohno. Automobile Driver Fingerprinting, *Proceedings on Privacy Enhancing Technologies* **2016**, n. 1 (Jan. 2016), pp. 34–50.

[108] Kashmir Hill. Camera Company That Let Hackers Spy On Naked Customers Ordered By FTC To Get Its Security Act Together, *Forbes Magazine* (Sept. 2013).

[109] Erik Buchmann, Klemens Böhm, Thorben Burghardt, and Stephan Kessler. Re-identification of Smart Meter Data, *Pers. Ubiquit. Comput.* **17**, n. 4 (Apr. 2013), pp. 653–662.

[110] Valentin Tudor, Magnus Almgren, and Marina Papatriantafilou. "A Study on Data De-pseudonymization in the Smart Grid", *Proceedings of the Eighth European Workshop on System Security*. EuroSec '15. Bordeaux, France: ACM, 2015, 2:1–2:6.

[111] Ross J Anderson and Shailendra Fuloria. "On the Security Economics of Electricity Metering", *WEIS*. 2010.

[112] Alfredo Rial and George Danezis. "Privacy-preserving Smart Metering", *Proceedings of the 10th Annual ACM Workshop on Privacy in the Electronic Society*. WPES '11. Chicago, Illinois, USA: ACM, 2011, pp. 49–60.

[113] Antonio Regalado. 2017 was the year consumer DNA testing blew up, *MIT Technology Review* (Feb. 2018).

[114] E Ayday, E De Cristofaro, J Hubaux, and G Tsudik. Whole Genome Sequencing: Revolutionary Medicine or Privacy Nightmare?, *Computer* **48**, n. 2 (Feb. 2015), pp. 58–66.

[115] Arif Harmanci and Mark Gerstein. Analysis of sensitive information leakage in functional genomics signal profiles through genomic deletions, *Nat. Commun.* **9**, n. 1 (June 2018), p. 2453.

[116] Arif Harmanci and Mark Gerstein. Quantification of private information leakage from phenotype-genotype data: linking attacks, *Nat. Methods* **13**, n. 3 (Mar. 2016), pp. 251–256.

[117] Meng Wang, Zhanglong Ji, Shuang Wang, Jihoon Kim, Hai Yang, Xiaoqian Jiang, and Lucila Ohno-Machado. Mechanisms to protect the privacy of families when using the transmission disequilibrium test in genome-wide association studies, *Bioinformatics* **33**, n. 23 (Dec. 2017), pp. 3716–3725.

[118] Network Advertising Initiative. *NAI Code of Conduct*. Tech. rep. June 2017.

[119] Federal Trade Commission. *Protecting consumer privacy in an era of rapid change: Recommendations for businesses and policymakers*. Tech. rep. Mar. 2012.

[120] Jules Polonetsky, Omer Tene, and Kelsey Finch. Shades of Gray: Seeing the Full Spectrum of Practical Data De-Identification, *Santa Clara Law Rev.* **56**, n. 3 (June 2016), pp. 593–629.

[121] Gregory J Matthews and Ofer Harel. Data confidentiality: A review of methods for statistical disclosure limitation and methods for assessing privacy, *Stat. Surv.* **5**, n. 0 (2011), pp. 1–29.

[122] Arvind Narayanan and Edward W Felten. *No silver bullet: De-identification still doesn't work*. http://randomwalker.info/publications/no-silver-bullet-de-identification.pdf. 2014.

[123] Kobbi Nissim and Alexandra Wood. Is privacy privacy?, *Philos. Trans. A Math. Phys. Eng. Sci.* **376**, n. 2128 (Sept. 2018).

[124] Isabel Wagner and David Eckhoff. Technical Privacy Metrics: A Systematic Survey, *ACM Comput. Surv.* **51**, n. 3 (June 2018), 57:1–57:38.

[125] P Samarati and L Sweeney. "Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression". 1998.

[126] Khaled El Emam, Fida Kamal Dankar, Romeo Issa, Elizabeth Jonker, Daniel Amyot, Elise Cogo, Jean-Pierre Corriveau, Mark Walker, Sadrul Chowdhury, Regis Vaillancourt, Tyson Roffey, and Jim Bottomley. A globally optimal k-anonymity method for the de-identification of

health data, *J. Am. Med. Inform. Assoc.* **16**, n. 5 (Sept. 2009), pp. 670–682.

[127] Bugra Gedik and Ling Liu. *A customizable k-anonymity model for protecting location privacy*. Tech. rep. Georgia Institute of Technology, 2004.

[128] Roberto J Bayardo and Rakesh Agrawal. "Data privacy through optimal k-anonymization", *21st International conference on data engineering (ICDE'05)*. 2005, pp. 217–228.

[129] J Domingo-Ferrer and V Torra. "A Critique of k-Anonymity and Some of Its Enhancements", *Availability, Reliability and Security, 2008. ARES 08. Third International Conference on*. Mar. 2008, pp. 990–993.

[130] Ashwin Machanavajjhala, Daniel Kifer, Johannes Gehrke, and Muthuramakrishnan Venkitasubramaniam. L-diversity: Privacy Beyond K-anonymity, *ACM Trans. Knowl. Discov. Data* **1**, n. 1 (Mar. 2007).

[131] N Li, T Li, and S Venkatasubramanian. "t-Closeness: Privacy Beyond k-Anonymity and l-Diversity", *2007 IEEE 23rd International Conference on Data Engineering*. ieeexplore.ieee.org, Apr. 2007, pp. 106–115.

[132] *How Google Anonymize Data*. https://policies.google.com/technologies/anonymization. Accessed: 2019-5-29. Aug. 2018.

[133] Fabian Prasser, Florian Kohlmayer, Ronald Lautenschläger, and Klaus A Kuhn. "ARX–A Comprehensive Tool for Anonymizing Biomedical Data", *AMIA Annu Symp Proc.* Vol. 2014. Nov. 2014, pp. 984–993.

[134] C C Aggarwal. "On k-anonymity and the curse of dimensionality", *Proceedings of the 31st international conference on Very large data bases*. dl.acm.org, 2005.

[135] Andreas Pfitzmann and Marit Hansen. "A terminology for talking about privacy by data minimization: Anonymity, unlinkability, undetectability, unobservability, pseudonymity, and identity management". 2010.

[136] Claudia Diaz. "Anonymity metrics revisited", *Dagstuhl Seminar Proceedings*. 2006.

[137] Andrei Serjantov and George Danezis. "Towards an Information Theoretic Metric for Anonymity", *Privacy Enhancing Technologies*. Ed. by Roger Dingledine and Paul Syverson. Vol. 2482. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer Berlin Heidelberg, June 2003, pp. 41–53.

[138] Edoardo M Airoldi, Xue Bai, and Bradley A Malin. An Entropy Approach to Disclosure Risk Assessment: Lessons from Real Applications and Simulated Domains, *Decis. Support Syst.* **51**, n. 1 (Apr. 2011), pp. 10–20.

[139] M Bezzi. "An entropy based method for measuring anonymity", *2007 Third International Conference on Security and Privacy in Communications Networks and the Workshops - SecureComm 2007*. Sept. 2007, pp. 28–32.

[140] Fabian Prasser, Raffael Bild, and Klaus A Kuhn. A Generic Method for Assessing the Quality of De-Identified Health Data, *Stud. Health Technol. Inform.* **228** (2016), pp. 312–316.

[141] Khaled El Emam and Luk Arbuckle. *Anonymizing Health Data*. O'Reilly, Dec. 2013.

[142] Bradley Malin. Re-identification of familial database records, *AMIA Annu. Symp. Proc.* (2006), pp. 524–528.

[143] Rosemary Braun, William Rowe, Carl Schaefer, Jinghui Zhang, and Kenneth Buetow. Needles in the haystack: identifying individuals present in pooled genomic data, *PLoS Genet.* **5**, n. 10 (Oct. 2009), e1000668.

[144] Sriram Sankararaman, Guillaume Obozinski, Michael I Jordan, and Eran Halperin. Genomic privacy and limits of individual detection in a pool, *Nat. Genet.* **41**, n. 9 (Sept. 2009), pp. 965–967.

[145] Peter M Visscher and William G Hill. The limits of individual identification from sample allele frequencies: theory and statistical analysis, *PLoS Genet.* **5**, n. 10 (Oct. 2009), e1000628.

[146] Rui Wang, Yong Fuga Li, Xiaofeng Wang, Haixu Tang, and Xiaoyong Zhou. "Learning Your Identity and Disease from Research Papers: Information Leaks in Genome Wide Association Study", *Proceedings of the 16th ACM Conference on Computer and Communications Security*. CCS '09. Chicago, Illinois, USA: ACM, 2009, pp. 534–544.

[147] Michael Backes, Pascal Berrang, Mathias Humbert, and Praveen Manoharan. "Membership Privacy in MicroRNA-based Studies", *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*. CCS '16. Vienna, Austria: ACM, 2016, pp. 319–330.

[148] Niklas Buescher, Spyros Boukoros, Stefan Bauregger, and Stefan Katzenbeisser. Two is not enough: Privacy assessment of aggregation schemes in smart metering, *Proceedings on Privacy Enhancing Technologies* **2017**, n. 4 (2017), pp. 198–214.

[149] Cynthia Dwork, Adam Smith, Thomas Steinke, and Jonathan Ullman. Exposed! A Survey of Attacks on Private Data, (Mar. 2017).

[150] Apostolos Pyrgelis, Carmela Troncoso, and Emiliano De Cristofaro. "Knock Knock, Who's There? Membership Inference on Aggregate Location Data", *Proceedings of the 25th Network and Distributed System Security Symposium*. 2018.

[151] Fengli Xu, Zhen Tu, Yong Li, Pengyu Zhang, Xiaoming Fu, and Depeng Jin. "Trajectory Recovery From Ash: User Privacy Is NOT Preserved in Aggregated Mobility Data", *Proceedings of the 26th International Conference on World Wide Web*. WWW '17. Perth, Australia: International World Wide Web Conferences Steering Committee, 2017, pp. 1241–1250.

[152] Apostolos Pyrgelis. *On Location, Time, and Membership: Studying How Aggregate Location Data Can Harm Users' Privacy – Bentham's Gaze*. https://www.benthamsgaze.org/2018/10/02/on-location-time-and-membership-studying-how-aggregate-location-data-can-harm-users-privacy/. Accessed: 2019-7-8. Oct. 2018.

[153] Office of the Australian Information Commissioner. *De-identification and the Privacy Act*. OAIC. Mar. 2018.

[154] P Samarati. Protecting respondents identities in microdata release, *IEEE Trans. Knowl. Data Eng.* **13**, n. 6 (Nov. 2001), pp. 1010–1027.

[155] Ali Amiri. Dare to share: Protecting sensitive knowledge with data sanitization, *Decis. Support Syst.* **43**, n. 1 (Feb. 2007), pp. 181–191.

[156] Lawrence H Cox. Network Models for Complementary Cell Suppression, *J. Am. Stat. Assoc.* **90**, n. 432 (Dec. 1995), pp. 1453–1462.

[157] Lawrence H Cox. A Constructive Procedure for Unbiased Controlled Rounding, *J. Am. Stat. Assoc.* **82**, n. 398 (June 1987), pp. 520–524.

[158] J M Gouweleeuw, Peter Kooiman, and P P De Wolf. Post randomisation for statistical disclosure control: Theory and implementation, *J. Off. Stat.* **14**, n. 4 (1998), p. 463.

[159] Stephen E Fienberg and Russell J Steele. Disclosure limitation using perturbation and related methods for categorical data, *J. Off. Stat.* **14**, n. 4 (1998), p. 485.

[160] Simson L Garfinkel. De-identification of personal information, *NISTIR* **8053** (2015), pp. 1–46.

[161] J Domingo-Ferrer and J M Mateo-Sanz. Practical data-oriented microaggregation for statistical disclosure control, *IEEE Trans. Knowl. Data Eng.* **14**, n. 1 (Jan. 2002), pp. 189–201.

[162] F Kohlmayer, F Prasser, C Eckert, A Kemper, and K A Kuhn. "Flash: Efficient, Stable and Optimal K-Anonymity", *2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Confernece on Social Computing*. Sept. 2012, pp. 708–717.

[163] Dan Zhu, Xiao-Bai Li, and Shuning Wu. *Identity disclosure protection: A data reconstruction approach for privacy-preserving data mining*. 2009.

[164] Natasha Gilbert. Researchers criticize genetic data restrictions, *Nature* **200810** (2008).

[165] Yves-Alexandre de Montjoye and Alex Sandy Pentland. Response to Comment on "Unique in the shopping mall: On the reidentifiability of credit card metadata", *Science* **351**, n. 6279 (Mar. 2016), p. 1274.

[166] J K Trotter. *Public NYC Taxicab Database Lets You See How Celebrities Tip*. http://gawker.com/the-public-nyc-taxicab-database-that-accidentally-track-1646724546. Accessed: 2019-2-7. Oct. 2014.

[167] James Siddle. *I Know Where You Were Last Summer: London's public bike data is telling everyone where you've been*. https://vartree.blogspot.com/2014/04/i-know-where-you-were-last-summer.html. Accessed: 2019-2-7. Apr. 2014.

[168] Chris Culnane, Benjamin I P Rubinstein, and Vanessa Teague. Health Data in an Open World, (Dec. 2017). arXiv: 1712.05627 [cs.CY].

[169] Jeffrey Mervis. Researchers object to census privacy measure, *Science* **363**, n. 6423 (Jan. 2019), p. 114.

[170] Ann Cavoukian and Daniel Castro. Big data and innovation, setting the record straight: De-identification does work, *White Paper, Jun* (2014), p. 20.

[171] Office for Civil Rights, HHS. *Standards for privacy of individually identifiable health information*. Federal Register. Aug. 2002, pp. 53181–53273.

[172] Bradley Malin, Kathleen Benitez, and Daniel Masys. Never too old for anonymity: a statistical standard for demographic data sharing via the HIPAA Privacy Rule, *J. Am. Med. Inform. Assoc.* **18**, n. 1 (Jan. 2011), pp. 3–10.

[173] Mark A Rothstein. Is deidentification sufficient to protect health privacy in research?, *Am. J. Bioeth.* **10**, n. 9 (Sept. 2010), pp. 3–11.

[174] L Sweeney. Weaving technology and policy together to maintain confidentiality, *J. Law Med. Ethics* **25**, n. 2-3 (1997), pp. 98–110, 82.

[175] Grigorios Loukides, Joshua C Denny, and Bradley Malin. The disclosure of diagnosis codes can breach research participants' privacy, *J. Am. Med. Inform. Assoc.* **17**, n. 3 (May 2010), pp. 322–327.

[176] M Douriez, H Doraiswamy, J Freire, and C T Silva. "Anonymizing NYC Taxi Data: Does It Matter?", *2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*. Oct. 2016, pp. 140–148.

[177] A Lavrenovs and K Podins. "Privacy violations in Riga open data public transport system", *2016 IEEE 4th Workshop on Advances in Information, Electronic and Electrical Engineering (AIEEE)*. Nov. 2016, pp. 1–6.

[178] Daniel Barth-Jones. *The 'Re-Identification' of Governor William Weld's Medical Information: A Critical Re-Examination of Health Data Identification Risks and Privacy Protections, Then and Now*. July 2012.

[179] K El Emam and L Arbuckle. *De-identification: A critical debate*. https://fpf.org/2014/07/24/de-identification-a-critical-debate/. 2014.

[180] David Sánchez, Sergio Martinez, and Josep Domingo-Ferrer. Comment on "Unique in the shopping mall: On the reidentifiability of credit card metadata", *Science* **351**, n. 6279 (Mar. 2016), p. 1274.

[181] Jerome P Reiter. Estimating Risks of Identification Disclosure in Microdata, *J. Am. Stat. Assoc.* **100**, n. 472 (Dec. 2005), pp. 1103–1112.

[182] Stephen E Fienberg and Ashish P Sanil. A Bayesian Approach to Data Disclosure: Optimal Intruder Behavior for Continuous Data, *J. Off. Stat.* **13**, n. 1 (Mar. 1997), p. 75.

[183] George Duncan and Diane Lambert. The Risk of Disclosure for Microdata, *J. Bus. Econ. Stat.* **7**, n. 2 (Apr. 1989), pp. 207–217.

[184] Steven Ruggles, Miriam L King, Deborah Levison, Robert McCaa, and Matthew Sobek. IPUMS-International, *Hist. Methods* **36**, n. 2 (Jan. 2003), pp. 60–65.

[185] J Bennett and S Lanning. "The Netflix Prize", *Proceedings of KDD Cup and workshop*. 2007, p. 35.

[186] Christian Genest and Jock Mackay. The Joy of Copulas: Bivariate Distributions with Uniform Marginals, *Am. Stat.* **40**, n. 4 (Nov. 1986), pp. 280–283.

[187] Umberto Cherubini, Elisa Luciano, and Walter Vecchiato. *Copula Methods in Finance*. John Wiley & Sons, Oct. 2004.

[188] Christian Genest and Anne-Catherine Favre. Everything You Always Wanted to Know about Copula Modeling but Were Afraid to Ask, *J. Hydrol. Eng.* **12**, n. 4 (July 2007), pp. 347–368.

[189] Weijing Wang and Martin T Wells. Model Selection and Semiparametric Inference for Bivariate Failure-Time Data, *J. Am. Stat. Assoc.* **95**, n. 449 (Mar. 2000), pp. 62–72.

[190] Nguyen Xuan Vinh, Julien Epps, and James Bailey. Information Theoretic Measures for Clusterings Comparison: Variants, Properties, Normalization and Correction for Chance, *J. Mach. Learn. Res.* **11**, n. Oct (2010), pp. 2837–2854.

[191] Alan Genz. Numerical Computation of Multivariate Normal Probabilities, *J. Comput. Graph. Stat.* **1**, n. 2 (June 1992), pp. 141–149.

[192] Alan Genz and Frank Bretz. *Computation of Multivariate Normal and t Probabilities*. Springer Science & Business Media, July 2009.

[193] Glen W Brier. Verification of Forecasts Expressed in terms of probability, *Mon. Weather Rev.* **78**, n. 1 (1950), pp. 1–3.

[194] Phee Waterfield and Timothy Revell. Huge new Facebook data leak exposed intimate details of 3m users, *New Scientist* (2018), pp. 82–100.

[195] W J Ewens. The sampling theory of selectively neutral alleles, *Theor. Popul. Biol.* **3**, n. 1 (Mar. 1972), pp. 87–112.

[196] W J Keller and J Pannekoek. Disclosure Control of Microdata, *J. Am. Stat. Assoc.* **85**, n. 409 (Mar. 1990), pp. 38–45.

[197] Jim Pitman. Random discrete distributions invariant under size-biased permutation, *Adv. Appl. Probab.* **28**, n. 2 (June 1996), pp. 525–539.

[198] C J Skinner and D J Holmes. Estimating the re-identification risk per record in microdata, *J. Off. Stat.* **14**, n. 4 (1998), p. 361.

[199] Chris Skinner and Natalie Shlomo. Assessing Identification Risk in Survey Microdata Using Log-Linear Models, *J. Am. Stat. Assoc.* **103**, n. 483 (Sept. 2008), pp. 989–1001.

[200] Laura Zayatz. *Estimation of the Percent of Unique Population Elements on a Microdata File Using the Sample*. Tech. rep. 91/08. Statistical Research Division, Bureau of the Census, 1991.

[201] Latanya Sweeney. k-anonymity: a model for protecting privacy, *Internat. J. Uncertain. Fuzziness Knowledge-Based Systems* **10**, n. 05 (2002), pp. 557–570.

[202] Adam Meyerson and Ryan Williams. "On the complexity of optimal k-anonymity", *Proceedings of the 23rd ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*. 2004, pp. 223–228.

[203] Giuseppe D'Acquisto, Josep Domingo-Ferrer, Panayiotis Kikiras, Vicenç Torra, Yves-Alexandre de Montjoye, and Athena Bourka. *Privacy by design in big data: An overview of privacy enhancing technologies in the era of big data analytics*. Tech. rep. European Union Agency for Network and Information Security, Dec. 2015.

[204] Hyunghoon Cho, David J Wu, and Bonnie Berger. Secure genome-wide association analysis using multiparty computation, *Nat. Biotechnol.* **36**, n. 6 (July 2018), pp. 547–551.

[205] Magnus Jändel. Decision support for releasing anonymised data, *Comput. Secur.* **46** (Oct. 2014), pp. 48–61.

[206] Yaniv Erlich and Arvind Narayanan. Routes for breaching and protecting genetic privacy, *Nat. Rev. Genet.* **15**, n. 6 (June 2014), pp. 409–421.

[207] Martin Scaiano, Stephen Korte, Andrew Baker, Geoffrey Green, Khaled El Emam, and Luk Arbuckle. "Re-identification risk measurement estimation of a dataset". 20180114037:A1. Apr. 2018.

[208] WSJ Staff. *The Information That Is Needed to Identify You: 33 Bits*. https://blogs.wsj.com/digits/2010/08/04/the-information-that-is-needed-to-identify-you-33-bits/. Accessed: 2019-3-24. Aug. 2010.

[209] A Korolova. "Privacy Violations Using Microtargeted Ads: A Case Study", *2010 IEEE International Conference on Data Mining Workshops*. IEEE, Dec. 2010, pp. 474–482.

[210] Ilya Nemenman, F Shafee, and William Bialek. "Entropy and Inference, Revisited", *Advances in Neural Information Processing Systems 14*. Ed. by T G Dietterich, S Becker, and Z Ghahramani. MIT Press, 2002, pp. 471–478.

[211] Evan Archer, Il Memming Park, and Jonathan W Pillow. Bayesian Entropy Estimation for Countable Discrete Distributions, *J. Mach. Learn. Res.* **15** (2014), pp. 2833–2868.

[212] Owen T Courtney and Ginestra Bianconi. Dense power-law networks and simplicial complexes, *Phys Rev E* **97**, n. 5-1 (May 2018), p. 052303.

[213] Erik B Sudderth and Michael I Jordan. "Shared Segmentation of Natural Scenes Using Dependent Pitman-Yor Processes", *Advances in Neural Information Processing Systems 21*. Ed. by D Koller, D Schuurmans, Y Bengio, and L Bottou. Curran Associates, Inc., 2009, pp. 1585–1592.

[214] Hemant Ishwaran and Lancelot F James. Generalized weighted Chinese restaurant processes for species sampling mixture models, *Stat. Sin.* **13**, n. 4 (2003), pp. 1211–1235.

[215] Raymond W Yeung. *A First Course in Information Theory*. Springer Science & Business Media, Dec. 2012.

[216] Raymond W Yeung. "The Science of Information", *Information Theory and Network Coding*. Ed. by Raymond W Yeung. Boston, MA: Springer US, 2008, pp. 1–4.

[217] Peter Eckersley. "How Unique Is Your Web Browser?", *Privacy Enhancing Technologies*. Springer Berlin Heidelberg, 2010, pp. 1–18.

[218] Nga Nguyen Thi Thanh, Khanh Nguyen Kim, Son Ngo Hong, and Trung Ngo Lam. Entropy Correlation and Its Impacts on Data Aggregation in a Wireless Sensor Network, *Sensors* **18**, n. 9 (Sept. 2018).

[219] R W Yeung. A framework for linear information inequalities, *IEEE Trans. Inf. Theory* **43**, n. 6 (Nov. 1997), pp. 1924–1934.

[220] President's Council of Advisors on Science and Technology. *Big Data and Privacy: a technological perspective*. Tech. rep. White House, Jan. 2014, pp. 1–76.

[221] Jeffrey Mervis. *How Two Economists Got Direct Access to IRS Tax Records*. https://sciencemag.org/news/2014/05/how-two-economists-got-direct-access-irs-tax-records. Accessed: 2019-8-15. Oct. 2014.

[222] D R Musicant, J M Christensen, and J F Olson. "Supervised Learning by Training on Aggregate Outputs", *Seventh IEEE International Conference on Data Mining (ICDM 2007)*. Oct. 2007, pp. 252–261.

[223] Novi Quadrianto, Alex J Smola, Tibério S Caetano, and Quoc V Le. Estimating Labels from Label Proportions, *J. Mach. Learn. Res.* **10**, n. Oct (2009), pp. 2349–2374.

[224] Jamie Hayes, Luca Melis, George Danezis, and Emiliano De Cristofaro. LOGAN: Membership inference attacks against generative models, *Proceedings on Privacy Enhancing Technologies* **2019**, n. 1 (2019), pp. 133–152.

[225] Cynthia Dwork. "Differential Privacy: A Survey of Results", *Theory and Applications of Models of Computation*. Ed. by Manindra Agrawal, Dingzhu Du, Zhenhua Duan, and Angsheng Li. Lecture Notes in Computer Science. Springer Berlin Heidelberg, Apr. 2008, pp. 1–19.

[226] Paul Francis, Sebastian Probst-Eide, Pawel Obrok, Cristian Berneanu, Sasa Juric, and Reinhard Munz. Extended Diffix, (June 2018). arXiv: 1806.02075 [cs.CR].

[227] Dorothy E Denning. "Are statistical data bases secure", *Proc. AFIPS*. Vol. 2978. 1978, pp. 525–530.

[228] L L Beck. A security mechanism for statistical database, *ACM Transactions on Database Systems (TODS)* (1980).

[229] Irit Dinur and Kobbi Nissim. *Revealing information while preserving privacy*. 2003.

[230] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. *Calibrating Noise to Sensitivity in Private Data Analysis*. 2017.

[231] Shiva P Kasiviswanathan and Adam Smith. On the 'Semantics' of Differential Privacy: A Bayesian Formulation, *J. Priv. Conf.* **6**, n. 1 (June 2014).

[232] Andreas Haeberlen, Benjamin C Pierce, and Arjun Narayan. "Differential Privacy Under Fire", *USENIX Security Symposium*. 2011.

[233] Ilya Mironov. "On significance of the least significant bits for differential privacy", *Proceedings of the 2012 ACM conference on Computer and communications security*. 2012, pp. 650–661.

[234] M Andrysco, D Kohlbrenner, K Mowery, R Jhala, S Lerner, and H Shacham. "On Subnormal Floating Point and Abnormal Timing", *2015 IEEE Symposium on Security and Privacy*. May 2015, pp. 623–639.

[235] Noah Johnson, Joseph P Near, and Dawn Song. Towards Practical Differential Privacy for SQL Queries, *Proceedings VLDB Endowment* **11**, n. 5 (Jan. 2018), pp. 526–539.

[236] Frank D McSherry. "Privacy Integrated Queries: An Extensible Platform for Privacy-preserving Data Analysis", *Proceedings of the 2009 ACM SIGMOD International Conference on Management of Data*. SIGMOD '09. Providence, Rhode Island, USA: ACM, 2009, pp. 19–30.

[237] Davide Proserpio, Sharon Goldberg, and Frank McSherry. Calibrating Data to Sensitivity in Private Data Analysis: A Platform for Differentially-private Analysis of Weighted Datasets, *Proceedings VLDB Endowment* **7**, n. 8 (Apr. 2014), pp. 637–648.

[238] Indrajit Roy, Srinath T V Setty, Ann Kilzer, Vitaly Shmatikov, and Emmett Witchel. "Airavat: Security and Privacy for MapReduce", *NSDI*. Vol. 10. 2010, pp. 297–312.

[239] Prashanth Mohan, Abhradeep Thakurta, Elaine Shi, Dawn Song, and David Culler. "GUPT: Privacy Preserving Data Analysis Made Easy", *Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data*. SIGMOD '12. Scottsdale, Arizona, USA: ACM, 2012, pp. 349–360.

[240] Frank McSherry. *Uber's differential privacy .. probably isn't*. https://github.com/frankmcsherry/blog/blob/master/posts/2018-02-25.md. 2018.

[241] Jun Tang, Aleksandra Korolova, Xiaolong Bai, Xueqiang Wang, and Xiaofeng Wang. Privacy Loss in Apple's Implementation of Differential Privacy on MacOS 10.12, (Sept. 2017). arXiv: 1709.02753 [cs.CR].

[242] Ilya Mironov. "On Significance of the Least Significant Bits for Differential Privacy", *Proceedings of the 2012 ACM Conference on Computer and Communications Security*. CCS '12. Raleigh, North Carolina, USA: ACM, 2012, pp. 650–661.

[243] Andreas Haeberlen, Benjamin C Pierce, and Arjun Narayan. "Differential Privacy Under Fire", *USENIX Security Symposium*. 2011.

[244] Kobbi Nissim, Sofya Raskhodnikova, and Adam Smith. "Smooth Sensitivity and Sampling in Private Data Analysis", *Proceedings of the Thirty-ninth Annual ACM Symposium on Theory of Computing*. STOC '07. San Diego, California, USA: ACM, 2007, pp. 75–84.

[245] Paul Francis, Sebastian Probst Eide, and Reinhard Munz. "Diffix: High-Utility Database Anonymization", *Privacy Technologies and Policy*. Springer International Publishing, 2017, pp. 141–158.

[246] Aircloak. *Building Trust*. https://web.archive.org/web/20180426104935/https://blog.aircloak.com/building-trust-35c74424efc6. July 2017.

[247] Aircloak. *Announcing our Seed investment*. https://web.archive.org/web/20180426131132/https://blog.aircloak.com/announcing-our-seed-investment-26f392f1068a. Oct. 2017.

[248] William Stallings and Lawrie Brown. *Computer security: principles and practice*. Pearson, Aug. 2017.

[249] Nabil R Adam and John C Worthmann. Security-control methods for statistical databases: a comparative study, *ACM Computing Surveys (CSUR)* **21**, n. 4 (Dec. 1989), pp. 515–556.

[250] Dheeru Dua and Casey Graff. *UCI Machine Learning Repository*. 2017.

[251] Frank McSherry. *Statistical inference considered harmful*. https://github.com/frankmcsherry/blog/blob/master/posts/2016-06-14.md. 2016.

[252] Aloni Cohen and Kobbi Nissim. *Linear Program Reconstruction in Practice*. TPDP 2018. Toronto, Canada, Oct. 2018.

[253] Cynthia Dwork, Frank McSherry, and Kunal Talwar. "The price of privacy and the limits of LP decoding", *Proceedings of the thirty-ninth annual ACM symposium on Theory of computing*. ACM, June 2007, pp. 85–94.

[254] Paul Francis. *Fix for the MIT/Georgetown Univ Attack on Diffix | Aircloak*. https://aircloak.com/fix-for-the-mit-georgetown-univ-attack-on-diffix/. Accessed: 2019-7-8. Oct. 2018.

[255] G A Young and Richard L Smith. *Essentials of statistical inference*. Cambridge University Press, 2005.

[256] Munuswamy Sankaran. Approximations to the non-central chi-square distribution, *Biometrika* **50**, n. 1–2 (June 1963), pp. 199–204.

[257] Peter I Frazier, Shane G Henderson, and Rolf Waeber. Probabilistic Bisection Converges Almost as Quickly as Stochastic Approximation, *arXiv preprint arXiv:1612.03964* (2016).

[258] Butler W Lampson. A Note on the Confinement Problem, *Commun. ACM* **16**, n. 10 (Oct. 1973), pp. 613–615.

[259] Daniel Kifer and Ashwin Machanavajjhala. "No Free Lunch in Data Privacy", *Proceedings of the 2011 ACM SIGMOD International Conference on Management of Data*. SIGMOD '11. Athens, Greece: ACM, 2011, pp. 193–204.

[260] Frank McSherry. *Lunchtime for Data Privacy*. https://github.com/frankmcsherry/blog/blob/master/posts/2016-08-16.md. 2016.

[261] Jim Giles. Making the links, *Nature* **488**, n. 7412 (2012), pp. 448–450.

[262] Carlos Sarraute, Pablo Blanc, and Javier Burroni. "A study of age and gender seen through mobile phone usage patterns in Mexico", *Advances in Social Networks Analysis and Mining, 2014 IEEE/ACM International Conference on*. 2014, pp. 836–843.

[263] Yves-Alexandre de Montjoye, Jordi Quoidbach, Florent Robic, and Alex Sandy Pentland. "Predicting personality using novel mobile phone-based metrics", *Social Computing, Behavioral-Cultural Modeling and Prediction*. Springer, 2013, pp. 48–55.

[264] Pål Sundsøy. Can mobile usage predict illiteracy in a developing country?, (July 2016). arXiv: 1607.01337 [cs.AI].

[265] Daniel Bjorkegren and Darrell Grissen. Behavior Revealed in Mobile Phone Usage Predicts Loan Repayment, *Available at SSRN 2611775* (2015).

[266] Pål Sundsøy, Johannes Bjelland, Asif M Iqbal, Yves-Alexandre de Montjoye, et al. "Big Data-Driven Marketing: How machine learning outperforms marketers' gut-feeling", *Social Computing, Behavioral-Cultural Modeling and Prediction*. Springer, 2014, pp. 367–374.

[267] Andrey Bogomolov, Bruno Lepri, Jacopo Staiano, Nuria Oliver, Fabio Pianesi, and Alex Pentland. "Once Upon a Crime: Towards Crime Prediction from Demographics and Mobile Data", *Proceedings of the 16th International Conference on Multimodal Interaction*. ICMI '14. Istanbul, Turkey: ACM, 2014, pp. 427–434.

[268] Robin Wilson, Elisabeth Zu Erbach-Schoenberg, Maximilian Albert, Daniel Power, Simon Tudge, Miguel Gonzalez, Sam Guthrie, Heather Chamberlain, Christopher Brooks, Christopher Hughes, Lenka Pitonakova, Caroline Buckee, Xin Lu, Erik Wetter, Andrew Tatem, and Linus Bengtsson. Rapid and Near Real-Time Assessments of Population Displacement Using Mobile Phone Data Following Disasters: The 2015 Nepal Earthquake, *PLoS Curr.* **8** (Feb. 2016).

[269] Vincent D Blondel, Adeline Decuyper, and Gautier Krings. A survey of results on mobile phone datasets analysis, *EPJ Data Science* **4**, n. 1 (2015), p. 1.