

HARVARD
Kenneth C. Griffin



**GRADUATE SCHOOL
OF ARTS AND SCIENCES**

DISSERTATION ACCEPTANCE CERTIFICATE

The undersigned, appointed by the

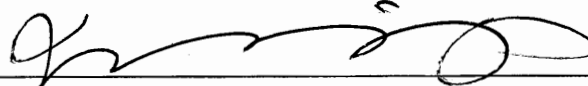
Department of Statistics

have examined a dissertation entitled

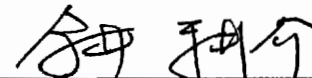
Topics in Privacy, Data Privacy and
Differential Privacy

presented by James Bailie

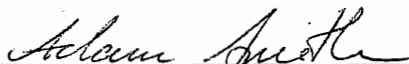
candidate for the degree of Doctor of Philosophy and hereby
certify that it is worthy of acceptance.

Signature 

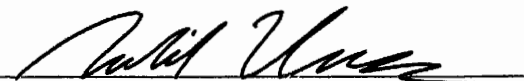
Typed name: Prof. Xiao-Li Meng

Signature 

Typed name: Prof. Kosuke Imai

Signature 

Typed name: Prof. Adam Smith

Signature 

Typed name: Prof. Salil Vadhan

Date: April 28, 2025

Topics in Privacy, Data Privacy and Differential Privacy

A DISSERTATION PRESENTED

BY

JAMES BAILIE

TO

THE DEPARTMENT OF STATISTICS

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

IN THE SUBJECT OF

STATISTICS

HARVARD UNIVERSITY

CAMBRIDGE, MASSACHUSETTS

APRIL 2025

© 2025 – JAMES BAILIE

THIS WORK IS LICENSED UNDER THE CREATIVE COMMONS ATTRIBUTION (CC-BY) 4.0 LICENSE.

Topics in Privacy, Data Privacy and Differential Privacy

ABSTRACT

In an era of unprecedented data availability and analytic capacity, the protection of individuals' privacy in statistical data releases is becoming an increasingly difficult problem. This dissertation contributes to the theoretical and methodological foundations of statistical data privacy, largely focusing on differential privacy (DP). We begin with a multifaceted investigation into privacy from legal, economic, social, and philosophical standpoints, before turning to a formal system of DP specifications built around five core building blocks found throughout the literature: the domain, multiverse, input premetric, output premetric, and protection loss budget. This system is applied to statistical disclosure control (SDC) mechanisms used in the US Decennial Census, analyzing both the traditional method of data swapping and the contemporary TopDown Algorithm. Beyond these case studies, this dissertation explores the inferential limitations posed by DP and Pufferfish privacy in both frequentist and Bayesian settings, establishing general bounds under mild assumptions. It further addresses the challenges of applying DP to complex survey pipelines, incorporating issues such as sampling, weighting, and imputation. Finally, it contextualizes DP within broader frameworks of data privacy, namely the Five Safes and contextual integrity, advocating for a more integrated approach to privacy that respects statistical utility, transparency, and societal norms.

Contents

TITLE PAGE	i
COPYRIGHT	ii
ABSTRACT	iii
TABLE OF CONTENTS	iv
LIST OF FIGURES	vii
LIST OF TABLES	viii
ACKNOWLEDGMENTS	x
o INTRODUCTION	I
I PRIVACY VIEWPOINTS	6
1.1 The Legal Understanding of Privacy	6
1.2 The Economics of Privacy	14
1.3 Privacy From the Social Sciences Perspective	25
1.4 Privacy Under a Philosophical Lens	40
I DIFFERENTIAL PRIVACY IN THE US CENSUS	43
2 FIVE BUILDING BLOCKS OF DIFFERENTIAL PRIVACY	45
2.1 Motivation: Why Do We Need To Identify Building Blocks for DP?	45
2.2 Contributions: Five Building Blocks of DP	47
2.3 An Etymological Account of DP	50
2.4 A System of DP Specifications	59
2.5 Post-Processing and Composition	90

3	INVARIANT-PRESERVING DEPLOYMENTS OF DIFFERENTIAL PRIVACY FOR THE US DECENNIAL CENSUS	94
3.1	Motivations and Contributions	94
3.2	A DP Analysis of Data Swapping	100
3.3	A DP Analysis of the TopDown Algorithm	113
3.4	Comparisons between the PSA and the 2020 DAS	117
3.5	Discussion	124
4	CAN SWAPPING BE DIFFERENTIALLY PRIVATE?	132
4.1	What Motivated This Stirred-Not-Shaken Trio?	132
4.2	Highlights of Part I: Five Building Blocks of DP	135
4.3	How to Reduce ‘Privacy Loss’ Without Adding More Noise: A Perverse Guide	142
4.4	Highlights of Part II: The US Census’s Evolving Data Protection	147
4.5	What Does It Mean If Swapping Is Differentially Private?	154
4.6	Invariants, Transparency and Data Utility	159
II	STATISTICAL INFERENCE UNDER PRIVACY CONSTRAINTS	168
5	GENERAL INFERENTIAL LIMITS UNDER DIFFERENTIAL AND PUFFERFISH PRIVACY	170
5.1	Introduction	170
5.2	Pure ϵ -Differential Privacy	174
5.3	Pure ϵ -Differential Privacy as an Interval of Measures	179
5.4	Bounds on the Privatised Data Probability	187
5.5	Frequentist Privacy-Protected Inference	193
5.6	Bayesian Privacy-Protected Inference	196
5.7	Pufferfish Privacy	204
5.8	An IP View of Pufferfish Privacy	209
5.9	Optimality of This Paper’s Results	216
5.10	Discussion	218
III	DIFFERENTIAL PRIVACY IN THE SURVEY CONTEXT	224
6	WHOSE DATA IS IT ANYWAY? TOWARDS A FORMAL TREATMENT OF DIFFERENTIAL PRIVACY FOR SURVEYS	226
6.1	Introduction	226
6.2	Background	230
6.3	DP Flavors for Survey Statistics	242
6.4	Utility Considerations	246
6.5	Privacy Considerations	255
6.6	Discussion	264
7	THE COMPLEXITIES OF DIFFERENTIAL PRIVACY FOR SURVEY DATA	266
7.1	Introduction	266

7.2	DP and the Multistage Process of Data Production	269
7.3	DP with Complex Sampling Designs	277
7.4	DP for Survey Weighted Estimates	278
7.5	DP and Weighting Adjustments	281
7.6	DP and Imputation	282
7.7	Discussion	284
IV BROADER PERSPECTIVES ON STATISTICAL DATA PRIVACY		286
8	THE FIVE SAFES AS A PRIVACY CONTEXT	288
8.1	Introduction	288
8.2	The Five Safes and the Information Flows They Govern	290
8.3	The Five Safes as a Privacy Context for Statistical Dissemination	294
8.4	Differential Privacy in the Context of the Five Safes	296
8.5	Ongoing Inquiries	299
APPENDICES		301
APPENDIX A APPENDICES TO CHAPTER 2		303
A.1	The Measure-Theoretic Definition of a Data-Release Mechanism T	303
A.2	What Can We Say About the Budget?	306
A.3	Connections Between the Input Premetric $d_{\mathcal{X}}$ and the Multiverse \mathcal{D}	307
A.4	Common Choices for the Input Premetric $d_{\mathcal{X}}$	310
A.5	The Post-Processing and Composition Mechanisms	316
A.6	Blackwell's Theorem and Post-Processing	318
A.7	Proofs	320
APPENDIX B APPENDICES TO CHAPTER 3		328
B.1	Background on Data Swapping	328
B.2	Other Related Work	330
B.3	Proof of Theorem 3.2.4	331
B.4	Optimality of Theorem 3.2.4	341
B.5	Proof and Discussion of Theorem 3.3.1	352
B.6	The 2010 US Census Disclosure Avoidance System	353
APPENDIX C APPENDICES TO CHAPTER 5		362
C.1	Definition of $\text{supp}(x \mid t, \theta)$	362
C.2	The Density Ratio Metric Is Well-Defined	363
C.3	Metric Spaces	365
C.4	Distorting Functions and Distortion Models	370
C.5	Supplementary Results	376
C.6	Proofs Omitted From the Main Text	398

APPENDIX D APPENDICES TO CHAPTER 6	410
D.1 Proofs	410
REFERENCES	413

List of Figures

3.1	Mean absolute percentage error in the two-way tabulation of dwelling ownership by county	112
3.2	Conversion between the nominal privacy loss budget (ϵ) and the swap rate (p) for the PSA	126
4.1	Schematic of a differential privacy specification	141
5.1	An illustration of the Laplace mechanism	178
5.2	Upper and lower bounds for the density $p(t \mid \theta)$ of the privatised binary sum	192
5.3	Upper and lower density bounds for $p(t \mid \theta)$ under randomised response	193
5.4	Density bounds for the posterior $p(\theta \mid t)$ from a privatised single count	203
6.1	The total privacy loss over two mechanisms T'_1 and T'_2 which share the same sampling step	261
7.1	Schematic of the survey data pipeline	270
7.2	Examples of where to start the data-release mechanism in the survey pipeline and which of the previous stages to take as invariant	272

List of Tables

2.1	Notation related to DP flavors and specifications	63
2.2	Notation related to data-release mechanisms	64
2.3	Notation related to the five building blocks of a DP specification	65
3.1	A comparison of two-way tabulations of dwelling ownership by county	110
3.2	Conversion of (expected) swap rate p to privacy loss ε	111
3.3	The privacy loss budget of the 2020 TDA	115
3.4	DP specifications for the 2020 US Census	118
3.5	The total nominal privacy loss ε for the PSA applied to the 2020 Decennial Census	121
6.1	Overview of the possible settings of DP in the survey pipeline	246
6.2	Overview of invariance’s implications on survey design weights	251
6.3	Overview of different DP setting’s implications on privacy amplification	252
8.1	The five contextual integrity parameters and their meanings in statistical dissemination, with reference to the Five Safes	295

Acknowledgments

IF IT TAKES A VILLAGE to raise a child, the same must also be true of raising a PhD thesis. In fact, unless a single village can span (at least) five continents, multiple must have raised the present thesis. Since any census is bound to have errors, be they in enumeration or for privacy protection, I will not attempt one for these villages – especially as an undercount in this situation is both permanent and costly.

Nevertheless, some general words of thanks are in order. Lest the previous paragraph be taken to imply the contrary, I want to start by emphasizing my heartfelt gratitude to the many, many individuals who helped me along this journey. I hope that, even though it does not enumerate them, the following will still justly honor all these individuals and their invaluable contributions.

Thank you to the staff at the Fulbright Program, the Institute of International Education and Fulbright Australia – and to my fellow Fulbright cohort – for your support, particularly in the early days of the program as we navigated COVID-19, lockdowns, remote study, travel and visas. I am also grateful to the Australian-American Fulbright Commission and the Kinghorn Foundation for their generous financial support throughout my PhD. Before the seeds of my PhD were even sown, my intelligent and kind colleagues in the Graduate Program and the Methodology Division of the Australian Bureau of Statistics (ABS) initiated my training as a survey statistician, helped me write my first academic papers and introduced me to the problem explored in this thesis. For this – and many other reasons – I am greatly indebted to them. After departing the ABS, the Australian National University’s Mathematical Sciences Institute generously hosted me and provided office space in the first year of my PhD while I was studying remotely and attending classes on Zoom in the middle of the night. A big thanks to the staff, faculty and students there, particularly my officemates who provided much needed support and companionship in those uncertain days.

The staff, faculty, postdocs, undergraduates and fellow PhD students at the Harvard University’s Department of Statistics have been equally supportive. I will cherish the years working, teaching and learning alongside all of you. Thank you for welcoming me to a new country; for helping me navigate a new place and a new chapter in my life; and – most of all – for the precious friendships we have formed and which have sustained me through the PhD. The same sentiments also apply to my housemates and dormmates, both in Australia and the USA, with whom I have been lucky to share many of life’s moments with during my PhD. I have also been fortunate to go on various research (and non-research) trips across Europe, Aus-

tralia, Asia and the Americas. My heartfelt thanks to all those who hosted (or accompanied) me on these trips, a number of whom generously welcomed me into their homes.

During these travels, as well as at the many conferences I have been able to attend, numerous colleagues have dispensed stimulating insights and feedback. It is no understatement to say that this thesis has greatly improved as a result of those conversations. Although any direct evidence is absent from this dissertation, my PhD experience has also been enriched by active and ongoing collaboration with colleagues in the AI and Global Development Lab. Furthermore, I have been lucky enough to have many other mentors and collaborators during my PhD, from whom I have learned so much. Indeed, while there is still so much more for me to know about how to be a productive academic, they have collectively taught me (among many other things) how to go from a vague hunch or spark of an idea into a concrete research direction, and then how to nurture and grow that research direction into a publishable paper. More generally, my teachers, of all shapes and sizes, both during and before my PhD – from my mum quizzing me on my times tables as she drove me to primary school, to the professors of my grad school classes – they have made this dissertation possible. Thank you.

Many thanks are also owed to my committee members, Salil Vadhan, Adam Smith and Kosuke Imai, for their service. I am appreciative of the time and energy you have spent reading my dissertation, engaging with my work, participating in my defense and asking questions – all with the genuine intention of improving my research. These remarks also apply to my advisor, Xiao-Li Meng, although they do not nearly capture your contributions to my PhD journey: From the time I started at Harvard, you have been boundlessly creative and steadfastly supportive of my research. Your optimism has encouraged me in the face of difficulty, disillusionment and futility. I am grateful and honored to have been advised by you.

To my family and friends, it is a truism, yet it cannot go without saying: I would not be who I am without you. In many ways, a PhD can be a solitary experience, but you have made it not so. Despite being spread across the world – with the inherent difficulties that entails – your encouragement, companionship and love have shown you always have my back and will always be there for me. The fun (whether type I, II or III) we have had together over the years has carried me through this PhD.

Finally, I want to acknowledge that ‘pale blue dot’ – our miraculous, fragile, singular home – whose beauty and awe inspires a deep sense of peace, contentment and humility; whose life-giving force sustains us all; and whose preservation needs our care and respect now more than ever.

This page intentionally left blank.



Introduction

IN THE CURRENT INFORMATION ERA, society is experiencing a paradigm shift in how data is generated, collected and used. The world today is witnessing an explosive growth of large-scale datasets containing personal information. Demographic and economic surveys, biomedical studies and massive online service platforms facilitate understanding of human biological functions and socio-behavioral environments.

At the same time, the ability to attack statistical outputs to reveal confidential information has never been higher. Today's attackers have at their fingertips an unprecedented level of computational power and access to data. Furthermore, with the proliferation of personal data available online, the mosaic effect – which describes the potential for privacy breaches by integrating many small pieces of innocuous data – is increasing the privacy risk of data publications.

This dissertation contributes to the burgeoning literature at the intersection of privacy and statistics. We focus on statistical perspectives of privacy, data privacy and differential privacy (DP). These topics have received extensive treatment across a range of fields. In particular, DP and associated concepts in the

formal privacy literature originated from – and have been extensively developed by – the computer science community. Our overall contribution is thus to build a statistical understanding of DP. Whereas the existing DP literature has focused on protecting individual data points, the statistical perspective is concerned with the underlying distributions for which the data contain sufficient variations to reveal partially.

In the first chapter, we survey the diverse understandings of privacy from the perspectives of law, economics, the social sciences and philosophy. We trace the legal notion of privacy through the 20th century back to the 1890 Harvard Law Review article “The Right to Privacy”, separating the two strands of decisional privacy and data privacy (also termed informational privacy). While technological advancements served as the catalyst, the notion of privacy arose naturally and inevitably during the 19th century following the social upheaval of the industrial revolution, during which Western nations shifted away from agrarian economies into societies with large, economically independent bourgeoisies. Herbert Simon’s characterization of today’s information economy as the consumption of attention is contrasted with an alternative interpretation of information as the consumption of privacy. Finally, we outline the philosophical contradictions of privacy: privacy as simultaneously a virtue and a vice; as a fundamental right versus as a derivative of more basic rights; and as a set of interrelated concepts, which can only be properly understood and constructed in context.

In the second chapter, we present a framework of *DP specifications*. Mathematically speaking, a DP specification is a Lipschitz condition on data-release mechanisms – functions that transform confidential input data into noise-injected output statistics. Thus, the core philosophy of DP is to manage disclosure risk by limiting the rate of change of variations in the output statistics as the input data are (counterfactually) altered. DP conceives of privacy protection specifically as control over the Lipschitz constant – i.e.

over this rate of change – and different DP specifications correspond to different choices of where and how to measure input alterations and output variations, in addition to the choice of how much to control this rate of variations-to-alterations. Following this line of thinking through existing literature leads to the five necessary building blocks of a DP specification. They are, in order of mathematical prerequisite, the protection domain, the scope of protection, the protection unit, the standard of protection, and the intensity of protection. In simple terms, these are respectively the “who,” “where,” “what,” “how” and “how much” questions of DP. This framework unveils the nuances and pitfalls in employing DP as a theoretical yardstick for statistical disclosure control (SDC), and provides the requisite mathematical language to faithfully extend the essence of DP to SDC procedures which were derived without consideration of DP.

Through the lens of the system developed in the previous chapter, the third and fourth chapters examine two SDC methods for the United States Decennial Census: the Permutation Swapping Algorithm (PSA), which is similar to the 2010 Census’s disclosure avoidance system (DAS), and the TopDown Algorithm (TDA), which was used in the 2020 DAS. To varying degrees, both methods leave unaltered some statistics of the confidential data – which are called the method’s invariants – and hence neither can be readily reconciled with DP, at least as it was originally conceived. Nevertheless, we establish that the PSA satisfies ϵ -DP subject to the invariants it necessarily induces, thereby showing that this traditional SDC method can in fact be understood within our more-general system of DP specifications. By a similar modification to ρ -zero concentrated DP, we also provide a DP specification for the TDA. Finally, as a point of comparison, we consider the counterfactual scenario in which the PSA was adopted for the 2020 Census, resulting in a reduction in the nominal privacy loss, but at the cost of releasing many more invariants. Therefore, while our results explicate the mathematical guarantees of SDC provided by the PSA, the TDA

and the 2020 DAS in general, care must be taken in their translation to actual privacy protection – just as is the case for any DP deployment.

Chapter Five concerns two important flavors of DP that are related yet conceptually distinct: pure ϵ -DP and Pufferfish privacy. We restate these flavors in terms of an object from the imprecise probability literature: the interval of measures. We use this reformulations to derive limits on key quantities in frequentist hypothesis testing and in Bayesian inference using data that are sanitised according to either of these two privacy standards. Under very mild conditions, the results in this work are valid for arbitrary parameters, priors and data generating models. These bounds are weaker than those attainable when analysing specific data generating models or data-release mechanisms. However, they provide generally applicable limits on the ability to learn from differentially private data – even when the analyst’s knowledge of the model or mechanism is limited. They also shed light on the semantic interpretations of the two DP flavors under examination, a subject of contention in the current literature.

Chapters Six and Seven turn to the question of implementing DP in the context of a statistical survey. Statistical agencies are increasingly considering DP to help manage the disclosure risk associated with survey data releases. Yet standard DP theory does not address how disclosure risk may be impacted by data collection and preprocessing procedures, which for survey data include sampling, non-response, weighting, imputation and the use of auxiliary data. To rectify this limitation and pave the way for its effective implementation, Chapter Six provides a formal treatment of DP in the context of the survey data pipeline. By reasoning about how DP should interact with survey data collection and preprocessing, this theory sheds new light on existing discussions – such as privacy amplification by sampling and the sensitivity of weighted estimators – and identifies new challenges – such as DP’s underestimation of disclosure risk under some

traditional statistical disclosure control attacker models. Chapter Seven takes a broader view, exploring the possibilities and limitations of DP for survey data. Specifically, we identify five aspects that need to be considered when adopting DP in the survey context: the multi-staged nature of data production; the limited privacy amplification from complex sampling designs; the implications of survey-weighted estimates; the weighting adjustments for nonresponse and other data deficiencies, and the imputation of missing values. We summarize key findings from the literature with respect to each of these aspects and also discuss some of the challenges that still need to be addressed before DP could become the new data protection standard at statistical agencies.

In the final chapter, we close with a broader perspective, examining the relationship between the Five Safes and contextual integrity as framing devices for DP. The Five Safes is a system used by national statistical offices (NSO) for assessing and managing the disclosure risk of data sharing. This chapter makes two main points: Firstly, the Five Safes can be understood as a specialization of the broader concept of contextual integrity, adapted to the situation of statistical dissemination by an NSO. We demonstrate this by mapping the five parameters of contextual integrity onto the five dimensions of the Five Safes. Secondly, the Five Safes contextualizes narrow, technical notions of privacy within a holistic risk assessment. We demonstrate this with the example of differential privacy (DP). This contextualization allows NSOs to place DP within their Five Safes toolkit while also guiding the design of DP implementations within the broader privacy context, as delineated by both their regulation and the relevant social norms.

1

Privacy Viewpoints

1.1 THE LEGAL UNDERSTANDING OF PRIVACY

ALTHOUGH PRIVACY CONCERNS HAD APPEARED EARLIER in English and American common law, the 1890 Harvard Law Review article “The Right to Privacy” (Warren and Brandeis, 1890) is generally attributed as initiating the legal study of privacy. Warren and Brandeis were largely motivated by contemporary technological developments – particularly the invention of the Kodak film camera and the increased circulation of sensationalist newspapers – which they viewed as invaders of “the sacred precincts of private and domestic life”. While they acknowledged that other legal claims had already been used in English and American courts to defend privacy, they argued for a new right-to-privacy law, “a right to be left alone”. And yet – despite being the “most influential law review article ever” (Kalven, 1966; Shapiro and Pearse, 2012) – many of the problems Warren and Brandeis raised remain unresolved today.

*Informational privacy*¹ – the strand of privacy law focussing on control over one’s personal information – has continued to evolve since 1890 by and large as a reactionary endeavour. Novel technologies throughout the twentieth century and into the present day have enabled increased surveillance, information-gathering, electronic storage and computing power. Modernising both common and statutory law has, and continues to be, necessary to protect against these privacy-invasive technologies. As two examples, wiretapping was criminalised by the US Congress in 1934 after significant criticism of its abuse (Solove and Schwartz, 2021, p.247, p.294)² and the EU adopted two broad-reaching privacy regulations in the last 30 years – the 1995 Data Protection Directive and the 2016 General Data Protection Regulation (GDPR).

A second strand of privacy law, *decisional privacy*, concerns an individual’s autonomy in making personal decisions, particularly regarding their body or actions within their home. This strand is deeply connected with rights of liberty; in fact, there is disagreement as to whether decisional privacy is a genuine issue of privacy, rather than solely a question of freedom (DeCew, 2018). In the United States, this strand is also known as the “constitutional right to privacy”,³ established by the Supreme Court in *Griswold v. Connecticut* (381 U.S. 479) and most famously applied in *Roe v. Wade* (410 U.S. 113).

Taken together, informational and decisional privacy cover the major subject matters of privacy law, although there are others.⁴ The remainder of this article will concentrate on informational privacy law.

¹In the EU and much of the rest of the world, the term *data protection* is used instead of informational privacy.

²While this law prohibited the use of wiretapping as evidence in court, it did not limit government wiretapping, which became more pervasive through World War II and the Cold War until 1968 when new legal restrictions were passed by the U.S. Congress (Solove and Schwartz, 2021, p.296).

³While the constitution does not mention the right of privacy, the Supreme Court in *Griswold v. Connecticut* (381 U.S. 479) conclude that a number of amendments, taken together, created the right to autonomy on certain decisions relating to an individual’s body and private life.

⁴See (Allen and Rotenberg, 2016, p.5) (which in turn is summarising (Allen, 2007)) for other legal senses of privacy: physical privacy (the right to solitude and seclusion in one’s own home or property; freedom from peeping Toms, search and seizures, surveillance in one’s home and trespass), proprietary privacy (the right to control the use

Issues in this subfield can be classified into three broad categories: information collection; storage and processing; and dissemination (Solove, 2008).

Privacy issues arising from information collection can be further understood in terms of surveillance – the covert or overt observation of an individual; and interrogation – the probing or questioning for information. Information storage and processing pose their own privacy pitfalls, particularly relating to the physical security and proper deletion of data; integration with other data; identification of individuals in the data; additional use of data beyond the originally stated intentions; and consent from the data subjects to store and process their data. Finally, there are privacy torts relating to information dissemination: the disclosure of confidential information, or simply the increased accessibility of such information; identity theft; defamation; and blackmail – the threat of potential information dissemination.

There has been limited legal study on the dissemination of potentially sensitive information when direct identifiers such as names, addresses and social security numbers, have been removed. In the commercial setting, this is a relatively new concern; only with the recent advent of the ‘information age’ have corporations been able to collect, store and process large amounts of data. And only with the new trend of machine learning in business has it become apparent that such information can still be valuable without direct identifiers.

In contrast, national statistical organisations (NSOs) – and more generally, government agencies – have collected personal information and published de-identified statistics for over a century; and for almost as

of one’s name, likeness, voice or other personal identifiers, called the right to publicity; protection from identity theft or appropriation), associational privacy (freedom of association and (private) assembly), intellectual privacy (‘the right to be let alone’ in one’s thoughts; freedom of thought; the protection of mental repose). Given their multitude and their creeping boundaries, torts of privacy have not escaped criticism, particularly when they overlap with other established legal claims such as defamation, trespass, and infliction of emotional distress (Prosser, 1960; Allen, 2010).

long, legal theory has recognised the privacy risks of such endeavours. Rather than attempting to legislate the nebulous concept of privacy, modern law governing NSOs in the West focuses on identity or identification.⁵ The US, UK, EU and Australia legislate against the dissemination of statistics when – possibly in combination with other information – either A) individuals’ *identities* can be deduced from the statistics, or B) individuals can be *identified* in the statistics. (See the third section of this article for a more detailed discussion of official statistics legislation.)

Due in part to a lack of court opinions testing these legislations,⁶ the legal meanings of ‘identity’ and ‘identify’ in this context are unclear (Finck and Pallas, 2020); it is even unclear whether A) and B) are legally equivalent.⁷ While the legal study of de-identified information dissemination is minimal, it is a topical issue and we can expect to see more attention from legal scholars in the coming years – in the meantime we will have to rely on other fields for guidance.

Legal scholars have, in contrast, long understood that studying the limitations of the right to privacy

⁵I assume the justification for this approach is based on the proposition that someone’s privacy cannot be compromised if they cannot first be identified in the data. Is this proposition valid? Certainly many computer scientists would think not.

⁶It is not surprising that NSOs take a conservative approach to privacy and are unwilling to test their legislation. An NSO’s social licence is critical to its functioning and thus any bad press, including court proceedings, will weaken its operations (Brick and Williams, 2013).

⁷By comparison, in the computer science literature, to identify an individual X means to determine some of X ’s attributes and show that X is the only unit in the data with these attributes (Sweeney, 2000; Narayanan and Shmatikov, 2008). Given this definition, is X ’s identity determined by these attributes? Certainly, the person on the street would not consider a potpourri of putative incidentals – even when they are unique – to define their identity. But then, what does defines X ’s identity?

While the legal study of identity and identification as it pertains to official statistics is limited, one can look more broadly in the law for guidance. For example, the Video Privacy Protection Act, 18 U.S.C §2710 prohibits the disclosure of “personally identifiable information” without written consent. This law has been tested in court (e.g. *Harris V Blockbuster*, covered in (Allen and Rotenberg, 2016, p.992), and *In re Netflix Privacy Litigation* (N.D. Cal. Mar. 18, 2013), covered in (Allen and Rotenberg, 2016, p.996)). In (Culnane et al., 2019), section 1.1 provides examples of purported re-identification of published de-identified data and section 6.3 lists cases where re-identification might be possible in non-open shared data; I should examine whether these cases were litigated.

is essential to a coherent theory.⁸ It is not sufficient to only know when a person is entitled to privacy; the question of when they are *not* entitled to privacy is equally important. In many circumstances, the right to privacy is outweighed by other rights; in other circumstances, privacy is simply too impractical to guarantee. For example, ‘the test of newsworthiness’ is used by many courts to determine whether some publication of private information was acceptable.⁹ Further, it is generally accepted that persons in the public sphere, such as politicians and celebrities, are entitled to less privacy. Finally, laws enabling the state, in certain circumstances, to arrest a citizen, listen in on their communications or search their property and body, are essentially limitations on privacy. Defining these limitations raises the important tradeoff between privacy on the one hand and security, criminal justice and public interest on the other.

Following the September 11 terrorist attacks, the legal give-and-take swung towards security, with new laws increasing surveillance (the USA PATRIOT Act) and privacy intrusions at airports (the Aviation and Transport Security Act). In recent years, a series of scandals – Snowden’s leaks of NSA surveillance, Cambridge Analytica political use of Facebook data and numerous commercial data breaches – have pushed the focus back towards privacy and promoted new types of legal protections. In particular, the GDPR established in EU the ‘right to be forgotten’, which allows consumers to request a company delete their personal data. The GDPR also requires data protection ‘by default’ – the strictest privacy settings must be the default – and ‘by design’ – privacy must be built-in to any operation from the start.¹⁰

⁸In fact, Warren and Brandeis in their original 1890 article ([Warren and Brandeis, 1890](#)) suggested six broad limitations: matters which are “of public or general interest” can be published; the right to privacy does not prohibit “privileged communications”; the invasion of privacy by oral communication should generally not be afforded redress; the right to privacy is forfeited when a person consents to publication of his personal information; and “the truth of the matter” or “the absence of ‘malice’ does not afford a defence”.

⁹See, for example, *Sipple V Chronicle Publishing* (Cal. Ct. App. 1984) in ([Allen and Rotenberg, 2016](#), p. 153).

¹⁰In comparison, the right to be forgotten has not been legislated in the US although it has limited recognition in the common law – see *Melvin v. Reid* (1931), *Sidis v. FR Publishing Corp.* (1940) and other cases in ([Allen and](#)

I.I.I SUMMARY OF NATIONAL STATISTICAL ORGANISATIONS' (NSO'S) PRIVACY REGULATIONS

1. The United States of America: Responses to censuses and surveys administered by the US Census Bureau (USCB) are protected under Section 9, Title 13 of the US Code (USC).¹¹ Responses can only be used for statistical purposes and any publication cannot release identifying information.¹² A good summary of this law and a history of privacy at the USCB is given in https://www.census.gov/history/www/reference/privacy_confidentiality/.¹³ Other federal statistical agencies¹⁴ are regulated by the Confidential Information Protection and Statistical Efficiency Act of 2002 (CIPSEA) (*National Academies of Sciences, Engineering, and Medicine*,

Rotenberg, 2016, section 1.E.2).

In general, the EU has a history of stronger privacy protection in business and commercial settings, while the US has concentrated more on protection from government intrusion. As one explanation for this, many in the U.S. see privacy protection as inefficient for market activity (Allen and Rotenberg, 2016, p. 11), while those in the EU value privacy as a requirement for individuals to engage in online commerce (DeCew, 2018). As such the EU has adopted an omnibus approach to data protection where a single, comprehensive privacy regulation applies across industry, government and business sectors (Allen and Rotenberg, 2016, p. 758). Under the US's predominately laissez faire approach, legislators have only acted on specific issues, resulting in a patchwork of sectorial regulations – for example the Video Privacy Protection Act (1988), the Employee Polygraph Protection Act (1988) and the Driver's Privacy Protection Act (1994).

¹¹<http://uscode.house.gov/browse/&edition=prelim>

¹²More specifically, “[the Department of Commerce may not:] (1) use the information [collected by a census or survey] for any purpose other than the statistical purposes for which it is supplied; or (2) make any publication whereby the data furnished by any particular establishment or individual ... can be identified; or (3) permit anyone other than the [officers of the Department] to examine the individual [records].”

Furthermore, the federal government are prohibited from obtaining copies of individual records (except in conducting a census or survey) and these records are inadmissible in court.

Under Section 2104, Title 44, USC, individual responses to the decennial Population Census are released to the public by the National Archives after 72 years (https://www.census.gov/history/www/genealogy/decennial_census_records/the_72_year_rule_1.html). (Section 2104, Title 44 supersedes Section 9, Title 13.)

¹³Title 13 also requires that the US Census Bureau release some statistics exactly (i.e. without any privacy protections applied). The computer science literature understands that any release of statistics results in some loss of privacy (Kifer and Machanavajjhala, 2011; Dinur and Nissim, 2003). Does this imply that Title 13's requirements for anonymity and exact statistics are contradictory? Perhaps, but not necessarily – anonymity is not the same as the ‘man-on-the-street’s definition of privacy; and the ‘man-on-the-street’s privacy is not the same as the computer scientist’s definition of privacy.

¹⁴There are approximately 125 federal agencies in charge of statistical activities, thirteen of whom – the Bureau of Economic Analysis; Bureau of Justice Statistics; Bureau of Labor Statistics; Bureau of Transportation Statistics; Census Bureau; Economic Research Service; Energy Information Administration; National Agricultural Statistics Service; National Center for Education Statistics; National Center for Health Statistics; National Center for Science and Engineering Statistics; Office of Research, Evaluation and Statistics in the Social Security Administration; and Statistics of Income Division in the Internal Revenue Service – have a primarily statistical function (*National Academies of Sciences, Engineering, and Medicine*, 2017a).

2017b). A summary of their legislative framework is found in (National Academies of Sciences, Engineering, and Medicine, 2017b, section 3). See also (National Academies of Sciences, Engineering, and Medicine, 2021).

2. The European Union: NSOs are bound by EU official statistics regulation¹⁵ as well as the General Data Protection Regulation (GDPR).¹⁶ In most cases, NSOs are also governed by their country's statistics act.¹⁷ Under EU regulations, any data which "allow statistical units to be identified [and] thereby disclos[e] individual information" cannot be disseminated except with approval of the European Parliament or consent of the respondent.¹⁸ The GDPR makes provisions for control of one's own personal data (and other rights of the 'data subject'); it restricts how data must be stored and processed; but it does not create obligations with regard to the dissemination of official statistics.
3. The United Kingdom: The GDPR remains British law as implemented by the Data Protection Act 2018, although this may change (Lillington, 2021). The confidentiality of personal information collected by the Office of National Statistics (ONS) is also protected by Section 39 of the Statistics and Registration Service Act (SRSA) 2007.¹⁹ Personal information cannot be released by an agent

¹⁵The EU regulation for NSOs is 'Regulation No 223/2009 of the European Parliament and of the Council on European statistics' (<https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32009R0223>)

¹⁶<https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:02016R0679-20160504>. See in particular, Article 4(1) and Recital 26 for GDPR's definition of personal data as that which is reasonably likely to be linked to an identified natural person.

¹⁷For example, Statistics Norway is governed by 'the Act relating to official statistics and Statistics Norway' https://www.ssb.no/en/omssb/lover-og-prinsipper/lover-og-prinsipper/_attachment/402255?_ts=176ad6e1f20 and associated regulations https://www.ssb.no/en/omssb/lover-og-prinsipper/lover-og-prinsipper/_attachment/448738?_ts=17834d9bc20; INSEE (France) is governed by Act no. 51-711 of 7 June 1951 (<https://www.insee.fr/en/information/2398930>); and INE (Spain) is governed by Act no. 12/1989 of 9 May (<https://www.ine.es/dyngs/AYU/index.htm?cid=130> and <https://www.boe.es/boe/dias/1989/05/11/pdfs/A14026-14035.pdf>).

¹⁸Confidentiality is protected under Chapter V of Regulation No 223/2009.

This regulation defines 'confidential data' to mean "data which allow [a person, household or corporation] to be identified, either directly [from their names or addresses] or indirectly [by any relevant means that might reasonably be used by a third party], *thereby* disclosing individual information" (emphasis added). Confidential data cannot be lawfully disclosed except with approval of the European Parliament or consent of the respondent.

"The use of confidential data [collected by a National Statistical Institute (NSI)] for purposes that are not exclusively statistical, such as administrative, legal or tax purposes, or for the verification against the statistical units [is] strictly prohibited".

"Access to confidential data which only allow for indirect identification of the statistical units may be granted to researchers carrying out statistical analyses for scientific purposes by the Commission (Eurostat) or by the NSIs or other national authorities, within their respective spheres of competence."

¹⁹<https://www.legislation.gov.uk/ukpga/2007/18/section/39>

of the ONS except as necessary for the ONS's functions or to a researcher for statistical purposes.²⁰ A summary of the legislation governing the ONS is given in <https://www.ons.gov.uk/about-us/transparencyandgovernance/datastrategy/relevantlegislation>.

4. Australia: Statistical information collected by the Australian Bureau of Statistics (ABS) “shall not be published or disseminated in a manner that is likely to enable the identification of a particular person or organization” (Subsection 12(2), Census and Statistics Act 1905).²¹
5. Canada: The Statistics Act states that “no person who has been sworn under section 6 shall disclose or knowingly cause to be disclosed, by any means, any information obtained under this Act in such a manner that it is possible from the disclosure to relate the particulars obtained from any individual return to any identifiable individual person, business or organization.”²²

²⁰Personal, identifying information is information that relates to a particular person *and* the identity of that person can be deduced from the information taken together with any other published information.

Before receiving access to personal information, a researcher must be approved by the ONS according to criteria which must be published by the ONS from time-to-time. Any approved researcher is also bound by section 39 of the SRSA.

An individual who makes an unauthorised disclosure is guilty of a criminal offence under the SRSA, *unless* that individual *reasonably believed* the information was not identifiable.

In addition to exceptions for the ONS's functions or statistical research, disclosure of personal information is also permitted if required by any other law; if ordered by a court; to facilitate a criminal investigation; if consented by the person to whom the information relates; or if the information is already lawfully public.

Note: The body regulated by the SRSA is the UK Statistics Authority – referred to as ‘the Board’ by the Act. The ONS is the executive office of the Authority.

²¹The ABS is governed under the Australian Bureau of Statistics Act 1975, the Census and Statistics Act 1905 and their legislative instruments (e.g. the Census and Statistics Determination 2018). It is also bound by the Privacy Act 1988.

The Australian Bureau of Statistics Act 1975 establishes the ABS and sets forth its functions which primarily are “to collect, compile, analyse and disseminate statistics and related information.”

The Census and Statistics Act 1905 states that information may be disclosed by the ABS if it is not “in a manner that is likely to enable the identification of a particular person or organization” (subsection 12(2)). If the information is “not of a personal or domestic nature relating to a person”, Part 3 of the Census and Statistics Determination 2018 provides a limited set of exceptions to this rule. For example, information may be disclosed if it is already available publicly (by an official body or the organisation of which the information relates); if the information is certain trade, agricultural or construction statistics, unless a respondent shows that their identification in the statistics is likely (this is called passive confidentiality); if the respondent gives consent; or if the information is disclosed to researchers for statistical purposes (under certain provisions given in Section 15 of the Determination).

Information collected by the ABS must not be divulged in a court or to any public service agency, other than in accordance with the provisions above. Breaching these laws is a criminal offence.

A summary of the ABS's legislative framework is <https://www.abs.gov.au/websitedbs/d3310114.nsf/home/abs+legislative+framework>. The ABS Privacy Policy can be accessed at <https://www.abs.gov.au/about/legislation-and-policy/privacy/privacy-abs/abs-privacy-policy-statistical-information>.

²²Source: <https://laws-lois.justice.gc.ca/eng/acts/S-19/page-2.html>.

6. Comparative Analysis: US, EU and Australian law concern whether individuals can be identified while the UK legislates against statistics where individuals' identities can be deduced. In order to breach the regulations, the EU and the UK also require that personal information is inferred from the released statistics. In contrast, Australian and US law require simply the possibility of identification.

US regulation is the strictest amongst those studied in the sense that there is no qualifications on how re-identification is achieved; no exceptions allowing access for outside researchers; and the toughest, explicit restrictions on the use of the collected data. UK law provides the defence of "reasonable belief" – it is not an offence if the ONS reasonably believed it was not possible for an individual's identity to be deduced. Similarly, identification has to be 'reasonably likely' or 'by reasonable means' in Australia and the EU respectively. On the other hand, Title 13 of the US Code doesn't provide any such defence.

Finally, it is crucial to understand that legislation is not the only factor governing NSOs. As public bodies, NSOs need also be cognisant of the privacy norms of the society in which they operate. Their existence is predicated on a social contract and their continued functioning depends critically upon their social licence to induce sufficient response rates (Brick and Williams, 2013).

1.2 THE ECONOMICS OF PRIVACY

In 1969, the economist and Nobel laureate Herbert Simon said:

In an information-rich world, the wealth of information means a dearth of something else: a scarcity of whatever it is that information consumes. What information consumes is rather obvious: it consumes the attention of its recipients.(Simon, 1971)

With this statement, Simon explained the business model of Facebook, Twitter and YouTube more than 30 years before these companies were even conceived. Attention economics – the field which sprung up from Simon's revelation – reveals why these companies (or their parent corporations) have enormous market capitalisations: every day the scarce resource of attention is given to them, willingly and at no monetary cost, by billions of users. And yet, Simon's statement presents only half the picture; yes, information

Microdata releases are permitted by Statistics Canada policy when "(a) the release substantially enhances the analytical value of the data collected; and (b) the Agency is satisfied all reasonable steps have been taken to prevent the identification of particular survey units" (Benschop and Welch, 2024).

consumes the attention of its recipients but – just as importantly – it also consumes the *privacy of its subjects*.

If today we are living in the information economy, then one would suspect that the economics of private information play an important role. We are thus well-motivated to investigate three central questions: 1) what is private information? 2) How is it valued? And 3) how do its buyers and sellers behave in the marketplace?

1.2.1 AN ECONOMIC CONSTRUCTION OF PRIVACY AND HOW IT IS VALUED

Economists generally avoid the first question, taking the definition of informational privacy²³ as given, but the following is a simple, economic answer: Information – or data – is private if its subject would want compensation to reveal it. Further, the larger the amount of compensation required, the more the information should be considered private. In essence, *privacy is the price* of divulging information. This definition casts a deliberately wide net. It also assumes that the data subject remains the owner (or, at least, the controller) of their data; this model is not currently realistic, although we are moving towards it, as exhibited in the EU's new General Data Protection Regulations (GDPR).

To illustrate the utility of this construction of privacy, consider the following examples: a patient may reveal private information to his doctor – on the condition of confidentiality – so that he may be compensated with an expert diagnosis. On the other hand, if there was no condition of patient-doctor confidentiality, then an expert diagnosis may not be sufficient compensation. As another example, the patient may

²³There are many other types of privacy (see the legal and philosophical summaries) whose kernels require the tools of economics (in conjunction with other fields) to crack. As just one brief example, if a celebrity is harassed by paparazzi – a breach of their right to seclusion and solitude (itself a form of privacy) – how should they be compensated? While these types of issues are no doubt important, we will focus on the privacy considerations that arise when sharing data between parties.

decide not divulge his medical information to a family member since they – not being a doctor – cannot provide the same compensation.

This working definition illuminates a number of dimensions of privacy (Nissenbaum, 2010) as it relates to the different aspects of information sharing: a) the data subject; b) the sender; c) the receiver; d) the type and sensitivity of the shared information; and e) what will be done with the information (including who it may reasonably be passed on to). It also acknowledges that privacy is not a black-and-white, all-or-nothing concept; privacy lives on a scale of many greys, which we may quantify through its price. (As we shall see, the privacy scale depends not only on the five aspects above but also – in statistical situations – on the resolution level of the private inference.) Finally, our construction of privacy indicates that it can, and should, be readily traded-off for other goods – broadly any form of compensation but specifically, the social utility of the data (Abowd and Schmutte, 2019), financial transparency (Flood et al., 2013), access to services such as Facebook or medical treatments (e.g. by participating in a medical trial), lowered insurance premiums (‘customised insurance’), etc.

However, for this definition of privacy to be workable, it must be possible to assign rational prices to the multitude of information sharing scenarios. At best, any pricing would be extremely context-specific and subjective – which would limit the usefulness of the resulting economic model; at worst, it is outright impossible. Unfortunately, there is wide agreement amongst economists on the difficulties of pricing privacy in today’s economy (Acquisti et al., 2016; Lindgreen, 2018). Many of today’s data sharing scenarios exhibit asymmetries of information, power and resources between the buyer and seller. In particular, the seller is often unaware of what data they are sharing, the data’s power and value in the hands of the buyer, how the data will be used and shared beyond the buyer, and what the consequences may be if the data

ends up in the hands of a malicious actor. In short, it is near impossible to make informed decisions in today’s information marketplace (Lindgreen, 2018). Further, the individual employs a variety of heuristics and cognitive biases (see the social science summary) which frustrates the idea that the rational pricing of privacy is realistic or even relevant.

There is one school of thought pushing back against this trend. A line of research – starting with (Ghosh and Roth, 2015) and with important contributions (Hsu et al., 2014; Abowd and Schmutte, 2019) – asserts that privacy loss to data subjects can be appropriately quantified using differential privacy (DP). However, we will argue that not only does this approach inherit all of the difficulties outlined above, it also makes a type error by confusing ε – the privacy loss of the publishing *mechanism* – with the privacy loss of the *respondents*.

Under (Ghosh and Roth, 2015), the price of individual i ’s privacy is $c(v_i, \varepsilon)$,²⁴ where c is some known function (e.g. $c(v_i, \varepsilon) = v_i \varepsilon$) and v_i is i ’s “privacy value”. While there is little explication of v_i in (Ghosh and Roth, 2015), I interpret it as reflecting the five aspects a)-e) of information sharing (Nissenbaum, 2010) outlined above, while ε reflects the sixth aspect: the resolution level of any private inference. On face value it appears that estimating v_i is just as difficult as estimating the price of privacy since it requires a consideration of a)-e); and so this approach has not solved any of the difficulties outlined above.

However, (Ghosh and Roth, 2015) does provides a utility-theoretic approach to understanding $c(v_i, \varepsilon)$. For any non-negative utility function u_i , differential privacy guarantees that

$$\mathbb{E}_{x \sim M(D)}[u_i(x)] \leq e^\varepsilon \mathbb{E}_{x \sim M(D')} [u_i(x)], \quad (1.1)$$

²⁴Technically, (Ghosh and Roth, 2015) allows ε to vary with i so that the mechanism can use i ’s data more or less accurately, at a greater or lesser price $c(v_i, \varepsilon_i)$. However this intricacy will not affect any of our discussion.

where the left expectation is over the outputs x from the mechanism M , computed on the dataset D which includes i ; and the right expectation is computed on the dataset D' which excludes i . As such, the expected loss of utility from i divulging their information is bounded by $(e^\epsilon - 1)\mathbb{E}_{x \sim M(D')} [u_i(x)] \approx \epsilon \mathbb{E}_{x \sim M(D')} [u_i(x)]$. (The approximation holds for small ϵ only.) (Ghosh and Roth, 2015) thus suggests cost functions of the form $c(v_i, \epsilon) = (e^\epsilon - 1)v_i$ or $c(v_i, \epsilon) = \epsilon v_i$ where $v_i = \mathbb{E}_{x \sim M(D')} [u_i(x)]$.

The question at hand is thus whether it is possible to estimate the privacy value $\mathbb{E}_{x \sim M(D')} [u_i(x)]$. One of the primary arguments in favour of DP is that we cannot know what statistical disclosure attacks may be possible in the future. If we accept this argument, we must also admit that we cannot know how the shared data may be used in the future. Data that is innocuous today may be privacy-breaching tomorrow. As one example, thirty years ago few people could imagine that your shopping history at a single department store chain (not a pharmacy!) could predict your pregnancy status, even during the first trimester. I suspect that even today this would be surprising to many people and yet, it was possible ten years ago (Duhigg, 2012). What may be possible thirty years from now? All this is to say, predicting the future utility $\mathbb{E}[u_i(x)]$ from sharing some data today is at least as hard as – if not harder than – predicting what the next statistical disclosure attack will be. The *paradox of differential privacy* is that as soon as you accept its necessity as a future-proof protection method, you consign it to the dustbin of the unimplementable, for setting an appropriate ϵ becomes a challenge at least as hard as the problem DP solves.

Further, (Ghosh and Roth, 2015) and subsequent papers in this line of research commit a type error: they conflate the privacy loss ϵ of the sharing mechanism M with the privacy loss experienced by the individuals who share their data. As we will illustrate, ϵ cannot be associated with individuals but only with *populations*. Under (Ghosh and Roth, 2015)'s model, if an individual, say Adam, chooses not to share

his data, then he experiences no privacy loss from the sharing mechanism M . However, imagine that his identical twin did decide to share her data. It would be nonsensical to assert that Adam experiences no privacy loss, particularly in the case of genetic data. Now suppose it wasn't his twin, but his cousin, sharing their data; Adam would still experience some privacy loss – perhaps not as much as before, but still a non-zero amount. Or maybe it is his work colleagues, divulging their salaries; this would naturally leak some information about Adam's salary too! Adam experiences privacy loss, even when he chooses not to share his data.²⁵ The point here is that – when we begin working from a statistical perspective, as DP does – a sharing mechanism M impinges on the privacy of populations – not individuals – regardless of the choices of each individual to share their data or not.²⁶ The economic model of (Ghosh and Roth, 2015) which restricts privacy loss to individuals and not populations is thus fundamentally flawed.

We end this section with a word of terminological caution. As we have mentioned, the “privacy budget” ε appears to capture a notion of the resolution level of a private inference; as such it plays an important role in evaluating the privacy of a data sharing event. However when interpreting values of ε , one must be cognisant of its scale. The term ‘budget’ suggests that ε is a linear measure, like money. If you doubled your money, you could buy twice as many apples; in the same way, if the budget doubles, you would expect the cost of privacy to also double. This is not the case – ε measures privacy loss on a logarithmic scale, as evidenced by the cost function $c(v_i, \varepsilon) = (\varepsilon^{\varepsilon} - 1)v_i$ from (Ghosh and Roth, 2015). For small $\varepsilon < 1$,

²⁵It may be easier to understand the contrapositive statement: An individual that chooses not to share their data is still protected by DP.

²⁶However, if all the individuals chose not to share their data, then the output of M would be vacuous and there could not possibly be any privacy loss. That is to say, the privacy loss for an individual depends on his own actions (obviously he suffers more privacy loss from sharing his data than not, all other things being equal) as well as the collective action of his peer group.

For a comprehensive picture, one would need to model the relationships between Adam's data and that of his peer group – as in the Pufferfish framework (Kifer and Machanavajjhala, 2014) – so that we could understand the effects to Adam's privacy from his peer group sharing their data.

this contention makes little difference, but increasing ε from 12 to 19 (as the US Census Bureau did) in fact results in an increase from 162,000 to 178,000,000 on a linear scale. We know that the public do not understand logarithmic graphs (Romano et al., 2020); yet reporting ε values is like using logarithmic graphs, while also using the wrong axis labels 1, 2, 3, . . . instead of 10, 100, 1000, . . .

Usage of ε may be justified by its convenient properties (e.g. composition). In this case, we have two recommendations. Firstly, a change in terminology: ε is more accurately described as the log-budget, rather than the budget. Secondly, the choice of base makes a practical difference. Since we want the likelihood ratio $\frac{\Pr(M(D) \in S)}{\Pr(M(D') \in S)}$ to remain small unless ε is exceptionally large, using the bound 1.1^ε or 1.01^ε in place of e^ε will lead to more interpretable results: an ε of 19 could then be more accurately expressed as 199 (with base 1.1) or 1909 (with base 1.01).

1.2.2 ECONOMIC TRADEOFFS: THE BEHAVIOUR OF PRIVACY BUYERS, SELLERS AND STATE ACTORS IN THE MARKETPLACE

We have discussed an economic construction of privacy under a data-sharing regime and the various complications in its valuation. This is only one facet of the decision-making for economic agents in the information marketplace. There are many other considerations relevant to the tradeoffs made by buyers, sellers and state actors. Here we consider state actors to be agents working for the greater social good – such as regulators, government statistical offices, police and other law enforcement, or central banks – who participate in the marketplace by weighing up the social costs and benefits of privacy in an (ideally) optimal manner. We summarises these facets (largely taken from (Lindgreen, 2018)):

1. Costs of greater privacy to individuals (data sellers): increased search costs, opportunity costs²⁷

²⁷By not providing their personal data to companies, consumers cannot benefit from product recommendations and instead must spend resources searching. Similarly, without product recommendations, a consumer may be un-

2. Benefits of greater privacy to individuals: improved negotiation ability, decreased respondent burden, greater well-being, reduced vulnerability, increased independence²⁸
3. Costs of greater privacy to commercial organisations (data buyers): opportunity costs, increased red-tape and other control costs in storing and maintain private data²⁹
4. Benefits of greater privacy to commercial organisations: protection of reputation, prevention of fines, increased consumer market participation³⁰
5. Costs of greater privacy to society: increased regulation and enforcement costs, decreased social utility of public data, decreased transparency (e.g. of financial markets) leading to illegal or inefficient market and social behaviour, oppression of vulnerable groups (see feminists critique of privacy in the philosophy summary), stagnation and straggling of innovation, increase in surveillance costs
6. Benefits of greater privacy to society: protection of human rights, decrease in surveillance costs³¹

We have no doubt that this is only a small sample of the many costs and benefits associated with privacy.

Many of the difficulties in trading off these costs and benefits are well-known in the literature:

1. Any tradeoff is highly context specific and inevitably subjective (Lindgreen, 2018).³²

aware of what they are missing out on – they may miss a concert from their favourite band for example. Greater privacy can help scammers or other malicious agents in situations where greater transparency could reveal their malpractice. These are opportunity costs.

²⁸By sharing less information on their desires, consumers have more leverage in the negotiating process. Greater privacy means that the consumer must provide less information reducing the respondent burden to surveys. Unscrupulous agents are not able to target vulnerable consumers if the information about their vulnerability is kept confidential. Greater privacy means that an individual will be influenced less by outside forces, e.g. personalised advertisement.

²⁹With less information on their potential customers, companies miss opportunities to market and sell their goods and services.

³⁰A company that values their consumers' privacy will maintain a better reputation and face less regulatory penalties. Greater privacy may mean that individuals are happier to participate in the market – this is an economic rationale behind the EU's data protection regulations.

³¹The societal cost of surveillance is included as a benefit and a cost of greater privacy, since it is unclear whether greater privacy will decrease surveillance costs since the state will collect less information, or increase costs since the state will be hampered by red-tape in their information collection.

³²In addition to the question of context and subjectivity of the tradeoff, there is a question of the correct *resolution* of the privacy cost-benefit tradeoff: For example, should an individual calculate a new tradeoff for each transaction they make in the marketplace? Or should the individual calculate the tradeoff in deciding whether to sign up to a service – such as Facebook – with which they may interact multiple times in the future? Or should the individual

2. Should the individual be modelled as a rational economic agent or should their utility function be based on an empirical understanding of human's flawed and biased cognition? (See the social sciences summary for more details on these flaws and biases; see (Adjerid et al., 2016) for an analysis of their effects in the privacy tradeoff.)
3. A related but currently unexplored (as far as I am aware) question is how biases in organisational psychology influence the privacy tradeoffs made by companies. Perhaps, like individuals, it is not realistic to model organisations as rational agents in their privacy tradeoffs.
4. The costs and benefits for society of privacy have not been quantified and a societal trade-off has not been attempted (Lindgreen, 2018, p.201).

There are also aspects of the privacy tradeoffs which are neglected in the current literature. In the case of national statistical organisations, the tradeoff for the survey respondent is usually framed as a balance between the individual's privacy and their data's utility to society. This misses a key aspect of the equation: respondent burden – the cost in time and effort for an individual to answer a survey (Yan et al., 2020).³³

An individual can reduce his costs – respondent burden and privacy – by not answering or lying on the survey form. On the other hand, DP deliberately ensures that the impact of such behaviours on social utility is small. In fact, analogous to DP's privacy utility guarantee (equation 1.1), the expected impact to social utility from an individual not responding, is bounded below by a multiplicative factor of $e^{-\epsilon}$.

DP therefore appears to provide a strong disincentive for responding to surveys. (See (Drechsler, 2023;

calculate the tradeoff in deciding whether to interact with a particular market segment (such as social media), or even with the privacy marketplace as a whole? At the lowest level, each transaction – such as uploading a photo to social media, or writing a tweet – may appear to have negligible privacy cost; and yet the sum of these costs can be significant. If the tradeoff is done at the company level, the individual must know all the ways in which they will interact with the company in the future, and what the costs and benefits of these future interactions are. This kind of prediction seems to be difficult in a landscape which is constantly changing (corporations rising and falling, individual and societal behaviours evolving, etc.); and so an informed tradeoff at the company-level is correspondingly difficult. Similar issues arise when making the tradeoff at the market segment level.

³³Another aspect to the tradeoff is an individual's sense of civic or scientific duty, which may compel an individual to answer the survey even when the social utility of their response is small. We will not consider this factor in our analysis, which we suspect is minimal in contemporary Western societies, relative to privacy, data utility and response burden.

Kreuter, 2019; Oberski and Kreuter, 2020) for similar arguments.) This may be particularly problematic in an era of declining response rates (Brick and Williams, 2013) and general apathy to surveys and censuses (United States Census Bureau, 2019).

Under DP, the expected impact to social utility from an individual's non-response is small. On the other hand, there are significant privacy costs – or at least, many people perceive there to be significant privacy costs (Mayer, 2002; Office of the Australian Information Commissioner and Lonergan Research, 2020; Stanford, 2020; United States Census Bureau, 2019). Additionally, respondent burden is recognised as a serious concern by national statistical organisations (NSOs) (Holzberg et al., 2021; Data Quality Hub, 2020). Thus, it is likely that the decrease in social utility is outweighed by the reduction in respondent burden and privacy costs. In this case, the optimal action for the individual is to not respond (or to lie). This is the *prisoner's dilemma of differential privacy*: the optimal action for every individual is to not respond, even though this will result in poorer outcomes for all the individuals.

As an example, suppose that the survey has a utility of \$100 for each individual if everyone responds and, following the requirements of DP, decreases by a multiplicative factor of $e^{-\epsilon} = 99/100$ for each person that does not respond; except that if no-one responds then the social utility is obviously \$0. Assume that the privacy and response burden costs are \$50 for each individual. The relative payoff for not responding compared to responding is at least $\$50 - \$100 \times e^{-\epsilon} = \$49 > 0$; hence not responding is the dominating strategy for all individuals, regardless of the behaviour of other individuals. However if all individuals respond, they receive a payoff of $\$100 - \$50 = \$50$; whereas if no one responds, they all receive \$0.³⁴

³⁴One may argue that – following the principles of DP – an individual's privacy utility only decreases by a factor of $e^{-\epsilon}$ if she does respond. As such, the relative payoff for not responding compared to responding should be $(\$50 - \$100)e^{-\epsilon} = -\$0.5 < 0$ and individuals should respond! Moreover, in any survey worth conducting, the baseline utility U_0 (in this case \$100) should be greater the baseline privacy cost P_0 (in this case \$50); so the relative payoff for not responding will always be negative – that is, the prisoner's dilemma can never hold in any realistic setting.

Thus, DP suffers from the prisoners dilemma: it encourages behaviour that will leave everyone worse off.

It is quite possible that other statistical disclosure controls (SDCs) are also susceptible to the prisoner's dilemma; the tragedy of the commons is an unfortunately common phenomenon. However, it is difficult to evaluate other SDCs since they lack the clean mathematical properties that DP enjoys.

1.2.3 A BRIEF HISTORY OF PRIVACY ECONOMICS

The study of the economics of privacy began in earnest in the late 1970s with the Chicago School of economic reasoning (Lindgreen, 2018). This initial line of research (Posner, 1981; Stigler, 1980) focused on privacy as a driver of market inefficiencies: since privacy allows dishonourable individuals to hide their true nature, it drives up costs in staff recruitment for example.

In the 1990s and early 2000s, concurrent with the growing concern in the legal and philosophical fields, economists such as (Varian, 1997) began to take a more nuanced view, acknowledging both benefits and costs of privacy. From there, the current mainstream economic thinking has evolved to emphasis the burgeoning complexities in understanding privacy in today's technological society. (See the survey articles (Acquisti et al., 2016; Lindgreen, 2018).)

At the same time, a separate group of scholars (Abowd and Schmutte, 2019; Hsu et al., 2014; Ghosh and Roth, 2015; Ligett and Roth, 2012; Li et al., 2017) are using DP to advance a theoretical solution to the difficult tradeoffs in the economics of privacy. A related but separate literature (McSherry and Talwar, 2007; Nissim et al., 2012) uses DP to incentivise truth-telling in a game-theoretic setting.

However, this ignores the role of respondent burden. The prisoner's dilemma still holds if the respondent burden is a significant cost – more specifically if it is greater than $e^{-\epsilon}(U_0 - P_0)$.

1.3 PRIVACY FROM THE SOCIAL SCIENCES PERSPECTIVE

Current social commentators are prone to decry the rapid disappearance of privacy in the 21st century. As more and more of our everyday life moves into the cyberspace, there has been a corresponding increase in privacy intrusions: NSA surveillance programs,³⁵ commercial and political infiltration of our private lives through social media,³⁶ and consumer tracking by big tech.³⁷

This concerning trend has rightly caused alarm in the last two decades. Yet we have been hearing similar cries for well over 100 years. Warren and Brandeis wrote in 1890 (Warren and Brandeis, 1890) that “numerous mechanical devices threaten to make good the prediction that ‘what is whispered in the closet shall be proclaimed from the house-tops’”. In 1964, privacy was “evaporating [and] under assault from many directions” (Packard, 1964, p.12) – so much so, in fact, that “we [were standing] on the threshold of what might be called the Age of the Goldfish Bowl” (Brenton, 1964, p.21). In 2001, “advances in technology endanger[ed] our privacy in ways never before imagined” (Garfinkel, 2001).

All of these concerns were voiced before the ubiquity of smart phones, social media, online shopping and Google. If privacy was on its deathbed before all of these new technologies, how is it that ‘the end of privacy’ has not yet been realised? A focus on technology alone will not be useful in answering this question – it is obvious that we have the technology to realise Orwell’s 1984 but we, collectively as a society, have so far chosen not to. Further, the law can at best receive partial credit in preventing the death of privacy, as it has largely failed to keep up with technology’s innovations. I posit that it is the social sciences

³⁵For example, PRISM, MUSCULAR and XKeyscore.

³⁶Cambridge Analytica’s and Russia’s political interference in the 2016 US presidential elections; facial recognition applied to photos uploaded to social media; and personalised advertising based on what an individual has shared and liked.

³⁷Third-party tracking cookies can follow your internet activity across different websites and thereby learn your consumer preferences or infer your personal information.

which best explain the continuing survival of privacy. It is the social sciences which explains the choices of individuals, organisations, government and society as a whole, to protect and adapt privacy in the face of today's technological capabilities.

1.3.1 HISTORY AND THE ARTS

In the above discussion, we have highlighted privacy's continued history in modern Western society.³⁸ Privacy has been a mainstay of political and social discussion following the industrial revolution, going hand-in-hand with the rise of technology. I have documented 20th and 21st century conceptions of privacy in the philosophical and legal summaries (see also (Westin, 2003) for a more detailed history), so this section will focus on earlier history.

Despite the prominence given by critics and proponents alike, technology was not the precursor to the West's modern understanding and value of privacy. Rather, technology played foil to the notion of individualism – both in the political and economic senses – which emerged with the growth of the bourgeoisie in the industrial revolution.³⁹ That “the ordinary human individual rather than the group or estate was the basic unit of human society” was a novel idea in the 19th century (Moore, 1985, p.25) – and a necessary one for today's understanding of privacy. The distribution of property to the bourgeoisie in the early industrial revolution and the resulting economic independence were further steps towards modern privacy. Amongst the lower class, the move from agrarian workers to the industrial proletariat also fostered a sense of individualism. Gone was communal farmwork where both tasks and outcomes were shared. In its place

³⁸The arts are included in this section even though they are part of the humanities rather than social sciences, because, as I will argue, they have had a significant impact on the historical and current discussions of privacy.

³⁹Some scholars see the growth of the middle class and individualism as starting earlier, in the Renaissance and Reformation periods (c.1450-1650), with the writings of Thomas More, John Locke and John Stuart Mill. See (Keulen and Kroeze, 2018) for a summary of this perspective.

were the discrete, particular responsibilities and personal paycheque of the factory worker.

It was this societal shift towards the individual that gave rise to the moral atmosphere conducive to modern privacy theorising. In contrast, technology's comparatively small role in the development of privacy rights is limited to the threat it poses to this moral atmosphere. One need only look to modern non-Western societies such as China (see in particular (McDougall and Hansson, 2002)) to see that a rise in technology is not a sufficient driver for Western notions of privacy.

Privacy in pre-industrial societies – both historically and in modern times – focuses on two narrow aspects: seclusion and protection against authority (including the right of resistance to an unjust ruler) (Moore, 1985). (Moore, 1984) charts these privacy norms in the Western world through the medieval ages back to their Greek and Hebrew traditions, along with ancient Chinese and other non-Western histories. (See the anthropology section for details on non-Western perceptions of privacy.)

The cultural influence of the arts has and continues to have a major impact on privacy discussions. Dystopian fiction – particularly 1984, *Brave New World*, *The Handmaid's Tale*, *A Clockwork Orange*, *Fahrenheit 451* and more recently *The Circle* and *Black Mirror*⁴⁰ – paint vivid futures detailing the surprising and often horrifying consequences of erosions of privacy.⁴¹ The struggle of the individual man (and it is almost always a man) to overcome a repressive government has been portrayed so frequently in

⁴⁰Unsurprisingly, literature has followed the dominating concern of the times. Twentieth century literature focussed primarily on privacy overreaches of totalitarian governments while twenty first century literature has focussed on the power of technology companies to disrupt privacy.

⁴¹An interesting thought experiment is to imagine Orwell or Huxley's reaction to the current state of privacy. Would they be as horrified by today's society as we are of 1984 or *Brave New World*? Perhaps we are only contented with today's society because of its familiarity (as suggested by psychology, see e.g. (Oulasvirta et al., 2012)) and horrified only by the novelty of 1984 and *Brave New World*. That is, maybe these works are concerning only because the privacy intrusions take different forms – telescreens rather than tracking virtual behaviour; government spying rather than commercial surveillance; and control of our emotions, opinions and actions by government rather than corporations – rather than different substances. (This possibility is hinted at by Simson in (Garfinkel, 2001).)

films – such as *Jason Bourne* and *V for Vendetta* – that it is now a trope.

The theme of privacy is not limited to literature and film, nor is it limited to portrayals of future dystopias. The historical-fiction play *The Crucible* describes the disaster befalling a society in which one lacks the privacy to one's thoughts.⁴² The lack of privacy is typically a major theme in any artistic portrayal of totalitarian regimes such as Nazi Germany, Stalin's Russia and the Eastern Bloc more generally, or North Korea. All this is to say that privacy is a concern found pervasively throughout the arts.

The impact of the arts on our collective understanding of privacy cannot be quantified, yet one can only assume it is great. While we have only examined English-language pieces, it is clear that privacy is front and centre in some of the most famous works of the last century. Through their mass consumption, these works have raised awareness of the various aspects of privacy and their importance. Through their compelling narratives, they have highlighted the consequences of privacy loss and spurred many into action. By their creation of cultural symbols (for example, Big Brother) they have built rallying points for privacy advocates. Clearly, an understanding of historical and current privacy trends would be acutely deficient without consideration of the arts' cultural influence.

1.3.2 PSYCHOLOGY, COGNITIVE AND BEHAVIOURAL SCIENCES

In general, human behaviour is multi-faceted and based on a wide range of contextual factors: time of day, mood and emotion, blood sugar level, environment both social and physical including peripheral cues, unconscious biases, previous behaviour both long- and short-term, priming stimulus, ego depletion, decision fatigue and so on (Kahneman, 2011; Baumeister and Tierney, 2011). It is complex to the point of appearing random or contradictory. Human behaviour as it relates to privacy is no exception. A complete

⁴²For another example of non-dystopian artworks on privacy, see Frazer's *Purity*.

understanding of why an individual took some action to increase or decrease their privacy is a chimera. However, there are many situations where privacy decision making is explainable and consistent across individuals and across time. While there are also many fuzzy, inconsistent boundary cases, an understanding of the explainable examples can be sufficiently enlightening.

What is clear is that a degree of privacy is necessary for individual well-being (Stuart et al., 2019), although the particular degree depends on the individual and their position in their society. Privacy provides the space to allow one to relax and escape from external stressors (Margulis, 2003). It allows an individual to regulate and negotiate social relationships (Agre and Rotenberg, 1997). In fact, a gradual increase in self-disclosure is one of the primary drivers for the strengthening of interpersonal relationships (Altman and Taylor, 1973). Without a starting position of privacy and the continuing control of one's own personal information, this interpersonal development would be impossible.

On the other hand, human beings are inherently social animals who have an innate desire to share information about themselves (Acquisti et al., 2015). There is thus a tradeoff at the most basic, psychological level: an individual is constantly balancing her desire for privacy against her desire for socialising. This tradeoff becomes increasingly more complicated as one understands the multitude of consequences – both positive and negative – resulting from any privacy decision. (The various considerations in this tradeoff are detailed in the economics summary.)

In today's environment, it is typically very difficult for the individual to evaluate the privacy tradeoff (Acquisti et al., 2020). As the academic community has come to accept this fact, discussion has moved away from the idea of a 'privacy calculus' (Laufer and Wolfe, 1977; Culnan and Armstrong, 1999; Dinev and Hart, 2006) – where the individual analytically, rationally and deliberately weighs up the various considera-

tions in the privacy tradeoff⁴³ – towards a greater emphasis on the psychological and cognitive mechanisms underpinning privacy decision-making (Acquisti et al., 2015; Lindgreen, 2018; Adjerid et al., 2016).⁴⁴

In reality, humans frequently resort to “low effort”, fast cognitive processes rather than “high effort”, slow deliberative calculations (Kahneman, 2011). In these situations, we employ a number of heuristics and suffer from a number of cognitive biases when making decisions regarding our privacy (Dinev et al., 2015). For example, people place a greater value on privacy when they have it, as compared to when they do not (Acquisti et al., 2015, p.510).⁴⁵ Further, repeated privacy breaches can lose their effect: a behaviour that was once felt as a privacy invasion can later be viewed with indifference, simply due to its repetition over time (Oulasvirta et al., 2012).

Another relevant cognitive bias is temporal discounting. Humans weigh the value of future events less than present events. This helps to explain why we willingly give up valuable private information in exchange for small rewards (Lindgreen, 2018). Any negative consequences of divulging personal information occur sometime in a distant, hypothetical future whereas the rewards are typically immediate. In fact, we know from construal level theory (Trope and Liberman, 2010) that it is difficult for people to even create a clear mental picture of future events, especially when there is uncertainty about whether they will

⁴³Even if humans did employ a privacy calculus, it is unclear that an objective metric of costs and benefits would be appropriate. Since the state of privacy is an inherently psychological phenomenon, subjective metrics which account for intangible goods such as the individual’s well-being would be more correct.

⁴⁴This evolution in understanding has occurred within a larger trend. Over the last 40 years, the economics principle of rational behaviour has been repeatedly questioned as our understanding of human behaviour and decision making has expanded to include concepts of bounded rationality – an agent only has access to a limited amount of information, time and cognition resources – cognitive biases and heuristics. In particular, there is now a well-developed theory – called prospect theory (Kahneman and Tversky, 1979) – on how people make decisions in irrational ways when there is risk and uncertainty (Lindgreen, 2018, p.182).

⁴⁵This is a specific case of the more general phenomenon called endowment bias or loss aversion: You value your possessions in part because they are yours. That is, you value an object you own more than an identical object you do not own. Humans feel a loss more than they feel an equivalent gain.

actually occur (Demmers, Joris, 2018).⁴⁶

An understanding of privacy heuristics and cognitive biases such as temporal discounting sheds light on the *privacy paradox* (Norberg et al., 2007), which asserts that:

People express concern about privacy in the abstract, but in reality readily give up their privacy for a small reward and fail to take easy steps to protect their privacy, even when such steps are obviously available (Solove and Schwartz, 2021).

In other words, there appears to be a disconnect between people's opinions and their actions. Psychological and economic perspectives largely explain this disconnect (Acquisti et al., 2020).⁴⁷ In particular, priming and framing effects mean that the situation and context matter when surveying people's attitudes or monitoring their behaviours (Acquisti et al., 2015).⁴⁸ By asking about privacy, a survey can prime respondents to increase their valuation of privacy.⁴⁹ On the other hand, companies will minimise or obfuscate privacy concerns so that consumers divulge their personal information willingly.

⁴⁶Construal level theory states that an individual will devalue an object when there is a large "psychological distance" between the individual and the object. Psychological distance has four dimensions: spatial, social, temporal and experiential.

Construal level theory gives insight into the tradeoff made by individuals in deciding to answer a survey. Respondent burden is immediate and real. On the other hand, privacy consequences occur in a distant, hypothetical future. Data utility is also temporally and experientially distant while also being socially distant – a survey is unlikely to result in a direct benefit to the participant or anyone in their immediate social circle. Thus, construal level theory suggests individuals will over-value respondent burden and undervalue privacy and utility factors.

⁴⁷The economic perspective is given in the relevant summary, but in brief it explains the privacy paradox by observing a) the informational and computational asymmetry which results in the consumer being unable to make an informed decision and b) the lack of options in the marketplace for the privacy-savvy consumer (Acquisti et al., 2020).

⁴⁸This literature relates generally to questionnaire design and the need for cognitive testing as is widely known by survey statisticians.

⁴⁹See also (Singer and Couper, 2010; Couper et al., 2008, 2010) which shows a different but related priming effect: "making explicit the possible harms that might result from disclosure also reduces willingness to participate" in a survey.

Finally, we end this section by considering the psychology of statistical privacy controls. The only literature on this subject (that I am aware of) relates to randomised response (Warner, 1965). To summarise this research, survey respondents typically do not understand nor trust randomised response as a privacy measure without significant work by the interviewer to build understanding and trust (Kirchner, 2015; Landsheer et al., 1999). This is further evidence that people do not estimate their privacy loss in an economically rational way and are unlikely to be persuaded to respond to surveys by the guarantee of complex mathematical protections such as differential privacy (Oberski and Kreuter, 2020; Drechsler, 2023).

1.3.3 ANTHROPOLOGY AND SOCIOLOGY

On face value, mainstream privacy concerns are relevant only for citizens of wealthy, developed countries. These concerns appear to rest on the assumption of a pervasive technological environment that facilitates massive data collection. Arguably, undeveloped countries do not have such technological environments and so would be unaffected by these privacy concerns. Further, personal data is only commercially valuable insofar as it can be used to influence consumer habits and improve a company's bottom line. Hence, citizens must be sufficiently rich for personal data collection to be commercially viable. Additionally, a government must have adequate resources – in terms of money, technology and expertise – to invest in mass surveillance, which may preclude third-world countries. Finally, people who are struggling to satisfy their basic needs surely would not be concerned with a lesser concern like privacy – or so the argument goes. All of this evidence indicates that the current crisis of privacy is unimportant except for the small minority of citizens who are fortunate to live in a rich, developed nation; that is, privacy is a 'first world problem'.

However, this argument supposes a narrow view of privacy in terms of commercial and governmental

surveillance.⁵⁰ We know (from the philosophical and legal summaries) that privacy has a much broader ambit.

Taking the wide view of privacy, there is much evidence supporting the universality of privacy – or at the very least, the universality of a desire for privacy – across cultures, societies and time (van der Geest, 2018; Acquisti et al., 2015; Moore, 1984; Altman, 1977). Privacy considerations arise in the Bible, the Talmud, the Quran and in Confucian and Taoist writings (Acquisti et al., 2015). Privacy has been observed in primates and other animals (Westin, 1967; Wilk, 2018), suggesting that privacy is fundamental to any functioning society, human or otherwise.

At the same time, privacy practices are highly contextual with complex, situational rules that differ widely between cultures (Altman, 1977). For example, in some cultures, privacy within families can be almost non-existent at the same time that privacy between families is strictly observed (Wilk, 2018). On the other hand, Western societies primarily value privacy in terms of the individual, rather than the family unit.⁵¹

It is not just the extent of privacy that varies between cultures. The moral responsibility and burden of privacy can variously be placed upon the subject – to guard their privacy – or the potential observer –

⁵⁰The argument also assumes that only developed nations have achieved a sufficient level of technology adoption to enable surveillance. This is becoming increasingly false as, for example, rates of smartphone ownership increase in low- and middle-income countries (Miller et al., 2021). There are already examples in the developing world of technology-enabled persecution of minorities or political opponents, such as in the Arab spring.

Further, it is unfortunately the case that many of the poorest countries are also amongst the most unstable and corrupt. Third-world nations often lack the strong institutions that protect the individual from over-reaching authority. They frequently do not have strong protections for freedom of speech and of thought (amongst other privacy rights). Thus, third-world citizens often have more to lose from privacy intrusions than those in first-world liberal democracies. (For example, they can face persecution for LGBTQI status or for expressing dissident opinions.)

⁵¹See Chapter 7 of (Francis and Francis, 2017) for an extended discussion on the questions of privacy and families. As another example of the varying social norms, “Americans, for example, are reputed to be more open about sexual matters than are the Chinese, whereas the latter are more open about financial matters (such as income, cost of home, and possessions)” (Acquisti et al., 2015). For more examples, see (Moore, 1985).

to not intrude another's privacy.⁵² Thus, all human societies have an understanding of privacy but with differing practices. In other words, privacy is "simultaneously culturally universal and culturally specific" ((Acquisti et al., 2015, p.512) paraphrasing (Altman, 1977)).

One explanation for this apparent paradox is that privacy is valued by all cultures, but its relative value amongst other competing priorities differs from culture to culture. For example, lack of within-family privacy may be due to the cohesiveness and communality of the family unit superseding the interests of the individual. Or perhaps the value of paternal care outweighs the independence of children or the elderly in the family (Miller, 2021).

Competing priorities exist at all levels of society, not just within families. At a broad level, there is a clash between the values of liberalism and socialism. How a cultural resolves this clash partly determines the culture's relative value of privacy. In a socialist society, "if the state can enhance social welfare by collecting information about individuals, this automatically supersedes individual rights" (Miller, 2020a). In liberal cultures on the other hand, individuals have an inalienable right to privacy.

However, a cultural valuation theory of privacy fails to acknowledge that there are varying perceptions

⁵²For example, in the Netherlands it is customary to keep curtains open so that, in theory, anyone could look inside another's private dwelling. However, it is very impolite to actually look inside (Van Der Horst and Messing, 2006; Wilk, 2018).

We can also see changes across time within the same society. In 1960, the American legal scholar William L. Prosser wrote (Prosser, 1960, p.422)

No doubt the cases thus far have been sufficiently extreme; but the question may well be raised whether there are not some limits, and whether, for example, a lady who insists upon sun-bathing in the nude in her own back yard should really have a cause of action for her humiliation when the neighbours examine her with appreciation and binoculars.

I can only assume that Prosser used this example as an 'obvious' case where the responsibility for maintaining privacy must lay with the subject (the sun-bathing lady) rather than the observer (her neighbours). Yet modern readers – with an understanding of victim blaming and a distaste for antiquated notions of modesty – could be forgiven for disapproving of the ogling neighbours. Although today's society may not necessarily endorse the lady's action, it is clear that modern privacy norms place a greater responsibility on the observer than past norms did.

of what constitutes a privacy intrusion. It is not simply the case that cultures tradeoff privacy in different ways, but that they also fundamentally perceive privacy in different ways. For example, a gross intrusion of the state into an individual's personal life in one culture is a caring act from a paternalistic government in another culture (Wang, 2019, 2020). There is a “fine line between care and surveillance” (Miller, 2020b) which changes between cultures and across time.⁵³

Understanding where this line stands requires knowing the norms – the socially acceptable behaviours – of the society in question. Even when taking the narrow view of privacy in terms of divulging information, there are many aspects of an information-sharing activity and its context which are relevant in judging whether it is socially acceptable (Nissenbaum, 2010). (See the economics summary for an outline of these aspects, as given in (Nissenbaum, 2010).) The social understanding of acceptable privacy-sharing activities is thus complex and nuanced.

Sociologists are interested in privacy in its own right – as one fundamental characteristic of a society – and in the ways it intersects with other fundamental societal characteristics, such as how privacy is used as a tool for social control; how – depending on the context – both privacy and lack of privacy can promote individual development and group cohesion; and how privacy relates to social stratification and more generally social order and inequality (Anthony et al., 2017; Kasper, 2007).

In the previous section, we observed a privacy tradeoff at the psychological level. We end this section by observing another privacy tradeoff, this time at the society-wide level. On one hand, privacy is a social good (Kasper, 2007) and fundamental to a well-functioning society ((DeCew, 2018) summarising a broad

⁵³The evolution of a culture is particularly apparent in times of crisis. One recent example is the acceptance of COVID19 tracking smartphone apps in liberal countries such as Australia. Pre-pandemic, there was little social appetite for government surveillance of where you have been and who you met up with. But the pandemic changed the privacy tradeoff, resulting in widespread adoption of these apps (Biddle et al., 2021; Miller, 2021; Sharma and Bashir, 2020).

literature including (Mead, 1928; Westin, 1967)). On the other hand, society by definition requires communality. Therefore, “without accepting some intrusions of privacy society cannot exist” (van der Geest, 2018). In summary, there is an inevitable tension between a lack of privacy as a cohesive force for society and a degree of privacy as a necessity for well-functioning society. In the words of the anthropologist David Miller,

“We cannot be for or against privacy. It must be a question of balance.” (Miller, 2020a)

1.3.4 POLITICAL SCIENCE

What is meant by the term ‘politics’ is highly contested but one mainstream definition takes it to denote the public affairs of a society (Heywood, 2013; Leftwich, 2004). More explicitly, this definition asserts that “the distinction between ‘the political’ and ‘the non-political’ coincides with the division between [the] public sphere of life and [the] private sphere” (Heywood, 2013, p.5).⁵⁴ In this case, politics and privacy are complementary concepts: what is political is not private, and visa versa. Could then political science – aka the study of politics – be characterised as the study of what is not, or should not be, private? Could the boundaries of the political inform us of the boundaries of the private?

Alternatively, we could take politics to mean the “exercise, legitimacy and organisation of power” (Raab, 2018, p.257) (see also (Heywood, 2013, p.9 “politics as power”)). While not immediately apparent, this definition is also complementary to privacy. The boundaries of power over the individual ascribe the rights of the individual “to be let alone” (Warren and Brandeis, 1890) – or, in other words, the right to privacy. And conversely, the right to be let alone delimits the boundaries of political power.

⁵⁴This conception of politics and privacy can be traced back to Aristotle’s *Politics*.

So far, I have been deliberately vague on who holds this power because, under this broad definition, political power can appear at many levels: a government, an organisation or a family. In fact, political power can appear in all social activity, in all groups, institutions and societies (Leftwich, 2004). Freedom of the press is a non-obvious example of political power; and the limits to journalist's free expression are curtailed by the rights to privacy (Lever, 2015). In saying that, the most recognisable politics-privacy dynamic concerns the power of the state as restricted by a bill of rights (see the constitutional right to privacy in the legal summary).⁵⁵

Indeed, the U.S. *Bill of Rights* and the French *Declaration of the Rights of Man and of the Citizen* were important drivers for both today's privacy and today's politics (Keulen and Kroeze, 2018, p.28).⁵⁶ More generally, democracy necessitates a degree of privacy and visa versa (Moore, 1985), since both concepts rest on the principle of the inviolate individual. Many of the tenets of democracy – freedom of voting, speech, assembly, religion and unwarranted government intrusion – are rights of privacy. At the same time, some level of transparency is required in a true democracy (Francis and Francis, 2017, p.290).⁵⁷ A

⁵⁵More concretely, the state is provided the power to acquire information necessary for the well-functioning of society – for example, search warrants, surveillance to identify anti-social activity but also official statistics. What is 'necessary for the well-functioning of society' is hotly contested and varies between nations but it is limited by the informational privacy rights of that society. The state is also provided the power to restrict individual behaviour to the benefit of society's functioning or morals – for example, controlling abortion rights; restricting child pornography; or implementing public health measures such as adding fluoride to drinking water (Keulen and Kroeze, 2018, p.36) or mandating vaccination. The state's power in this regard complements the legal notion of decisional privacy rights.

⁵⁶These documents were created around the time of the industrial revolution with the growth of the bourgeoisie. As I argued in a previous section, this was the environment in which the modern notion of privacy was conceptualised. It was also the environment which birthed the modern liberal democracy.

⁵⁷Primarily, a democracy requires a transparent government, secure freedom of information rights and a press that is free to be politically critical. Secondly, a democracy also requires individuals to sacrifice their privacy. In order to limit corruption – which is antithetical to democracy – some prying into the private lives of politicians and public servants is inevitable. Finally, national security concerns require the private citizen – not just the public official – to give up some of her privacy in order to secure the democratic state against insurgent forces (Francis and Francis, 2017, p.290).

privacy tradeoff – termed the “paradox of the liberal state” (Keulen and Kroeze, 2018, p.31) – therefore arises in the political context as well the psychological and sociological.

THE POLITICS OF ALLOCATING PRIVACY

A third definition of politics states that it is the decision-making of resource allocation within a society (Boswell, 2020). If privacy is a resource, how should it be distributed? This question is fundamentally political.

As embodied in the UN’s *Universal Declaration of Human Rights*, the legal ideal is that privacy is a basic right. As such, all individuals should enjoy some minimum level of privacy. In other words, privacy should be distributed equally. However, the political reality may be very different to the legal ideal. There is a strong argument that privacy is now a luxury good: The base price of a typical good or service is subsidised by the collection of personal information; and privacy is available only to those who can afford the (often expensive) extra premiums (Papacharissi, 2010).⁵⁸ Maintaining one’s privacy then requires one to repeatedly pay a premium for each good or service. Privacy is thus a luxury in the sense that the aggregate cost of these premiums is beyond the budget of the average citizen.

If politics is resource allocation decision-making, then technological artefacts have politics (Winner, 1980; Bowles, 2018).⁵⁹ That is, a technology confers a certain distribution of resources when embedded in a society. It is true that how a technology furnishes this distribution amongst members of society is

⁵⁸The extra premiums are not necessarily monetary; they could be the extra time required to monitor and disable settings to share your personal information, or the requirement to be savvy and computer-literate. Alternatively, it may be that it is impossible to access the good or service without degrading privacy. In this case, privacy is a luxury good since maintaining privacy would require the often-significant cost of forgoing this good or service. For example, one must divulge their personal information to access Facebook. On the other hand, forgoing Facebook can come at significant cost of social capital. As such, it may only be an option to people with sufficient social resources.

⁵⁹See also the literature on algorithmic fairness for more examples of how technology can have unintended and unexpected political consequences.

often implicit. But despite it being hidden from plain view, a technology's implied resource allocation is no less important than more explicitly allocations, such as a government's policy agenda.

Differential privacy (DP) is no exception in this regard. (In its vanilla version) it is equalitarian since the same protection is given to everyone. Traditional statistical disclosure control techniques are typically more equitable than they are equal: more protection is given to unique or distinguishable records, and minimal or no protection is given to the 'average' person.

DP is also political in its focus on the individual as the unit of protection. The question of whether individuals, groups, or sub-populations deserve protection, is a question of privacy resource allocation. It is thus a question of politics. DP's implied answer to this question is therefore political.

Marginalised populations are defined by the fact that they are discriminated or excluded based on their population-level characteristics. Thus, almost by definition, marginalised populations can be hurt by group-level inference – not just individual-level inference. On the other hand, members of majority populations are not hurt by group-level inference, since they are not discriminated by their population-level characteristics. DP is therefore political in its choice of protecting the individual and its absence of protection for marginalised groups.

This is not to say that DP cannot be modified so as to adapt its politics. This discussion solely serves to highlight the politics implicit in a technology, particularly as they relate to the question of privacy. It is a reminder that a well-developed technology has hidden political biases just as a human expert has; and that we must be cognisant to these biases as we implement the technology in society, lest we cause harm and reinforce systemic discrimination by doing so. My aim with this discussion is to open the door to further political analysis of privacy technologies (particularly by experts who know more about politics than I).

SUMMARY

To summarise, privacy and politics are inseparable. It is not simply that changes in one influence the other; rather any change in one *is* a change in the other. A shift in how the state interacts with the citizen (for example, personalised e-government service, surveillance or national identifiers) is simultaneously a change of both politics and privacy. Yet, the study of privacy has been largely neglected by political scientists (Raab, 2018).

Political scientists are concerned with resource allocation – the questions of “who gets what, when and how” (Boswell, 2020). They have a well-developed theory of the exercise, legitimacy and organisation of power. They elucidate the distinction between public and private affairs. These matters naturally arise as we grapple with today’s privacy issues. Thus, the political perspective on privacy is integral to balancing the many competing interests in the privacy debate.

1.4 PRIVACY UNDER A PHILOSOPHICAL LENS

SOLOVE BEGINS HIS 2008 MONOGRAPH “Understanding Privacy” (2008) by stating:

Privacy ... is a concept in disarray. Nobody can articulate what it means. Currently privacy is a sweeping concept.... Philosophers ... have frequently lamented the great difficulty in reaching a satisfying conception of privacy.

As a further complication, we must weigh privacy’s virtues – for example, its protection of human autonomy, amongst other things – against its ability to enable corruption and abuse.⁶⁰ Moreover, many philosophers argue that privacy can, at least in most cases, be explained in terms of other rights or moral goods (Thomson, 1975). Thus, the concept of privacy is afflicted by slipperiness in its multiple meanings

⁶⁰That privacy could be negative has been known since antiquity through the myth of the ring of Gyges (Laird, 2001), and remains relevant today as evidenced by the feminist critique of privacy (MacKinnon, 1989).

and uncertain scope; in its sweeping reach across many facets of life and society; in its position as a virtue on the precipice of vice; and in its frequent reduction to other rights.

Philosophical discourse on privacy dates back to Aristotle's distinction of the public and private spheres. Yet the philosophical development of privacy only began in earnest with the late 19th century legal discussions (Warren and Brandeis, 1890) and continues to this day closely entwined with the law's understanding of privacy.

The 20th century saw much debate over how to conceptualise and ascribe value to privacy. One popular view, as described in (Parent, 1983), characterises privacy as control over information: privacy is defined “as the condition of not having undocumented personal information known or possessed by others” (DeCew, 2018). A related, but in some ways broader, conceptualisation defines privacy in terms of access – to an individual's body, thoughts, personal information or attention; or to a physical location (Bok, 1982; Gavison, 1980; Allen, 1988; Moore, 2003). Other philosophers took the reductionist view (Thomson, 1975) that privacy is not a concept in and of itself, but simply a proxy for a variety of other moral interests; as such any perceived right to privacy is no more than a derivative of other rights.

Accounts of the philosophical value of privacy are similarly varied: One view considers privacy as essential for one's autonomy and for developing a concept of self as an independent agent (Gross, 1971; Henkin, 1974). Others see privacy as *instrumentally* valuable⁶¹ for human dignity (Bloustein, 1964); put another way, “invasion of privacy is best understood, in sum, as affront to human dignity” (DeCew, 2018). A third view takes privacy as important because intimacy is impossible without it (Inness, 1992), or more generally because the ability to be self-expressive is critical for developing any interpersonal relationship (Rachels, 1975). Philosophers have used these arguments to develop *values-based* definitions of privacy.

⁶¹ *X* is instrumentally valuable if there is a valuable good which cannot exist without *X*, or degrades as *X* degrades.

Privacy has not been free of criticism. Most prominently, feminist critiques (MacKinnon, 1989; Allen, 1988) worry that privacy, particularly in the home, enables domestic abuse. More generally, privacy can aid a wide variety of crimes, such as financial malpractice or acts of terrorism. However, privacy can be useful for facilitating actions which are acceptable yet run against the prevailing social norms, such as LGBTQI relationships. Therefore, there is a continuing difficulty in drawing the fine line separating the public/private dichotomy (Moore, 2015).

21st century trends appear to move away from an abstract understanding of privacy in terms of an essential characterisation. Many contemporary authors see any such approach – including values- and concept-driven definitions – leading inevitably to flawed theories of privacy. Instead, modern understandings have moved towards pluralistic views of privacy as a set of inter-related concepts (DeCew, 2018), which can only be properly understood and constructed in context (Moore, 2008). Broadly these approaches (see (Nissenbaum, 2010; Barth et al., 2006; Rubel, 2006, 2011; Solove, 2002, 2008; DeCew, 1997)) can all be thought of as normative theories⁶² of privacy and are collectively termed *family resemblance views* of privacy.⁶³

⁶²Normative theories of privacy describe what *should* be private based on an understanding of a given society/culture’s norms. In contrast, a descriptive theory of privacy details what *is* private. Both types of theory can be useful in describing the reality of privacy as it exists in society – although one should obviously build systems which uphold the ideal of a normative theory, rather than the reality of a descriptive theory.

⁶³Our ontology of privacy also follows this modern trend, by specifying, in a given context, A) what should be kept private (the ‘secrets’) and to whom (the ‘attackers’); B) how, and by what extent, privacy can be lost (the ‘measure of increase in attacker’s confidence’); and C) what the contextual value of that privacy loss is (the ‘loss function’).

There is a parallel here between the historical trends in the philosophical and computer science understandings of privacy: Computer scientists, like philosophers, began by trying to distil privacy into a single characterisation. (In the case of computer science, this characterisation is the single equation of differential privacy $\Pr(\mathcal{M}(D) \in \mathcal{A}) \leq e^\epsilon \Pr(\mathcal{M}(D') \in \mathcal{A})$.) After about 50 years, philosophers gave up on this elusive ideal and moved to a context-driven approach. Now, almost 20 years later, the computer science literature – our work included, along with (Nissim and Wood, 2018; Kifer and Machanavajjhala, 2011, 2014) – is beginning to argue that a single characterisation is impossible and a context-driven approach is, in fact, required. Philosophers can be forgiven for feeling a bit exasperated.

PART I

DIFFERENTIAL PRIVACY IN THE US CENSUS

This page intentionally left blank.

2

Five Building Blocks of Differential Privacy¹

2.1 MOTIVATION: WHY DO WE NEED TO IDENTIFY BUILDING BLOCKS FOR DP?

DIFFERENTIAL PRIVACY EXEMPLIFIES A FORMAL APPROACH to data privacy protection. It advocates a mathematical formulation of the otherwise elusive concept of ‘privacy,’ which in turn provides design principles to data release algorithms that provably satisfies the stated formulation. Less than two decades since the inception of the first differential privacy definition, *ϵ -indistinguishability* (Dwork et al., 2006b), differential privacy has grown into a vast literature, with a simple Google Scholar search of the phrase returning over 8 million results as of this writing. The interest in differential privacy is fueled by renewed threats to the confidentiality of individual data contributors as one of the core challenges to the modern reality of increased data collection, sharing, and dissemination.

¹Based on work coauthored with Ruobin Gong and Xiao-Li Meng.

As differential privacy finds its application in a variety of scientific, commercial, and administrative contexts, much scholarly attention has been devoted to refining the original definition in ways that suit the unique characteristics of the data and relevant usage. For example, straightforward mathematical generalizations to ϵ -indistinguishability are found to be conducive to thinner-tailed privacy noise distributions, which can be less harmful to the data quality. Certain structured databases, such as network, graph, and geospatial data, present complications in conceptualizing the notion of an “individual” and gave rise to varied treatments. Also prevalent, particularly so in the context of official statistics, are external constraints that the data product must satisfy. These practical restrictions gave rise to branches of work that recapitulate DP under invariants, as well as other empirically defined or derived standards or hybrids. We provide an extended, albeit still selected, review of some of these variants in Section 2.3.

The myriad of modifications made to the original definition of differential privacy, as much as they may be justifiably motivated by the use case to which they apply, insinuates the worry that the essence of its protection becomes diluted (Dwork et al., 2019). Without standardization and comparability, it becomes challenging for stakeholders to tell when a definition becomes “too” diluted. Separately exacerbating is the tremendous, though ironic, success that differential privacy has achieved in streamlining privacy loss into a single numerical quantity. Indeed, proposals that discuss the “privacy-utility tradeoff” frame privacy loss as a one-dimensional value that can be depicted on an axis (see e.g. Abowd and Schmutte, 2019, Fig. 1). While convenient, an oversimplified narrative of privacy protection induces forgetfulness that “privacy loss” is no more than a shorthand representation of a set of complex probabilistic characteristics associated with a data product to which real people contributed their information.

2.2 CONTRIBUTIONS: FIVE BUILDING BLOCKS OF DP

Our starting point is the question, *who is eligible for privacy protection?* Under our framework, this question is answered by specifying the actual, potential or counterfactual datasets that are to be protected. The collection of all such datasets is called the data space, or *domain*, and is denoted by \mathcal{X} . From an attacker’s perspective, the confidential dataset x – which always belongs to \mathcal{X} by design – is the unknown “parameter” to be inferred from the released data, hence the choice of \mathcal{X} is conceptually analogous to the choice of a parameter space in standard statistical inference. Based on the confidential dataset x , some output statistics are computed and published via a *data-release mechanism* T – a random function of $x \in \mathcal{X}$.

In this work, we take \mathcal{X} as fixed, for it is sufficient for all subsequent developments. However, explicating \mathcal{X} is necessary but insufficient. As we discuss repeatedly, permitting invariants necessitates the concept of a data *universe* $\mathcal{D} \subset \mathcal{X}$, which is a collection of all datasets that are deemed possible for a specific world in which we live. For example, when the enumerated US population size N_{US} is 330 million, then any data set that does not lead to its total count N to be 330 million will not be eligible for being in the data universe when N_{US} is an invariant.

A reader may wonder why we do not restrict \mathcal{X} in the first place to eliminate any dataset that violates $N = 330,000,000$. The answer is because the particular value of the observed national total is *accidental* — even if it was a completely accurate enumeration (which it never is), it is only the total at the time of the Census. The DP mechanism needs to work regardless of the actual value of the total. This leads to the concept of a data *multiverse* \mathcal{D} , which is a collection of all possible data universes \mathcal{D} – be they actual or hypothetical – to which privacy protection can be extended. Hence the data multiverse is specified by *essential quantities* (e.g., N_{US}), while a data universe corresponds to the *accidental values* these quantity take

(e.g., $N_{US} = 330,000,000$), a distinction necessary to ensure the generality of the theoretical guarantees of DP.

The data multiverse \mathcal{D} answers the question, *to where does the protection extend?* Next is the question of *what is the granularity of protection*, which can be a source of confusion, as well as potential for manipulation in capable but malicious hands. The granularity is given by the protection units – the entities whose data changes when the data x is counterfactually altered. Individual persons or business entities are common choices for the protection units, but they are not the only ones. For example, for the privacy protection of electronic communications, the unit may be defined as a single message sent by a person, rather than the sender herself. As an individual may send many messages a day, such a fine-grained privacy unit allows a social media platform to declare a privacy loss budget that is impressively small on a nominal level, even though the actual risks of identification of the sender remain exponentially large.

In DP, a protection unit is formally conceptualized via a premetric $d_{\mathcal{X}}(x, x')$, which is a measure of the difference between two datasets x and x' in \mathcal{X} . A unit of privacy protection corresponds conceptually to a unit difference in $d_{\mathcal{X}}$. That is, the difference between datasets x and x' with $d_{\mathcal{X}}(x, x') = 1$ – which might be, for example, the deletion of one record, or the alteration of a single attribute – is the formal definition of a protection unit.

Fundamentally, DP quantifies privacy protection as the rate of change in output variations. This rate is calculated with respect to the premetric $d_{\mathcal{X}}$ on the data space \mathcal{X} but, *how is the change in output variations measured?* A premetric is also used for this, except in this case the premetric $D_{Pr}(P_x, P_{x'})$ is a measure of difference between probability distributions P_x and $P_{x'}$. Here, P_x denotes the probability distribution of the released statistics, as a function of the confidential data x , where the randomness in P_x is introduced solely by the data-release mechanism T . This is a sensible approach to privacy quantification: as all statistical in-

formation is created by variations in the data, by limiting the relative changes in the output distributions, we limit the changes in variations due to the change from x to x' as measured by $d_{\mathcal{X}}(x, x')$. We will show in Section 2.4 that the most common DP definitions, including the classic *pure* ε -DP, *approximate* (ε, δ) -DP, and *zero-concentrated* ρ -DP (zCDP), are all special cases of the general formulation:

$$\frac{D_{\text{Pr}}(\mathbf{P}_x, \mathbf{P}_{x'})}{d_{\mathcal{X}}(x, x')} \leq \varepsilon, \quad \text{or, more correctly,} \quad D_{\text{Pr}}(\mathbf{P}_x, \mathbf{P}_{x'}) \leq \varepsilon d_{\mathcal{X}}(x, x'), \quad \text{for all } x, x', \quad (2.1)$$

with different choices for D_{Pr} and $d_{\mathcal{X}}$. We provide two expressions here because the first one resembles the familiar notation of taking derivative, and hence the term “differential privacy”; while the second shows that mathematically a DP specification is simply a Lipschitz continuity condition on \mathbf{P}_x as a function of the input data x . (Informally speaking, Lipschitz continuity is simply a generalization of differentiation.)

Only with answers to the above questions – “who” “where”, “what” and “how” – can we render the privacy loss budget (the answer to “how much”) a concrete meaning. Taken together, these five answers form the building blocks of a *differential privacy specification*:

- The protection domain (*who* is eligible for protection?), as defined by the set \mathcal{X} .
- The scope of protection (*where* does the protection extend to?), as instantiated by the multiverse \mathcal{D} , which is a collection of universes $\mathcal{D} \subset \mathcal{X}$.
- The protection unit (*what* is the granularity of protection?), as conceptualized by the input pre-metric $d_{\mathcal{X}}$ on the domain \mathcal{X} .
- The standard of protection (*how* to measure change in the output variations?), as captured by the output pre-metric D_{Pr} on the released data’s possible probability distributions.

- The intensity of protection (*how much* protection is afforded?), as quantified by the privacy-loss budget $\varepsilon_{\mathcal{D}}$ for each data universe \mathcal{D} (where smaller budgets correspond to a higher intensity of protection).

2.3 AN ETYMOLOGICAL ACCOUNT OF DP

2.3.1 THE ORIGIN OF DP: ε -INDISTINGUISHABILITY

A data custodian is interested in publishing a privacy-protected (i.e. *sanitized*) statistic $T \in \mathcal{T}$ based some data $x \in \mathcal{X}$. The data x is some representation of a population – a collection of *individual entities*, which need not be persons, but could be, for example, households or business entities; and the statistic T is simply a function (i.e. a transformation) of the data x .

Although the data x is frequently held in confidence by the data custodian, this is not always the case. For example in randomized response (Warner, 1965) (or in local DP (Kasiviswanathan et al., 2011) more generally), the data custodian does not have access to x . We therefore use the term ‘*data custodian*’ in this paper to refer to the entity responsible for designing and implementing the function T .

We name T the *data-release mechanism* to emphasize that, in addition to privacy protection, T may encompass many other data processing steps (such as cleaning, coding, imputation, etc.) from data collection – or even earlier – through to data publication (see Subsection 2.4.2). In fact, the dual role of T as simultaneously a statistic about the population from the data user’s perspective and a data privatization mechanism from the privacy analyst’s perspective creates a fundamental tension in its design, one that has come to be known as the *privacy-utility tradeoff* (see e.g. Abowd and Schmutte, 2016).

For a concrete example, T may be the mean of the realized data x . This may not be sufficiently privacy-protecting (indeed, according to the standards of DP, it is typically not), in which case the data custodian

may add some noise to the mean before release, so that T is a noisy mean of x .

The first definition of DP, called ε -indistinguishability (and, later, pure ε -DP), is given by [Dwork et al. \(2006b\)](#) and is paraphrased below.

Definition 2.3.1 (Definition 1 of [Dwork et al. \(2006b\)](#)). Let the dataset x be a vector of n records from some domain \mathcal{R} , typically of the form $\{0, 1\}^d$ or \mathbb{R}^d . A data-release mechanism T is ε -indistinguishable if for all neighbors – i.e. pairs $x, x' \in \mathcal{R}^n$ of datasets which differ in exactly one record – and for all outputs $t \in \mathcal{T}$:

$$\left| \ln \frac{\mathbb{P}(T(x, U) = t)}{\mathbb{P}(T(x', U) = t)} \right| \leq \varepsilon. \quad (2.2)$$

Here, using our generic notation, \mathcal{T} is the space of possible outputs, or simply the output space. Typically $\mathcal{T} \subset \mathbb{R}^d$, but it can be more complex. For example, if T is an algorithm that turns a confidential dataset into a synthetic one, then \mathcal{T} contains all possible configurations and values of the synthetic dataset.

It is important to emphasize that T is a *random* map from \mathcal{X} to \mathcal{T} . The probability \mathbb{P} as it appears in (2.2) is induced solely by the auxiliary randomness (the *seed*) U in T and not in x , which is treated as fixed. To pinpoint this source of randomness we assume (without loss of generality) that $T: \mathcal{X} \times \mathcal{U} \rightarrow \mathcal{T}$ is a function of both x and some auxiliary random variable $U \in \mathcal{U}$. When the dependence of T on U or on both (x, U) is apparent from the context, we write $T(x)$ or just T for simplicity.

Since it is not modeled, the dataset x – as the object of an attacker’s inference – plays the role of the *parameter* in privacy analysis. An immediate consequence is that any distribution placed on x can be viewed either as a posited generative model for x or as a prior distribution for x , or a blend of both in its construction. We emphasize the crucial role of x by denoting the law of $T(x, U)$ by \mathbb{P}_x .

Definition 2.3.1 allows for an intuitive understanding of DP as a condition on the data-release mecha-

nism T . It is the requirement that if the data x change slightly, then the output $T(x)$ – or more precisely, the distribution $P_x(T \in \cdot)$ of the output – also only changes slightly. In other words, DP requires that the stochastic behavior of T be robust to small perturbations in x .

To state the intuition mathematically, a data-release mechanism T satisfies ε -DP if the “derivative” of $\ln P_x$ with respect to x is bounded within $[-\varepsilon, \varepsilon]$, for all *permissible* datasets x . The considerations of what constitute permissible lead to concepts of *multiverse* \mathcal{D} and universe \mathcal{D} , as we shall discuss in detail shortly. Furthermore, the choice of $\ln P_x$ has the statistical interpretation of controlling information loss via limiting the power of any hypothesis test for distinguishing x and x' (Wasserman and Zhou, 2010; Kifer et al., 2022; Bailie and Gong, 2023a). But without casting this choice in a broader framework that permits potentially other choices, we deprive ourselves the opportunity to seek the optimal mathematical representations of privacy control, including forming criteria and guiding principles for defining the optimality. Clearly the meaning of ε – which in this context is called the *privacy loss budget* – depends on which derivative we choose (as well as the broader social context (Nissenbaum, 2010)). These are the issues that we will examine one by one in Section 2.4, after a brief overview (Subsection 2.3.2) of the major works extending pure ε -DP, as ε -indistinguishability is now termed.

2.3.2 GENERALIZATIONS, RELAXATIONS AND VARIATIONS OF PURE ε -DP

Nearly two decades of research following Dwork et al. (2006b) has taken DP to the academic forefront of data privacy protection. Central to this work is the quest to delineate the very notion of “privacy” that is encapsulated by the mathematical definition of DP. While a plethora of literature exists on the subject (see Cuff and Yu (2016); Tschantz et al. (2020); Kifer et al. (2022) and the references cited therein), there remains a significant misperception in some of the common (non-technical) narratives concerning what

DP truly safeguards. This misperception largely stems from the slipperiness in translating the Lipschitz bound (2.1) into more intuitive (i.e. less mathematical) notions of privacy. Such translations sometimes rely on the “strong adversary” (SA) assumption, which appears on the surface to correspond to the “worst-case” scenario. This assumption posits that the adversary knows with certainty the entire population except for the data belonging to a single unit, which is the adversary’s intended target (Tschantz et al., 2020). (For the purposes of an informal exposition, the reader may think of a unit as an individual person; more generally, protection units are defined via the choice of $d_{\mathcal{X}}$ in a DP specification – see Section 2.4.4.) In place of the strong adversary assumption, other work use any of a number of commensurate assumptions – which we collectively term generalized strong adversary (GSA) assumptions to highlight their similarity to SA – such as: the non-target units are conditioned upon, or independent of, the target unit; or the target unit can be counterfactually manipulated or deleted (e.g. via the *do*-operator (Pearl, 1995)); or the target unit can be counterfactually resampled from a distribution which can depend arbitrarily on all other units in the population (Dwork et al., 2006b; Wasserman and Zhou, 2010; Kifer and Machanavajjhala, 2011; Kasiviswanathan and Smith, 2014; Tschantz et al., 2020; Kifer et al., 2022).

If DP protects against the strong adversary, who has access to the maximal amount of auxiliary information, it should stand to reason that DP would also protect against weaker adversaries. Yet the greater the knowledge we attribute to an adversary, the more we limit the remaining information which the adversary can attack. Consequently, DP actually offers diminished protection against adversaries whose information-seeking goes beyond the confines of the GSA assumptions – assumptions we contend are rarely met in reality, a point we will elaborate on in a forthcoming article (Bailie et al., 2025e). In fact, regardless of how one measures the adversary’s knowledge-gain from observing a DP output T (e.g. ac-

cording to the power of the adversary’s hypothesis test relative to its size; or the increase in their posterior relative to their prior, or relative to a counterfactual posterior where the target unit did not contribute to the data; or their improvement in reconstructing the target unit’s record (Jarmin et al., 2023; Kifer et al., 2022; Bailie and Gong, 2024)), a GSA assumption is necessary to ensure that pure ϵ -DP bounds the attacker’s knowledge-gain by the nominal level $\exp(\epsilon)$. More generally, a DP specification directly protects only those units which can be counterfactually altered (as determined by $d_{\mathcal{X}}$), and only when such alterations are not just mechanistically conceivable (i.e. in the same data universe \mathcal{D}), but are in fact statistically feasible, as made possible by a GSA assumption. Any one of the GSA assumptions works because they all have the common effect of reducing the attackable information to that which is completely unique to the target unit – i.e. to the variations unexplained by any other unit in the database, or by knowledge on (and beyond) the database population. When the literature states that DP provides relative privacy, this is what is meant: protection of what is left of your personal data after excluding any information that can be inferred from your relatives, colleagues, neighbors and broader community (Hotz and Salvo, 2022; Jarmin et al., 2023). (This is in contrast with the alternative definition of relative privacy as the protection of information which is a-priori unknown to the attacker.)

Additional information may still be protected by a DP specification, but at degraded levels of protection, where the level of degradation depends on the specification’s “group privacy” properties.² DP’s relative protection of this additional information is modelled by an adversary’s knowledge-gain under some relaxation of a GSA assumption. This relaxation might be, for example, the assumption that the adversary knows all but $k > 1$ units. So these sorts of relaxations limit the information-seeking of the attacker not

²Group privacy (Dwork and Roth, 2014) refers to the bound on $D_{\text{Pr}}(P_x, P_{x'})$ when $d_{\mathcal{X}}(x, x') > 1$.

to an individual unit’s unique information, but instead to that of a group of units (Kifer and Machanavajjhala, 2011).

It is crucial for discussions on the essence of DP to steer clear of perpetuating misinterpretations and misconceptions. On this front, we briefly make two points. Firstly, since data privacy is a multi-dimensional problem (to hint at just a few dimensions, consider the adversary, the legitimate data user, the data-generating mechanism, the probability induced by the data-release mechanism and the a-posteriori observed output of the data-release mechanism), the worst-case scenario is not well-defined without first specifying A) how to rank scenarios when deciding which is worst; B) which dimensions are allowed to vary when determining the “worst-case”; and C) which dimensions are being held constant – and indeed there must be some dimensions held constant in order to speak about non-trivial worst-case protection (Dwork and Naor, 2010; Kifer and Machanavajjhala, 2011). Secondly, we emphasize that when we write throughout this article about the privacy protection afforded by a DP specification, we are referring solely to the Lipschitz continuity bound (2.1), rather than bounds on the knowledge gained by an adversary – unless we explicitly state otherwise. The translation of (2.1) into a measure of protection against an adversary³ is the subject of much work (see Dwork et al. (2006b); Ganta et al. (2008); Wasserman and Zhou (2010); Dwork and Naor (2010); Bassily et al. (2013); Hall et al. (2013); Kasiviswanathan and Smith (2014); Cuff and Yu (2016); Dwork et al. (2016); Kairouz et al. (2017); Balle et al. (2019); Tschantz et al. (2020); Desfontaines et al. (2020); Protivash et al. (2022); Kifer et al. (2022); Bailie and Gong (2023a); Bailie et al. (2025e)). In this article, we limit our attention on such work to the preceding discussion and one additional remark: Any non-vacuous translation requires assuming a GSA, or one of its relaxations (Dwork and Naor, 2010; Kifer

³Such measures of protection are called ‘privacy semantics’ in the literature, following the terminology of ‘semantic security of encryption’ (Goldwasser and Micali, 1984).

and Machanavajjhala, 2011). This is not a criticism of DP since such an assumption is necessary in order to artificially demarcate a boundary between individual- and population-level information (which, by the sorites paradox, are naturally two ends of a continuum) and hence these complications are inevitable to any workable statistical-inferential conceptualization of data privacy which permits population-level learning.

Even putting such fundamental issues aside, as differential privacy has been applied to a myriad of practical contexts, researchers have come to realize the limits of the original definition of Dwork et al. (2006b), and to recognize the need for variants and relaxations of pure ε -DP. In this subsection, we briefly review some of this literature, grouped in terms of the four building blocks D_{Pr} , $d_{\mathcal{X}}$, \mathcal{D} and \mathcal{X} .

A first branch of work seeks to augment and relax the ways that the change in output distributions are measured (Subsection 2.4.5). Pure ε -DP requires that the log-likelihood ratio between P_x and $P_{x'}$ is universally bounded between $-\varepsilon$ and ε . This requirement is too stringent in many practical settings, such as the US census, since it requires fat-tailed (i.e. $e^{-O(|x|)}$ density) noise distributions. As a result, notions such as (ε, δ) -approximate DP (Dwork et al., 2006a), computational DP (Beimel et al., 2008; Dwork et al., 2006a; Mironov et al., 2009), Rényi DP (Mironov, 2017), concentrated DP (Bun and Steinke, 2016), f -divergence privacy (Barber and Duchi, 2014; Barthe and Olmedo, 2013) and f -DP (including Gaussian DP) (Dong et al., 2022) relax this requirement by considering different choices for D_{Pr} in place of the multiplicative distance (defined below in equation (2.9)) which is mandated by pure ε -DP. These different choices for D_{Pr} trade-off the strictness of pure ε -DP for an increase in efficiency of statistical inference and estimation.

A second branch seeks to clarify what databases constitute neighbors and to give flexibility to this choice (Subsection 2.4.4). For example, the data custodian may want an asymmetric neighbor relation (Kotso- giannis et al., 2020; Takagi et al., 2022). More generally, $(\mathcal{R}, \varepsilon)$ -generic DP (Kifer and Machanavajjhala,

2011) defines neighboring datasets with an arbitrary relation on \mathcal{X} . In our terminology, this line of work explores different choices of the input premetric $d_{\mathcal{X}}$ and hence investigates the concept of the protection unit. Further examples include applications to social networks and graph data which conceptualize neighbors according to modifications of a single edge or node (Hay et al., 2009; McSherry and Mahajan, 2010). Understanding neighbors as datasets separated by a unit distance lead to d -metric DP (Chatzikokolakis et al., 2013) – a generalization of pure ε -DP which uses a metric d instead of neighbors. Flexibility in the choice of neighbors has been used to accommodate the need to allow for structured databases for which some existing information already exists. Blowfish privacy (He et al., 2014) considers neighbors induced by a policy graph, which is designed to encode known constraints on the dataset. Related variants include element-level DP (Asi et al., 2022), distributional privacy (Zhou et al., 2009), one-sided DP (Kotsogiannis et al., 2020), asymmetric DP (Takagi et al., 2022), event-level vs user-level DP (Dwork et al., 2010a), and others (see the many examples listed in Section 4 of Desfontaines and Pejó (2020)).

A third branch is concerned with restricting the scope of protection (Subsection 2.4.3). Blowfish privacy addresses this concern by allowing the data custodian to limit the set of potential datasets – that is, it allows the data custodian to specify data universes. In addition to the existing literature on privacy under invariants (Ashmead et al., 2019; Gong and Meng, 2020; Gao et al., 2022; Dharangutte et al., 2023), other work in this branch include conditioned or empirical DP (Abowd et al., 2013; Charest and Hou, 2016), personalized DP (Ebadi et al., 2015; Jorgensen et al., 2015), individual DP (Soria-Comas et al., 2017; Feldman and Zrnic, 2022), bootstrap DP (O’Keefe and Charest, 2019), stratified DP (Bun et al., 2022), per-record DP (Seeman et al., 2023) and per-instance DP (Wang, 2018; Redberg and Wang, 2021). Such work is concerned with the specification of the multiverse \mathcal{D} , and of the privacy loss budget $\varepsilon_{\mathcal{D}}$ as a

function of the universe $\mathcal{D} \in \mathcal{D}$.

A fourth branch is concerned with what the protection objects $x \in \mathcal{X}$ are (Subsection 2.4.2). While DP was traditionally concerned with tabular data (see Definition 2.3.1), $x \in \mathcal{X}$ could instead be, for example, a graph encoding network data (Hay et al., 2009), or geospatial data (Andrés et al., 2013). (Note that alternative data structures often necessitate new choices for $d_{\mathcal{X}}$, not just \mathcal{X} .) Alternatively, \mathcal{X} could be artificially restricted to make DP easier to implement – for example, by assuming the domain of every possible dataset record is $[-a, a]^d$ (for some large a), rather than \mathbb{R}^d . Another possibility is to permit randomness in the confidential dataset, by generalizing the protection domain \mathcal{X} to be a set of probability measures on the data space, rather than the data space itself. Pufferfish privacy (Kifer and Machanavajjhala, 2014; Bailie and Gong, 2024) uses this generalization to encode the background knowledge of an attacker via a probability distribution. Special cases of the same idea can be found in Bhaskar et al. (2011) and Seeman et al. (2022). A related vein of work studies how DP protection varies as the domain \mathcal{X} moves along the data life cycle. (We use the terms ‘data pipeline’, ‘data journey’ and ‘data life cycle’ synonymously, although we prefer the latter since it highlights the circular nature of, and feedback loops inherent to, the processes of data conceptualization, generation, processing, analysis, etc.) This includes the effect of privacy amplification by shuffling (Cheu et al., 2019; Erlingsson et al., 2019; Feldman et al., 2022; Cheu, 2022) and by sampling (Beimel et al., 2010; Balle et al., 2020; Bun et al., 2022), as well as the effects of data pre-processing steps in general (Debenedetti et al., 2024; Hu et al., 2024). However, work studying pre-processing steps is still largely primarily, and a broad understanding of how – for example – various common survey sampling steps (e.g. non-response imputation, editing and survey weight adjustment) alter privacy is still an open area of research (Reiter, 2019; Drechsler, 2023; Das et al., 2022; Bailie and

Drechsler, 2024; Drechsler and Bailie, 2024).

This subsection is a soupçon of the literature on DP which is limited to a cursory discussion on a few of the many research areas in this large and active field – the ‘systematization of knowledge’ article Desfontaines and Pejó (2020) found approximately 225 published DP specifications which are generalizations, relaxations or variants of pure ε -DP. Our intentions here are threefold: to give credit to existing research which informed our understanding of a DP specification as a tuple of five components; to provide an initial demonstration of why this system of five building blocks may be useful for comparing different DP formulations; and to hint at how these formulations all share a common spirit as Lipschitz continuity conditions. This discussion also explains the phrase “stirred, not shaken” as it appears in this paper’s title. As this system of DP specifications is not itself a novel conceptualization, the way that we discuss DP throughout this article stays faithful to its spirit as exemplified in the original formulation of Definition 2.3.1. That is, DP – as a Lipschitz condition – studies the change of the variations in the output statistics with respect to alterations in the input space. Granted this way of thinking, however, we will argue and demonstrate that small distinctions in the choices for the components of a DP specification can make a substantial difference in the quality and the strength of the resulting privacy guarantee in practice.

2.4 A SYSTEM OF DP SPECIFICATIONS

2.4.1 A DP SPECIFICATION: FLAVOR AND INTENSITY

As Section 2.2 overviews, there are five building blocks to a DP specification. We begin with the “who”, “where”, “what”, and “how” questions, which define the flavor of a DP specification.

Definition 2.4.1. A *differential privacy flavor* is a four-tuple $(\mathcal{X}, \mathcal{D}, d_{\mathcal{X}}, D_{\text{Pr}})$ where

1. The *domain* \mathcal{X} is a set, whose elements are called *protection objects*;

2. The *multiverse* $\mathcal{D} \subset 2^{\mathcal{X}}$ is a collection of *universes* $\mathcal{D} \subset \mathcal{X}$;
3. The *input premetric* $d_{\mathcal{X}}$ is a premetric (see Definition 2.4.14 below) on \mathcal{X} ; and
4. The *output premetric* D_{Pr} is a probability premetric (Definition 2.4.16).

A flavor of DP is therefore a distinct set of choices for the domain \mathcal{X} , the multiverse \mathcal{D} , the input premetric $d_{\mathcal{X}}$, and the output premetric D_{Pr} . The domain \mathcal{X} is typically the data space – the space of all (theoretically-conceivable) datasets – but it can take other forms (as explained in the previous subsection). We typically denote protection objects – i.e. elements of \mathcal{X} – by x or x' .

After fixing a particular flavor of DP, the “how much” question becomes relevant, as it affords a measurement of the intensity of privacy protection. Importantly this intensity is relative to the chosen flavor and hence must always be interpreted within its context. The flavor and the intensity together constitutes a DP specification, as defined below.

Definition 2.4.2. A *differential privacy specification* is a quintuple $(\mathcal{X}, \mathcal{D}, d_{\mathcal{X}}, D_{\text{Pr}}, \varepsilon_{\mathcal{D}})$ consisting of a DP flavor $(\mathcal{X}, \mathcal{D}, d_{\mathcal{X}}, D_{\text{Pr}})$ and a *privacy-loss budget* $\varepsilon_{\mathcal{D}} : \mathcal{D} \rightarrow [0, \infty]$ (or *privacy budget* for short). We denote a DP specification by $\varepsilon_{\mathcal{D}}\text{-DP}(\mathcal{X}, \mathcal{D}, d_{\mathcal{X}}, D_{\text{Pr}})$ (following the notational style of [Kifer and Machanavajjhala \(2014\)](#)).

Definition 2.4.3. A *data-release mechanism* (or *mechanism* for short) is a function $T : \mathcal{X} \times \mathcal{U} \rightarrow \mathcal{T}$, along with a probability $\mathbb{P}(U \in \cdot)$ on the (secret) seed $U \in \mathcal{U}$. The probability $\mathbb{P}(U \in \cdot)$ induces a distribution \mathbb{P}_x on \mathcal{T} in the standard way:

$$\mathbb{P}_x(T(x, U) \in S) = \mathbb{P}(U \in \{u \in \mathcal{U} : T(x, u) \in S\}).$$

P_x is the *distribution of the data-release mechanism* $T(x, U)$'s *output*, given a fixed $x \in \mathcal{X}$. (We omit some non-trivial measure-theoretic details of this definition – see Appendix A.1.)

Definition 2.4.4. A data-release mechanism $T : \mathcal{X} \times \mathcal{U} \rightarrow \mathcal{T}$ *satisfies* the DP specification $\varepsilon_{\mathcal{D}}\text{-DP}(\mathcal{X}, \mathcal{D}, d_{\mathcal{X}}, D_{\text{Pr}})$ if, for all data universes $\mathcal{D} \in \mathcal{D}$, and all protection objects $x, x' \in \mathcal{D}$,

$$D_{\text{Pr}}[P_x(T \in \cdot), P_{x'}(T \in \cdot)] \leq \varepsilon_{\mathcal{D}} d_{\mathcal{X}}(x, x'). \quad (2.3)$$

Definition 2.4.5. Given a DP flavor $(\mathcal{X}, \mathcal{D}, d_{\mathcal{X}}, D_{\text{Pr}})$, the *privacy loss* of a data-release mechanism T is the smallest $\varepsilon_{\mathcal{D}}$ such that T satisfies $\varepsilon_{\mathcal{D}}\text{-DP}(\mathcal{X}, \mathcal{D}, d_{\mathcal{X}}, D_{\text{Pr}})$.

Remark 2.4.6. To resolve the edge case where $\varepsilon_{\mathcal{D}} = 0$ but $d_{\mathcal{X}}(x, x') = \infty$, we define $0 \times \infty = \infty$. This means DP never controls the difference between P_x and $P_{x'}$ when $d_{\mathcal{X}}(x, x') = \infty$, even in the case of complete privacy ($\varepsilon_{\mathcal{D}} = 0$).

Remark 2.4.7. In Definition 2.4.4, the domain of the data-release mechanism T and the domain of the DP specification $\varepsilon_{\mathcal{D}}\text{-DP}(\mathcal{X}, \mathcal{D}, d_{\mathcal{X}}, D_{\text{Pr}})$ are exactly the same set \mathcal{X} . This is a necessary precondition for a data-release mechanism to satisfy a DP specification.

Remark 2.4.8. Throughout this article, we use the terms ‘DP definition’ and ‘DP formulation’ informally to refer generally to an existing notion of DP or one of its variants, extensions or generalisations. We reserve the terms ‘DP flavor’ and ‘DP specification’ for the precise meanings given in Definitions 2.4.1 and 2.4.2. A ‘DP mechanism’ is a data-release mechanism which satisfies some (implicit) DP specification. When we speak of the ‘DP guarantee’, or the ‘privacy guarantee’ of a DP mechanism T , we are referring to the assurance that T satisfies the Lipschitz condition (2.3), under a given (implicit) DP specification.

The privacy loss budget $\varepsilon_{\mathcal{D}}$ controls the intensity of protection that is guaranteed by a DP specification. A larger budget allows for mechanisms with less intense protection; whereas a small budget corresponds to requiring a higher intensity of privacy protection.

Note that the privacy loss budget is a component of a DP specification, whereas the privacy loss is an attribute of a DP mechanism. The privacy loss budget should be interpreted as the maximum *possible* privacy loss that is considered acceptable by the data custodian. (The maximum acceptable privacy loss can be in absolute terms, or it can be relative to the statistical accuracy achievable under the corresponding DP specification (Abowd and Schmutte, 2019).) In contrast, the privacy loss of a DP mechanism T is the *actual* reduction in privacy that results from releasing data via T . As with the budget, the privacy loss is always relative to the underlying DP flavor.

The privacy loss budget $\varepsilon_{\mathcal{D}}$ is allowed to vary with the universe \mathcal{D} . That is, $\varepsilon_{\mathcal{D}}$ is a function with domain the multiverse \mathcal{D} . Since each universe is allowed to have its own privacy loss budget, some universes may be afforded more protection than others. This property is important in many applications, such as stratified DP (Bun et al., 2022) or per-record DP (Seeman et al., 2023).

The DP property (2.3) is achieved by injecting artificial noise into the data-release process. By decreasing the dependence of P_x on x , noise injection flattens the ‘derivative’ $\frac{dP_x}{dx}$ of the data-release mechanism and hence one can view DP as a stability, or a robustness, condition (Dwork and Lei, 2009; Avella-Medina, 2020, 2021; Asi et al., 2023; Hopkins et al., 2023).

Definition 2.4.9. Let $\mathcal{M}(\mathcal{X}, \mathcal{D}, d_{\mathcal{X}}, D_{\text{Pr}}, \varepsilon_{\mathcal{D}})$ denote the set of data-release mechanisms which satisfy the DP specification $\varepsilon_{\mathcal{D}}\text{-DP}(\mathcal{X}, \mathcal{D}, d_{\mathcal{X}}, D_{\text{Pr}})$.

We say that one DP specification $\varepsilon_{\mathcal{D}}\text{-DP}(\mathcal{X}, \mathcal{D}, d_{\mathcal{X}}, D_{\text{Pr}})$ is *stronger* than another specification $\varepsilon'_{\mathcal{D}}\text{-DP}(\mathcal{X}, \mathcal{D}, d_{\mathcal{X}}, D_{\text{Pr}})$ if $\varepsilon_{\mathcal{D}} \leq \varepsilon'_{\mathcal{D}}$.

Notation	Description	Reference
\mathcal{X}	The domain	Subsection 2.4.2
$\mathcal{D} \subset 2^{\mathcal{X}}$	The multiverse; a set of subsets of the domain \mathcal{X}	Subsection 2.4.3
$d_{\mathcal{X}} : \mathcal{X} \times \mathcal{X} \rightarrow [0, \infty]$	The input premetric; a premetric on the domain \mathcal{X}	Subsection 2.4.4
$D_{\text{Pr}} : \mathcal{P} \times \mathcal{P} \rightarrow [0, \infty]$	The output premetric; a probability premetric	Subsection 2.4.5
$\varepsilon_{\mathcal{D}} : \mathcal{D} \rightarrow [0, \infty]$	The privacy-loss budget; when $D_{\text{Pr}} \neq D_{\text{MULT}}$, it may be denoted by other Greek letters, e.g. $\rho_{\mathcal{D}}$.	Subsection 2.4.1
$(\mathcal{X}, \mathcal{D}, d_{\mathcal{X}}, D_{\text{Pr}})$	A generic DP flavor	Definition 2.4.1
$(\mathcal{X}, \mathcal{D}, d_{\mathcal{X}}, D_{\text{Pr}}, \varepsilon_{\mathcal{D}})$ or $\varepsilon_{\mathcal{D}}\text{-DP}(\mathcal{X}, \mathcal{D}, d_{\mathcal{X}}, D_{\text{Pr}})$	A generic DP specification	Definition 2.4.2
$\mathcal{M}(\mathcal{X}, \mathcal{D}, d_{\mathcal{X}}, D_{\text{Pr}}, \varepsilon_{\mathcal{D}})$	The set of all data-release mechanisms satisfying the DP specification $\varepsilon_{\mathcal{D}}\text{-DP}(\mathcal{X}, \mathcal{D}, d_{\mathcal{X}}, D_{\text{Pr}})$	Definition 2.4.9

Table 2.1: Notation related to DP flavors and specifications.

$\text{DP}(\mathcal{X}, \mathcal{D}', d'_{\mathcal{X}}, D'_{\text{Pr}})$ if

$$\mathcal{M}(\mathcal{X}, \mathcal{D}, d_{\mathcal{X}}, D_{\text{Pr}}, \varepsilon_{\mathcal{D}}) \subset \mathcal{M}(\mathcal{X}, \mathcal{D}', d'_{\mathcal{X}}, D'_{\text{Pr}}, \varepsilon'_{\mathcal{D}}).$$

2.4.2 THE PROJECTION OBJECTS: THE DOMAIN \mathcal{X}

The domain \mathcal{X} is the set consisting of all the protection objects – that is, all those objects which are eligible for protection by a DP mechanism. Supposing that the protection objects are datasets, \mathcal{X} can be understood before data collection as the set of all potential datasets that are a-priori realizable. (We use the term ‘potential’ in the same sense as ‘potential outcome.’) After the actual dataset has been realized, \mathcal{X} may be interpreted as the set of all possible, counterfactual datasets.

However, this definition of the domain \mathcal{X} is not completely satisfactory because it leaves indeterminate what makes one dataset potentially or counterfactually realizable and another dataset not so. To address

Notation	Description	Reference
$x, x' \in \mathcal{X}$	Two generic protection objects, i.e. elements of the domain \mathcal{X} ; typically (for most DP flavors) x, x' are datasets	Subsections 2.3.1 & 2.4.1
$\mathcal{D} \subset \mathcal{X}$	A universe; an element of the multiverse \mathcal{D}	Subsection 2.4.3
T	A data-release mechanism; a function $\mathcal{X} \times \mathcal{U} \rightarrow \mathcal{T}$; Also denoted by $T(x)$ or $T(x, U)$	Subsection 2.3.1, Definition 2.4.3
\mathcal{T}	The space of all possible outputs; the codomain of the data-release mechanism T	Subsection 2.3.1
t	An element of \mathcal{T} ; a generic, realized output (as opposed to the (unrealized) random variable T)	Subsection 2.3.1
U and \mathcal{U}	The random seed of a data-release mechanism T and its space	Subsection 2.3.1
P, Q	Two generic probability distributions	–
$P(U \in \cdot)$	The probability of the random seed U	Definition 2.4.3
P_x and $P_{x'}$	The probability distributions of $T(x, U)$ and $T(x', U)$, respectively, as induced by the random seed U	Subsection 2.3.1
p_x	The density of P_x (with respect to some dominating measure)	–
$L(x t)$	The likelihood of x given output t ; $p_x(T = t)$	Subsections 2.4.5
(Ω, \mathcal{F})	A generic measurable space where Ω is a set and \mathcal{F} is a σ -algebra on Ω	–
$\mathcal{P}_{(\Omega, \mathcal{F})}$	The collection of all probability measures on (Ω, \mathcal{F})	–
\mathcal{P}	The collection of all probability measures	Definition 2.4.16

Table 2.2: Notation related to data-release mechanisms.

Notation	Description	Reference
$\mathcal{D}(\cdot)$ and $\mathcal{D} = \mathcal{D}(x)$	A universe function $\mathcal{X} \rightarrow 2^{\mathcal{X}}$ and the universe associated to x	Subsection 2.4.3
$c : \mathcal{X} \rightarrow \mathbb{R}^l$	A generic invariant function	Subsection 2.4.3
$\mathcal{D}_c(\cdot) : \mathcal{X} \rightarrow 2^{\mathcal{X}}$ and $\mathcal{D}_c = \{\mathcal{D}(x)\}_{x \in \mathcal{X}}$	The invariant-induced universe function and the corresponding invariant-induced multiverse	Subsection 2.4.3
$d_{\text{HamS}}^u(x, x')$	The Hamming distance on unordered datasets at the resolution u ; u may be, for example, person-records p or household-records h (d_{HamS}^u is an example of $d_{\mathcal{X}}$)	Equation (A.3)
D_{MULT}	The multiplicative distance; the probability premetric used by pure ε -DP	Equation 2.9
D_{NoR}	The normalized Rényi metric; the probability premetric used by ρ -zero concentrated DP	Subsection 2.4.5

Table 2.3: Notation related to various choices for the five building blocks of a DP specification.

this criticism, one may define the domain \mathcal{X} according to a pre-specified database schema. (A schema is a set of logical rules describing the “blueprint” of a database, which includes a declaration of the variables of the database and the set of values each variables can take.) From this perspective, \mathcal{X} is the set of all datasets that are compliant with the given database schema. Yet this definition, while concrete, is not particularly illuminating, since it begs the questions, why is the database schema the way it is, and why this database schema, rather than some other schema?

To fully appreciate the domain \mathcal{X} and its role in DP’s privacy protection, we need to examine the *data life cycle*: the journey of the data within the context of its environment, from conceptualization, generation, collection and processing, through to analysis, visualization, re-use and beyond. (This is not an exhaustive list of the various components of a data life cycle. For a more comprehensive treatment, see [Wing \(2019\)](#), [Leonelli and Tempini \(2020\)](#), [Gitelman \(2013\)](#) and [Borgman \(2019\)](#).)

The data-release mechanism – as a procedure which processes and transforms the data – is an integral phase of the data life cycle. But what steps of the data life cycle are included as part of the data-release mechanism is a subjective decision – one which must be made by the data custodian in the course of implementing DP. In fact, the data custodian is, consciously or unconsciously, forced to categorize the steps of the data life cycle into three consecutive phases: 1) the *data-recording phase*, which comprises the initial steps of the data life cycle up until the data-release mechanism starts; 2) the *data-processing phase*, which consists of the steps that are performed by the data-release mechanism; and 3) the *post-release phase*, which covers all the subsequent steps of the data life cycle. Although it is not always the case (since it depends on the data custodian’s choice of data-release mechanism), the data-recording phases usually includes data conceptualization and generation, as well as the recording (or measurement) step; the data-processing phase might involve data cleaning, preprocessing and wrangling, along with the privacy-noise injection; and the post-release phase may incorporate analysis, visualisation and data re-use.

By delineating where the data-release mechanism ends in the data life cycle, the data custodian defines the *processing-disseminating boundary* – the boundary between the second and third phases. And in deciding where the data-release mechanism starts in the data life cycle, the data custodian defines the *recording-processing boundary*, which demarcates the first and second phases. This boundary provides us with a fully-rigorous definition of the domain \mathcal{X} : Each $x \in \mathcal{X}$ is a potential input to the data-release mechanism and these inputs are, by definition, the outputs of the data-recording phase. Therefore, the domain \mathcal{X} is the set of all a-priori potential, or a-posteriori counterfactual, outputs of the data-recording phase.

While important for the data analyst (as illustrated by the debate surrounding the 2020 US Census “noisy measurement files” (McCartan et al., 2023)), the processing-disseminating boundary is inconse-

quential from the perspective of DP, because the post-processing property (Subsection 2.5) implies that the DP guarantee applies to all data downstream of this boundary. However, the same point does not apply to the recording-processing boundary: moving this boundary upstream – by re-categorizing data-recording steps as data-processing – can dilute, or even invalidate, the DP guarantee (Das et al., 2022; Bun et al., 2022; Hu et al., 2024). In fact, we will see that the recording-processing boundary plays an important role in determining the privacy protection actually afforded by a DP mechanism.

The data life cycle of the 2020 US Decennial Census provides an illustrative example. After respondents provide their information (which, to be clear, is not the start of the life cycle), the resulting data is passed through a number of complex processes, including data coding, editing (to correct implausible or illogical responses) and imputation (to fill in missing responses) (Cantwell, 2021; Ramirez and Borman, 2021; Marks and Rios-Vargas, 2021). The dataset outputted at the end of these processes is termed the Census Edited File (CEF), which is the input to the data-release mechanisms of the 2020 Census (e.g. the TopDown algorithm) (Abowd et al., 2022a). Therefore, the “data-recording” phase of the 2020 Census includes key steps which would ordinarily be classified as data processing, such as imputation and editing.

The reader may wonder why discussion on the processing of Census data is important to privacy. The key realization is that, because it is the dataset $x \in \mathcal{X}$ at the recording-processing boundary which is subject to the Lipschitz condition (2.3), it is this dataset which is protected by a DP mechanism. As such, DP’s semantics (i.e. measures of DP’s protection against an adversary – see Subsection 2.3.2) are framed in terms of inference about this dataset. That is to say, DP semantics assume that the attacker’s target of inference is the data at the recording-processing boundary (or some subset of these data).

While the Lipschitz condition (2.3) holds for data at the recording-processing boundary whenever the data-release mechanism is DP, there is no guarantee that this condition holds for data elsewhere in the data

life cycle. In fact, there are simple examples which illustrate that, without additional assumptions on the data life cycle, a DP mechanism does not ensure that the Lipschitz condition (2.3) holds for data upstream or downstream of the recording-processing boundary: Data downstream of any noise-injection will clearly not satisfy (2.3), at least not with the same privacy loss, and neither will data which is upstream of steps that impact a query’s sensitivity, such as data-dependent clipping⁴ (Kamath et al., 2023) or imputation (Das et al., 2022). (For a more general discussion on this point, see the *linkage inequality* of privacy (Wang et al., 2017).) It follows that only the data at the recording-processing boundary – not data elsewhere in the data life cycle – are directly covered by DP’s privacy semantics.

The domain \mathcal{X} of a DP mechanism’s flavor is therefore important since it specifies which data is directly protected by the mechanism. Returning to the 2020 US Census as an example, the domain of the TDA’s flavor is the set of all possible CEFs, not the set of all respondents’ possible data. As such, it is not the respondents’ data (i.e. their ‘raw’ Census responses) which are directly protected by the TDA, but rather it is the edited and imputed data (i.e. the CEF) which receives the DP guarantee. (Moreover, by their choice of $d_{\mathcal{X}}$, the USCB’s DP analysis treats the CEF *as if* it were the ‘raw’ data – see Subsection 2.4.4.) The TDA’s privacy semantics therefore model an attacker who is interested in learning the edited and imputed responses, not the respondents’ actual answers.

Because pre-processing steps are pervasive in machine learning and data analysis, similar complications also arise in many other scenarios beyond the 2020 Census. In addition to the examples of coding, editing imputation and data-dependent clipping mentioned above, data deduplication, scaling and quantization are common pre-processing steps which can reduce privacy. That is to say, data downstream of these pre-

⁴Clipping is also called winsorizing or top- and bottom-coding, depending on the context.

processing steps will have weaker (or no) privacy guarantees than data upstream; or equivalently, moving these steps from the data-recording phase to the data-processing phase would require additional noise to be added in order to maintain the same level of privacy protection (Debenedetti et al., 2024; Hu et al., 2024).

The data custodian faces a number of important considerations when determining the recording-processing boundary. If there were a point in the data life cycle when the data were ‘raw,’ then one might reasonably suppose that the recording-processing boundary should be drawn at that point. Yet ‘raw data’ is an oxymoron (Bowker, 2005); data never ‘just exist’ but are always manufactured (Gitelman, 2013). Even so, data are generally ‘rawer’ at earlier points in their life cycle. This suggests that to effect better privacy protection a data custodian should set the recording-processing boundary towards the start of the data life cycle. For example, by including editing and imputation inside the data-processing phase, the data custodian protects a form of the data which is a closer representation of the collected data (although perhaps not a closer representation of the respondents’ ‘actual’ – i.e. ‘true’ – information). An extreme example of this approach is Pufferfish privacy (Kifer and Machanavajjhala, 2014), which mandates that the data-generation process is included in the data-processing phase. In this case, the data-recording phase ends before the data is even generated (but crucially, this phase still includes the data conceptualization step), and the protection objects $x \in \mathcal{X}$ are not datasets, but rather they are the potential probability distributions which could (potentially or counterfactually) generate the data (Bailie and Gong, 2024). Alternatively, the goal may often be to protect the ‘untransformed’, ‘unprocessed’ responses of the data providers (when it is clear what are the ‘untransformed’ responses). If so, the recording-processing boundary should be set at point of the life cycle where the data are these ‘untransformed’, ‘raw’ responses.

We have described why the recording-processing boundary should be set early in the data life cycle. However, there are also compelling reasons to set the recording-processing boundary later in the data life cycle. Firstly, some stages of the data life cycle may be out of the data custodian's control. Pragmatically, it may be difficult to incorporate these stages in a DP mechanism. If that is the case, the recording-processing boundary will need to be drawn after these stages. Secondly, DP privacy semantics usually assume implicitly that an attacker does not have knowledge of data intermediate to the data-processing phase. This assumption is reasonable when the data-processing phase is completely contained within a secure computing environment, but it is harder to justify when the phase includes real-world processes such as survey sampling. In fact, if an attacker has auxiliary information about a data-processing step (such as knowledge of who is in the sample), they can partially undo the DP mechanism's protection, rendering the typical DP semantic guarantees invalid (Bailie and Drechsler, 2024). Thirdly, because the recording-processing boundary determines what data is protected by a DP mechanism, there is an ethical consideration when setting this boundary: What form of the data is the custodian obligated to protect? Returning to the US Decennial Census as an example, does the data custodian have a responsibility to protect the data as it is reported by respondents (in which case the data-processing phase should begin immediately after data collection and include editing and imputation), or should the data custodian protect their best guess of the respondents' actual data (in which case the data-processing phase should only begin after processing the raw responses into the data custodian's best guess of the actual data and thus editing and imputation should be included in the data-recording phase)? In summary, it is debatable where the data custodian should draw the recording-processing boundary. As such, the data custodian's segmentation of the data life cycle into the three phases of recording, processing and post-release is a value-laden, but necessary,

designation.

To conclude this subsection, we reiterate three points. Firstly, the domain \mathcal{X} of a DP flavor matters because it encodes where the recording-processing boundary is drawn. Secondly, the recording-processing boundary is important because a DP mechanism protects the data as it is conceived at this boundary. In the terminology of Seeman and Susser (2023), this is one of the *framing effects* of a DP flavor. (In fact, we will see throughout this section that all four components of a flavor contribute to its framing effects.) Thirdly, understanding the recording-processing boundary is crucial because this places the protection objects $x \in \mathcal{X}$ within the broader context of their data life cycle. Contextualization determines how the protection objects $x \in \mathcal{X}$ are given meaning and value (Leonelli, 2019). It defines how x relates to real-world entities (e.g. the survey respondents) and what real-world quantities x measures (e.g. the respondents’ characteristics). Therefore, the choice of \mathcal{X} , as the set of possible data within the context of the recording-processing boundary of the data life cycle, partly determines the substance of a DP flavor’s protection by specifying “who” and “what form and kind of their data” is eligible for protection.

2.4.3 THE SCOPE OF PROTECTION: THE MULTIVERSE \mathcal{D}

The multiverse \mathcal{D} is a collection of subsets of the domain \mathcal{X} . The elements of \mathcal{D} are the potential universes \mathcal{D} , and each universe \mathcal{D} is a set of mutually-plausible datasets. (For ease of interpretability, we assume throughout this subsection that the protection objects $x, x' \in \mathcal{X}$ are datasets, although the discussion generalizes beyond this case.)

For example, in the 2020 U.S. Decennial Census, any applicable DP flavor must entertain universes which are delineated by values for the state population totals, as mandated by the U.S. Constitution (see Theorem 3.3.1 specifically and Proposition 2.4.11 below more generally). Any particular combination

of state population totals would be accidental upon observation: there does not exist a universal law that dictates that these totals could not have been realized in a different way. On the other hand, that the published Decennial Census accords to the known state population totals is essential to the Census because it is made to happen by design. It is this essential quality (hence the multiverse \mathcal{D}), rather than any accidental observation (i.e. any particular universe \mathcal{D} attained), that must be respected by an applicable DP mechanism. Therefore, the multiverse commands careful treatment in a DP specification, to ensure that no information of accidental nature leaks through privacy protection, such as those illustrated in [Gong and Meng \(2020\)](#) which render the celebrated composition property of differential privacy inapplicable.

One may see the distinction between the multiverse and a universe as analogous to the difference between the conditional probability $P(\cdot|Y)$ conditioning on a random variable Y and the conditional probability $P(\cdot|Y = y)$ conditioning on an event $\{Y = y\}$. The random variable Y is the essential quality and the event $\{Y = y\}$ an accidental observation; in the same way, a DP specification is concerned with the essential nature of a data-release mechanism T , rather than the properties of T within some particular, but accidental, universe \mathcal{D} .

A DP specification requires that, for every data universe $\mathcal{D} \in \mathcal{D}$, the Lipschitz continuity condition must hold in the space \mathcal{D} – but not necessarily in the space \mathcal{X} . This distinction is crucial because x typically has high dimension, and we could plausibly demand for Lipschitz continuity along every dimension – i.e. we could require that the ‘derivative’ $\nabla_v P_x$ is bounded by ε in every ‘direction’ $v = \frac{x' - x}{d_{\mathcal{X}}(x, x')}$. (Here we write the derivative as the gradient $\nabla_v P_x$ to emphasise the intuition that the derivative can be taken in multiple directions.) By limiting the choices of x' , restricting \mathcal{D} decreases the directions v for which the directive $\nabla_v P_x$ is bounded. Therefore, reducing the size of the universes strictly weakens the privacy protection in

two senses: 1) it limits the datasets x that are protected to those in some universe \mathcal{D} ; and 2) it reduces the protection afforded to each of these datasets $x \in \mathcal{D}$ by limiting comparisons to other datasets x' which are also in \mathcal{D} . This is the sense in which \mathcal{D} defines the scope of the protection given by a DP flavor.

In practice, it is often the case that the multiverse \mathcal{D} is induced via a set-valued function that we call the *universe function* $\mathcal{D}(\cdot) : \mathcal{X} \rightarrow 2^{\mathcal{X}}$, which associates every potential dataset $x \in \mathcal{X}$ with a universe $\mathcal{D} = \mathcal{D}(x) \subset \mathcal{X}$. In this case, the multiverse $\mathcal{D} = \{\mathcal{D}(x)\}_{x \in \mathcal{X}}$ is the image of the universe function \mathcal{D} .

An important class of universe functions encodes *invariants*: exact quantities calculated from the confidential dataset. Due to legal and policy mandates, or other guidance, invariants are published as-is. From the perspective of data utility, invariants are thus restrictions on the output of a mechanism. Conversely, from the perspective of data privacy, invariants are restrictions on the input, or more exactly, on the data multiverse \mathcal{D} . For this work, we are particularly interested in universe functions of the form

$$\mathcal{D}_c(x) = \{x' \in \mathcal{X} : c(x') = c(x)\}, \quad (2.4)$$

for a given deterministic function $c : \mathcal{X} \rightarrow \mathbb{R}^l$. Here the function c describes the features $c(x)$ of the dataset x which are taken to be invariant. We call $\mathcal{D}_c(\cdot)$ the *invariant-induced universe function* and its image $\{\mathcal{D}_c(x)\}_{x \in \mathcal{X}}$ the *invariant-induced multiverse* \mathcal{D}_c .

Note that invariants of this form define an equivalence relation \sim over its domain \mathcal{X} , defined by $x \sim x'$ if $c(x) = c(x')$. Hence, the data universe function (2.4) induces a *partition* of \mathcal{X} indexed by the image of the invariant function c .

Example 2.4.10. Let the dataset be an contingency table of $m \times n$ records taking non-negative integer values: $\mathcal{X} = (\mathbb{N}^+)^{m \times n}$. Suppose the function $c : (\mathbb{N}^+)^{m \times n} \rightarrow (\mathbb{N}^+)^{m+n}$ tabulates the column- and

row-margins:

$$c(x) = \left(\sum_{i=1}^m x_{i1}, \dots, \sum_{i=1}^m x_{in}, \sum_{j=1}^n x_{1j}, \dots, \sum_{j=1}^n x_{mj} \right). \quad (2.5)$$

The data curator may treat the column- and row-margins of the confidential dataset as invariant. This would be equivalent to employing the universe function $\mathcal{D}_c(\cdot)$ as defined in (2.4) using the function c from (2.5), since $\mathcal{D}_c(\cdot)$ ensures that only pairs of datasets x, x' with the same column- and row-margins are subject to the Lipschitz condition (2.3) of the DP specification.

In some applications, there are also inequality invariants (Abowd et al., 2022a). As an example of such an invariant, the 2020 US Decennial Census requires that the reported number of group quarters in any geographical unit is at most the number of persons in that unit. More generally, an inequality invariant is of the form $f(x) \leq 0$ for some function $f: \mathcal{X} \rightarrow \mathbb{R}$. Such an invariant can be incorporated in our framework by defining

$$c(x) = \begin{cases} 1 & \text{if } f(x) \leq 0, \\ 0 & \text{if } f(x) > 0. \end{cases} \quad (2.6)$$

While weakening the Lipschitz condition (2.4.4) via a non-vacuous multiverse \mathcal{D} leads to a reduction in actual privacy protection, this complication is necessary in many real-world applications of DP. (A multiverse is *vacuous* if, for all distinct $x \neq x' \in \mathcal{X}$ with $d_{\mathcal{X}}(x, x') < \infty$ or $d_{\mathcal{X}}(x', x) < \infty$ there exists a universe $\mathcal{D} \in \mathcal{D}$ with $x, x' \in \mathcal{D}$.) In addition to the examples from the literature referenced in Subsection 2.3.2, we prove in Subsection 3.3 that an invariant-induced multiverse \mathcal{D}_c is necessary for describing the DP guarantee of the 2020 U.S. Census. Furthermore, the practice of empirically restricting the data universe is typical in statistical disclosure control and data analysis more broadly. Top-coding – where one sets a maximum limit on a continuous variable, usually after looking at the raw data – is one common

example.

In the following two propositions, we will see that the interpretation of the value of ε cannot be isolated from the multiverse \mathcal{D} , and indeed this complicates the comparison of privacy loss budgets across different applications. For these two results, fix a domain \mathcal{X} and invariants $c : \mathcal{X} \rightarrow \mathbb{R}^l$.

Proposition 2.4.11. *For any $d_{\mathcal{X}}$ and D_{Pr} , the mechanism $T(x) = c(x)$ that releases the invariants exactly satisfies $\varepsilon_{\mathcal{D}}\text{-DP}(\mathcal{X}, \mathcal{D}_c, d_{\mathcal{X}}, D_{\text{Pr}})$ with privacy loss budget $\varepsilon_{\mathcal{D}} = 0$ for all $\mathcal{D} \in \mathcal{D}$.*

Now suppose $D_{\text{Pr}}(\mathbf{P}, \mathbf{Q}) = \infty$ if $d_{\text{TV}}(\mathbf{P}, \mathbf{Q}) = 1$.⁵ Let \mathcal{D} be a multiverse such that there exists some universe $\mathcal{D}_0 \in \mathcal{D}$ and some $x, x' \in \mathcal{D}_0$ with $d_{\mathcal{X}}(x, x') < \infty$ and $c(x) \neq c(x')$. Then T does not satisfy $\varepsilon_{\mathcal{D}}\text{-DP}(\mathcal{X}, \mathcal{D}, d_{\mathcal{X}}, D_{\text{Pr}})$ whenever $\varepsilon_{\mathcal{D}_0} < \infty$.

The results of Proposition 2.4.11 also hold if \mathcal{D} is any multiverse with c constant within every universe $\mathcal{D} \in \mathcal{D}$ (i.e. if $c(x) = c(x')$ for all $x, x' \in \mathcal{D}$ and all $\mathcal{D} \in \mathcal{D}$). The following result is the converse of Proposition 2.4.11.

Proposition 2.4.12. *Suppose that a mechanism T varies within some universe $\mathcal{D}_0 \in \mathcal{D}$ in the sense that there exists $x, x' \in \mathcal{D}_0$ with $d_{\mathcal{X}}(x, x') < \infty$ but $\mathbf{P}_x \neq \mathbf{P}_{x'}$. When D_{Pr} is a metric, T satisfies $\varepsilon_{\mathcal{D}}\text{-DP}(\mathcal{X}, \mathcal{D}, d_{\mathcal{X}}, D_{\text{Pr}})$ only if $\varepsilon_{\mathcal{D}_0} > 0$.*

These two propositions demonstrate that in order to formulate DP with invariants c , it is necessary and sufficient to limit the Lipschitz condition via the invariant-induced universe function \mathcal{D}_c . Necessity follows from the second half of Proposition 2.4.11: if there are datasets $x, x' \in \mathcal{D}_0$ with different values on

⁵We write d_{TV} to denote the total variation distance (or statistical distance). $d_{\text{TV}}(\mathbf{P}, \mathbf{Q}) = 1$ means that the probability measures \mathbf{P} and \mathbf{Q} have no common support. The assumption $d_{\text{TV}}(\mathbf{P}, \mathbf{Q}) = 1 \Rightarrow D_{\text{Pr}}(\mathbf{P}, \mathbf{Q}) = \infty$ is satisfied by most common choices of D_{Pr} .

the invariants, then releasing the invariants exactly would require $\varepsilon_{\mathcal{D}} = \infty$. Sufficiency is described in two parts: 1) the invariants can be released exactly without privacy loss (by the first half of Proposition 2.4.11); but 2) any additional information (not logically equivalent to the invariants) cannot be released without incurring privacy loss (by Proposition 2.4.12).

As a concrete example of how the meaning of the privacy loss budget $\varepsilon_{\mathcal{D}}$ changes with \mathcal{D} , consider evaluating the same mechanism T against two DP flavors $(\mathcal{X}, \mathcal{D}_c, d_{\mathcal{X}}, D_{\text{Pr}})$ and $(\mathcal{X}, \mathcal{D}_{c'}, d_{\mathcal{X}}, D_{\text{Pr}})$ which differ only on their invariants. Suppose the second set of invariants are nested within the first; that is, c is strictly more constraining than c' . (For example, c are population counts at the block level and c' are counts at the county level.) Then T 's privacy loss $\varepsilon'_{\mathcal{D}'}$ under $\mathcal{D}_{c'}$ cannot be smaller and may be strictly larger than T 's loss $\varepsilon_{\mathcal{D}}$ under \mathcal{D}_c , for any $\mathcal{D}' \subset \mathcal{D}$. We formalize this statement in Proposition 2.4.13.

As we repeatedly emphasize, it is dangerous to think that the c -release is indeed afforded with less privacy protection than the c' -release because there is privacy leakage due to specifying additional invariants, which is not captured by the “within-system” privacy evaluation ε . Indeed in the extreme example where c is an injective function so that the universes \mathcal{D} are singletons, there is no privacy protection afforded by the DP specification $\varepsilon_{\mathcal{D}}\text{-DP}(\mathcal{X}, \mathcal{D}_c, d_{\mathcal{X}}, D_{\text{Pr}})$ regardless of the choices of $d_{\mathcal{X}}$, D_{Pr} and $\varepsilon_{\mathcal{D}}$. This point is crucial to understanding the comparative analysis between the PSA and the 2020 TDA as presented in Section 3.4.

Proposition 2.4.13. *Suppose that \mathcal{D} and \mathcal{D}' are nested in the sense that, for all $\mathcal{D}' \in \mathcal{D}'$, there exists some $\mathcal{D} \in \mathcal{D}$ such that $\mathcal{D}' \subset \mathcal{D}$. (That is, \mathcal{D}' is a refinement of \mathcal{D} .) Then, for all privacy loss budgets $\varepsilon'_{\mathcal{D}'} : \mathcal{D}' \rightarrow [0, \infty]$, we have*

$$\mathcal{M}(\mathcal{X}, \mathcal{D}, d_{\mathcal{X}}, D_{\text{Pr}}, \varepsilon_{\mathcal{D}}) \subset \mathcal{M}(\mathcal{X}, \mathcal{D}', d_{\mathcal{X}}, D_{\text{Pr}}, \varepsilon'_{\mathcal{D}'}), \quad (2.7)$$

where $\varepsilon_{\mathcal{D}} = \inf\{\varepsilon'_{\mathcal{D}'} : \mathcal{D}' \in \mathcal{D}' \text{ with } \mathcal{D}' \subset \mathcal{D}\}$ is the budget under \mathcal{D} . Further, (2.7) holds for all privacy

loss budgets $\varepsilon_{\mathcal{D}} : \mathcal{D} \rightarrow [0, \infty]$ when $\varepsilon'_{\mathcal{D}'} = \inf\{\varepsilon_{\mathcal{D}} : \mathcal{D} \in \mathcal{D} \text{ with } \mathcal{D}' \subset \mathcal{D}\}$ is the budget under \mathcal{D}' .

This proposition illustrates that refining the data multiverse weakens the protection provided by a DP specification. As we have stated, this is (intuitively) because reducing the comparisons between protection objects reduces the dimensions along which Lischitz continuity is required. The operation of refining the data multiverse is mathematically equivalent to redefining $d_{\mathcal{X}}$, as shown by the following proposition.

2.4.4 THE PROTECTION UNITS: THE INPUT PREMETRIC $d_{\mathcal{X}}$

Because information is generated by variations – and differential privacy is the control of variations – we need a way of measuring how P_x varies as $x \in \mathcal{X}$ is altered. To do so mathematically, we first need to have a measure of change in $x \in \mathcal{X}$. In the context of DP, the general notion of *premetric* is useful.

Definition 2.4.14 (premetric). A *premetric* d on a set S is a function $S \times S \rightarrow [0, \infty]$ satisfying $d(x, y) = 0$ if $x = y$.

A premetric generalizes the mathematical concept of a metric, because it is not required to be symmetric, positive for distinct points, nor does it need to satisfy the triangle inequality. While many of the premetrics used in DP are metrics – such as the multiplicative distance adopted in the original formulation of DP (Dwork et al., 2006b) – the more general notion of premetric is needed because it has been used in forming various approximations to the multiplicative distance, as we shall see in the next subsection, and because it has been used in generalizing Dwork et al. (2006b)’s notion of neighboring datasets, as we explore in this subsection.

The input premetric $d_{\mathcal{X}}$ of a DP flavor $(\mathcal{X}, \mathcal{D}, d_{\mathcal{X}}, D_{\text{Pr}})$ is a mathematical tool which encodes the flavor’s protection units – that is, the entities that an attacker is assumed to be interested in, or the entities

that the data custodian wants to directly protect. Although common narratives focus on an individual’s personal information as the adversary’s target, many applications are concerned with protecting the privacy of other kinds of units, such as households, businesses, messages, transactions or even single attributes. In the literature, the protection units are ordinarily implied by the *neighboring* datasets, that is, the pairs of datasets that “differ by one entry” (as stated in Definition 2.3.1). Here “an entry” refers to the data associated with a single unit and “differ by one entry” refers to changing a unit’s attributes as recorded in the dataset, or to changing the indicator of whether a unit is included in the dataset or not. In this way, exactly one protection unit’s data changes when x is counterfactually altered to a neighboring x' .

The input premetric $d_{\mathcal{X}}$ provides a way to encode neighboring datasets: $x, x' \in \mathcal{X}$ are neighbors if and only if $d_{\mathcal{X}}(x, x') = 1$. Hence, $d_{\mathcal{X}}$ also provides a way to describe the protection units: A *protection unit* is an entity i for which there exists $x, x' \in \mathcal{X}$ with $d_{\mathcal{X}}(x, x') = 1$ such that x and x' differ only on i ’s data. Put simply, a protection unit is an entity whose data is the difference between neighboring datasets.

While we have described how to derive the protection units from the input premetric $d_{\mathcal{X}}$, typically in practise, the protection units are actually chosen first. In the process of implementing DP, the data custodian will usually determine the entities that they want to directly protect. These entities will then serve as the protection units. As mentioned earlier, they could be persons, businesses, households, interactions with a website or service, etc. After deciding upon the protection units, the data custodian will then select an input premetric which encodes their choice of protection units. By ‘encoding protection units’, we mean that this input premetric $d_{\mathcal{X}}$ satisfies the requirement: $d_{\mathcal{X}}(x, x') = 1$ if and only if x and x' differ on a single protection unit’s data. For example, if the data custodian’s desired protection units are individual persons, then they could select the Hamming distance d_{Ham}^p on person-records as their input premetric.

The Hamming distance $d_{\text{Ham}}^p(x, x')$ on person-records between two equal-length databases x and x' is the number of records – where each person is associated with a single record – which differ between x and x' . (See Appendix A.4 for a precise definition, along with the definitions of other common choices for $d_{\mathcal{X}}$.) Hence $d_{\text{Ham}}^p(x, x') = 1$ if and only if x and x' differ on the attributes of one individual, which is one of the data custodian’s protection units.

In the above discussion, we did not explain what is meant by ‘an entity’s data.’ Since this is critical for understanding what protection units are encoded by an input premetric, and hence what entities are protected by a DP mechanism, this deserves some attention. In simple cases, it is obvious which entities each data point concerns. Hence, a rudimentary definition of ‘entity i ’s data’ is all the data points which are concerned with i .

In more complex cases, it may not be clear which entities a data point concerns. This can make it difficult to determine the protection units. To illustrate this complexity, we return to the 2020 US Decennial Census. As discussed in Subsection 2.4.2, the domain \mathcal{X} is the set of all Census Edited Files (CEFs), and each CEF $x \in \mathcal{X}$ can be represented as a dataset of person-records. The input premetric $d_{\mathcal{X}}$ is the Hamming distance d_{Ham}^p on these person-records (Abowd et al., 2022a). Therefore, one might naturally suppose that the protection units are individuals. However, in producing the CEF, the Census Bureau copies the records of a *donor* household to impute the missing records of a nearby, *recipient* non-responding household (Ramirez and Borman, 2021). In this way, the data of a donor individual can be associated with multiple person-records in the CEF: the donor’s original record; and zero, one or more imputed records. Plausibly, an imputed record may concern either 1) the recipient only (since it contains the geographical information of the recipient, not the donor, and since it may be considered as the Census Bureau’s best

guess of the recipient’s data); 2) the donor only (since it is a copy of the donor’s response); or 3) both the recipient and the donor (by a combination of the previous two rationales). The protection units of the 2020 Census are individual persons only in the first of these three possibilities, because, in the other two cases, $d_{\text{Ham}}^p(x, x') > 1$ whenever x and x' differ on a donor’s data, and so donor individuals cannot be protection units.

Since in certain scenarios it is unclear which entities a data point concerns, a more sophisticated definition of ‘an entity’s data’ is required. At the core of DP’s philosophy is the conceptualization of privacy as indistinguishability: under DP, an attacker should not be able to distinguish between two protection objects x and x' (which are in the same universe and have smaller input premetric). Equivalently, DP is concerned with masking the difference between such x and x' . If we want to mask an entity, what really matters is the influence of the entity on the protection objects. Therefore, we properly define ‘an entity’s data’ as the data points which change when the entity counterfactually alters their behavior or attributes during the data-recording phase. By the definition of the protection units above, it follows that the protection units are those entities with the following property:

There exist two counterfactual runs of the data-recording phase which are the same in every regard except that the entity behaves differently (or has different attributes) in these two runs. Suppose these two runs produce the protection objects x and x' . Then $d_{\mathcal{X}}(x, x') = 1$.

In making this definition, we must assume that the probability distributions of the protection units are mutually independent (except for deterministic or logical constraints between units, such as those in graph data). This allows a protection unit to change their behavior without affecting the rest of the data-recording phase, except in deterministic ways. (If non-deterministic dependencies were allowed, the resulting outputs x and x' would not even be well-defined.) This is a necessary assumption to avoid modelling the data-recording phase, as is typically desired. Yet it is also why a generalized strong adversary assump-

tion (or one of its relaxations) is necessary for translating DP specifications into privacy semantics (see Subsection 2.3.2).

The above definition of ‘an entity’s data’ resolves the ambiguity in the 2020 Census’s protection units. A donor can behave differently in the data-recording phase by altering their response to the Census. Their behavior affects multiple records – their own and their recipients’ – and hence the resulting CEFs x and x' would have $d_{\text{Ham}}^p(x, x') > 1$. The 2020 Census’s protection units are thus ‘post-imputation persons,’ those (fictional) entities with data that is exactly one record in the CEF, so that the attributes of a ‘post-imputation person’ can be altered in the data-recording phase without affecting other records in the CEF.

The distinction between ‘post-imputation persons’ and real world persons is important. Notably, a donor individual’s privacy is not protected at the nominal level given by the privacy loss budget of the 2020 Census. In fact, because the Hamming distance induced by persons as protection units is smaller than d_{Ham}^p , donor individuals receive strictly less protection than prima facie indicated: a donor’s privacy loss is increased by a factor equal to the number of their recipients, as demonstrated by Proposition 2.4.15 below. The important realization is that d_{Ham}^p treats the CEF as if it consists of ‘raw’ Census responses; yet an examination of the data-recording phase belies this misconception. Fundamentally, an input premetric $d_{\mathcal{X}}$ measures changes in data – but, as we discuss extensively in [Bailie et al. \(2025e\)](#), data are *accidental* representations of *essential* information ([Robertson Ishii and Atkins, 2023](#)). As such, what really matters is the information encoded in the data, not the data itself. If a premetric $d_{\mathcal{X}}$ allows for the manipulation of data values without respecting the underlying data-generating process, it contradicts the information in the data and so cannot encode real-world protection units, but only fictitious ones. As we also saw in Subsection 2.4.2, there is therefore the potential for a disconnect between DP’s protection of data values

and the protection of an individual unit’s information. We re-emphasize Subsection 2.4.2’s point of the necessity of placing the data in the context of its life cycle in order to understand this disconnect and, hence, to understand the privacy actually afforded by a DP mechanism.

The protection unit is a useful concept for translating the input premetric $d_{\mathcal{X}}$ into a measure of real-world privacy protection. Yet $d_{\mathcal{X}}$ cannot always be interpreted in terms of protection units. There are cases where there are no protection objects $x, x' \in \mathcal{X}$ with $d_{\mathcal{X}}(x, x') = 1$; and there are cases where a value of $d_{\mathcal{X}}(x, x') = 1$ is not imbued with any special significance (see Chatzikokolakis et al. (2013) and (Desfontaines and Pejó, 2020, Section 4.3), as well as the references in Subsection 2.3.2). However, there is a second, more general, interpretation of the input premetric $d_{\mathcal{X}}$, which is more broadly applicable than the notion of the protection units: $d_{\mathcal{X}}$ is the yardstick (i.e. unit of measurement) against which the rate of change in output variations is measured. By this statement, we mean that $d_{\mathcal{X}}$ serves the role of the denominator in the ‘derivative’ $\frac{dP_x}{dx}$. Mathematically, $d_{\mathcal{X}}$ specifies which counterfactual input alterations are permissible – an alteration of $x \in X$ is permissible whenever it results in some x' with $d_{\mathcal{X}}(x, x') < \infty$ – while also quantifying these alterations according to the corresponding value of $d_{\mathcal{X}}(x, x')$. This quantification is the yardstick of data alterations – “the dx ” – against which changes in the output variations – “the dP_x ” – are benchmarked.

Moreover, shrinking (or enlarging) this yardstick corresponds to decreasing (or increasing) the *granularity* of privacy protection. That is, decreasing the value of $d_{\mathcal{X}}(x, x')$ – which intuitively corresponds to increasing the size of the protection units – translates to protecting larger changes in x . For example, the data custodian might want household-level data to be the protection units, instead of person-level data. This would be equivalent to using an input premetric $d_{\mathcal{X}}^h$ with the property $d_{\mathcal{X}}^h(x, x') = 1$ whenever x

and \mathcal{X}' differ on the data belonging to a single household. This household-level premetric $d_{\mathcal{X}}^b$ is smaller than the analogous person-level premetric $d_{\mathcal{X}}^p$:

$$d_{\mathcal{X}}^b \leq d_{\mathcal{X}}^p \leq b d_{\mathcal{X}}^b,$$

where b is an upper bound on the number of people in a single household. Hence, a unit with respect to $d_{\mathcal{X}}^b$ can correspond to b units with respect to $d_{\mathcal{X}}^p$. By decreasing the granularity, more data is being protected per unit change in the yardstick $d_{\mathcal{X}}$. This results in a stronger DP flavor, as the following proposition demonstrates:

Proposition 2.4.15. *Fix a domain \mathcal{X} and let $d_{\mathcal{X}}, d'_{\mathcal{X}}$ be two input premetrics satisfying*

$$l d_{\mathcal{X}} \leq d'_{\mathcal{X}} \leq u d_{\mathcal{X}},$$

for some constants $0 < l \leq u \leq \infty$. Then

$$\mathcal{M}(\mathcal{X}, \mathcal{D}, d_{\mathcal{X}}, D_{\text{Pr}}, \varepsilon_{\mathcal{D}}) \subset \mathcal{M}(\mathcal{X}, \mathcal{D}, d'_{\mathcal{X}}, D_{\text{Pr}}, \varepsilon_{\mathcal{D}}/l),$$

and

$$\mathcal{M}(\mathcal{X}, \mathcal{D}, d'_{\mathcal{X}}, D_{\text{Pr}}, \varepsilon_{\mathcal{D}}) \subset \mathcal{M}(\mathcal{X}, \mathcal{D}, d_{\mathcal{X}}, D_{\text{Pr}}, u \varepsilon_{\mathcal{D}}).$$

Thus, the input premetric $d_{\mathcal{X}}$ provides a notion of *granularity* (or resolution) of the protection afforded by a DP specification $\varepsilon_{\mathcal{D}}\text{-DP}(\mathcal{X}, \mathcal{D}, d_{\mathcal{X}}, D_{\text{Pr}})$. A specification with lower granularity is stronger: If $d_{\mathcal{X}} \leq d'_{\mathcal{X}}$, then $\varepsilon_{\mathcal{D}}\text{-DP}(\mathcal{X}, \mathcal{D}, d_{\mathcal{X}}, D_{\text{Pr}})$ is a stricter condition than $\varepsilon_{\mathcal{D}}\text{-DP}(\mathcal{X}, \mathcal{D}, d'_{\mathcal{X}}, D_{\text{Pr}})$. The intuitive justification of this result is that $\varepsilon_{\mathcal{D}}\text{-DP}(\mathcal{X}, \mathcal{D}, d_{\mathcal{X}}, D_{\text{Pr}})$ protects more data per unit change in $d_{\mathcal{X}}$ than $\varepsilon_{\mathcal{D}}\text{-DP}(\mathcal{X}, \mathcal{D}, d'_{\mathcal{X}}, D_{\text{Pr}})$.

In addition to $d_{\mathcal{X}}$'s ability to generalize the notion of neighboring datasets (as in the examples presented

in Subsection 2.3.2), another advantage of using a divergence $d_{\mathcal{X}}$ in the definition of DP – instead of the notion of neighboring datasets – is that it enables the quantification of privacy protection beyond what is required by the original DP formulation of [Dwork et al. \(2006b\)](#). Comparing Definition 2.3.1 with Definition 2.4.2, we can see that the former places restrictions only on datasets x and x' that are neighbors (that is, only when $d_{\mathcal{X}}(x, x') = 1$), while the latter holds for all pairs regardless of the value of $d_{\mathcal{X}}(x, x')$. However, this does not necessarily imply that the latter definition is a more stringent requirement. A simple way to see this is that we can always define $d_{\mathcal{X}}(x, x') = \infty$ when x and x' are not neighbors, in which case Definition 2.4.2 would be satisfied trivially for any non-neighboring x and x' .

One real value of using $d_{\mathcal{X}}$ instead of the notion of neighbors perhaps is best seen by considering what will happen for datasets x and x' which differ by k entries, under the original Definition 2.3.1. For simplicity, let's assume $k = 2$, and $x = \{x_1, x_2, c\} \in \mathcal{D}$ and $x' = \{x'_1, x'_2, c\} \in \mathcal{D}$, where c represents the common part of x and x' . Let $\tilde{x} = \{x_1, x'_2, c\}$ and assume it is also in \mathcal{D} (this is not a trivial assumption in general). Then because \tilde{x} is a neighbor of both x and x' , we have from (2.2) that

$$\left| \ln \frac{\mathbb{P}(T(x, U) = t)}{\mathbb{P}(T(x', U) = t)} \right| \leq \left| \ln \frac{\mathbb{P}(T(x, U) = t)}{\mathbb{P}(T(\tilde{x}, U) = t)} \right| + \left| \ln \frac{\mathbb{P}(T(\tilde{x}, U) = t)}{\mathbb{P}(T(x', U) = t)} \right| \leq \varepsilon + \varepsilon = 2\varepsilon. \quad (2.8)$$

We can see that the factor 2 corresponds to $k = 2$, and the above derivation is easily replicated with general $k \in \mathbb{N}$. Hence when we adopt the Hamming distance for $d_{\mathcal{X}}$, we see that (2.2) holds for all (legitimate) pairs, as long as we replace ε by $d_{\mathcal{X}}(x, x')\varepsilon$, which is a special case of (2.3). (This is related to the concept of group privacy; see [Dwork and Roth \(2014\)](#).) Two other values of using $d_{\mathcal{X}}$ instead of neighboring datasets is 1) in applications where there is no appropriate notion of neighbors, for example, in geospatial data ([Andrés et al., 2013](#)) and 2) when there are invariants (or more generally when there is a non-vacuous multiverse), because invariants often imply that there are no pairs of datasets x and x' in the same data

universe \mathcal{D} with $d_{\mathcal{X}}(x, x') = 1$ and hence restricting to such datasets would result in a vacuous DP flavor.

In a nutshell, using a premetric $d_{\mathcal{X}}$ makes explicit the protection bounds for all pairs of datasets, without losing the intuition of neighbors as the protection units. It also allows us to explicitly write a DP specification as a Lipschitz condition, thereby explicating that the essence of DP is to limit the change of variations per unit change in x , where “unit change” is precisely defined by $d_{\mathcal{X}}$ taking the unit value – one.

2.4.5 THE STANDARD OF PROTECTION: THE OUTPUT PREMETERIC D_{Pr}

Having formulated changes in the input x via $d_{\mathcal{X}}$, we now explicate the measure of change in the output variations. Output variations are captured by the probability distribution of the data-release mechanism T given an input x , so we need to measure change in terms of these probability distributions. Pure ε -DP (Definition 2.3.1) uses the multiplicative distance D_{MULT} (which is also termed the max-divergence in some contexts) to measure this change:

$$D_{\text{MULT}}(P, Q) = \begin{cases} \sup_{S \in \mathcal{F}} \left| \ln \frac{P(S)}{Q(S)} \right| & \text{if } P \text{ and } Q \text{ are on the same measurable space } (\Omega, \mathcal{F}), \\ \infty & \text{otherwise,} \end{cases} \quad (2.9)$$

where we define $\frac{P(S)}{Q(S)} = 1$ when $P(S) = Q(S) = 0$. (The multiplicative distance is strongly equivalent (in the sense of metrics) to the density-ratio metric (Wasserman, 1992).) For $b < \infty$, the condition $D_{\text{MULT}}(P, Q) \leq b$ is equivalent to P and Q being mutually absolutely continuous and having densities (with respect to a common dominating measure μ) whose ratio is μ -a.e. bounded by $\exp(b)$ (Bailie and Gong, 2024). When P and Q are probabilities $P_x, P_{x'}$ for the output of T , this condition translates to bounding the log-likelihood ratio (or, in other terms, the log-Bayes factor) between x and x' . That is, pure

ε -DP is equivalent to the condition:

$$\ln \left[\frac{L(x | t)}{L(x' | t)} \right] \leq \varepsilon d_{\mathcal{X}}(x, x'),$$

uniformly for almost all $t \in \mathcal{T}$ and all $x, x' \in \mathcal{X}$, where $L(x | t) = p_x(T = t)$ is the likelihood function. (Here p_x is the density of P_x .) The use of the multiplicative distance to formulate a mathematical formalization of privacy is therefore justified by the law of likelihood (Hacking, 1965), which asserts that this likelihood ratio $L(x | t)/L(x' | t)$ is the degree to which the output t supports the hypothesis that the true input dataset is x rather than x' . By bounding this likelihood ratio, pure ε -DP restricts the degree of support for x against x' to at most $\exp[\varepsilon d_{\mathcal{X}}(x, x')]$.

Yet requiring that this property hold for all outputs t – even those t which are highly-improbable – is a stringent condition. For additive noise, it requires a fat-tailed noise distribution, with density $e^{-O(|x|)}$, which rules out Gaussian noise for example. This motivates the study of different ways to measure the change in the output variations. We capture these various different ways with the unifying concept of a ‘probability premetric’:

Definition 2.4.16. Let $\mathcal{P}_{(\Omega, \mathcal{F})}$ be the collection of all probability measures on the measurable space (Ω, \mathcal{F}) . Let $\mathcal{P} = \bigcup_{(\Omega, \mathcal{F})} \mathcal{P}_{(\Omega, \mathcal{F})}$ (where the union is over all measurable spaces (Ω, \mathcal{F})) be the collection of all probability measures. A *probability premetric* D_{Pr} is a function $\mathcal{P} \times \mathcal{P} \rightarrow [0, \infty]$ satisfying

- $D_{\text{Pr}}(P, Q) = 0$ if $P = Q$; and
- $D_{\text{Pr}}(P, Q) = \infty$ if P and Q are on different measurable spaces (that is, $P \in \mathcal{P}_{(\Omega, \mathcal{F})}$ and $Q \in \mathcal{P}_{(\Omega', \mathcal{F}')}$ with $(\Omega, \mathcal{F}) \neq (\Omega', \mathcal{F}')$).

Most distances and divergences encountered in statistics and probability theory (including the total

variation distance, the KL-divergence, the Rényi divergences, the Hellinger distance, the χ^2 -divergence, the integral probability metrics and the Wasserstein distances) are probability premetrics.

Many of the most popular variants of DP simply replace the multiplicative distance D_{MULT} for some other probability premetric D_{Pr} . (Note that technically these variants refer to families of DP flavors, since they leave the other three building blocks unspecified.) For example, approximate (ε, δ) -DP (Dwork et al., 2006a) uses the δ -approximate multiplicative premetric $D_{\text{MULT}}^\delta(P, Q)$ for D_{Pr} :

$$D_{\text{MULT}}^\delta(P, Q) = \sup_{S \in \mathcal{F}} \left\{ \ln \frac{[P(S) - \delta]^+}{Q(S)}, \ln \frac{[Q(S) - \delta]^+}{P(S)}, 0 \right\},$$

for $P, Q \in \mathcal{P}_{(\Omega, \mathcal{F})}$, where $[x]^+ = \max\{x, 0\}$. (Clearly $D_{\text{MULT}}^\delta(P, Q)$ reduces to $D_{\text{MULT}}(P, Q)$ when $\delta = 0$, but for $\delta > 0$, it is generally not a metric. D_{MULT}^δ is the symmetrization of the δ -approximate max-divergence (Dwork et al., 2010b).) And ρ -zero concentrated differential privacy (ρ -zCDP) (Bun and Steinke, 2016) uses the *normalized Rényi metric* D_{NoR} for D_{Pr} :

$$D_{\text{NoR}}(P, Q) = \sup_{\alpha > 1} \frac{1}{\sqrt{\alpha}} \max \left\{ \sqrt{D_\alpha(P||Q)}, \sqrt{D_\alpha(Q||P)} \right\},$$

where D_α is the Rényi divergence of order α :

$$D_\alpha(P||Q) = \begin{cases} \frac{1}{\alpha-1} \ln \int \left[\frac{dP}{dQ} \right]^\alpha dQ, & \text{if } P \text{ is absolutely continuous wrt. } Q, \\ \infty & \text{otherwise.} \end{cases}$$

Here $\frac{dP}{dQ}$ is the Radon-Nikodym derivative of P with respect to Q . (Note that we reparameterize ρ so that D_{NoR} is a metric (Bailie et al., 2025a). This is similar to the parameterization of zCDP given in Canonne et al. (2022). The original parameterization ρ in Bun and Steinke (2016) is equivalent to ρ^2 in our formulation of zCDP.)

Conceivably, when defining a DP specification, any probability premetric D_{Pr} could be used, such as the total variation distance or a Wasserstein distance. But some choices of D_{Pr} are more appropriate for capturing the notion of data privacy than others. For one, DP flavors are commonly required to be well-behaved, in the sense that 1) the aggregate privacy loss increases smoothly with additional DP data releases; and that 2) transformations of the released data do not have weaker DP guarantees than the released data itself. We will show in Subsection 2.5 that these properties of a DP flavor – called ‘closure under composition’ and ‘immunity to post-processing’ respectively – are consequences of the choice of D_{Pr} . (For example, a DP flavor with any of the above choices for D_{Pr} will always satisfy composition and post-processing.) It is natural to select a probability premetric D_{Pr} which ensures the composition and post-processing axioms are satisfied. This is one argument for using some probability premetrics D_{Pr} over others.

More fundamentally, DP’s goal is to limit a specific type of inferential disclosure and the probability premetric D_{Pr} plays a crucial role in doing so. For example, all of the above probability premetrics are concerned with the likelihood ratio $L(x \mid t)/L(x' \mid t)$ in some way and hence they limit likelihood-based inference (such as hypothesis testing and most Bayesian inference). (For specifics, see the references on DP semantics given in Subsection 2.3.2.) Adopting choices of D_{Pr} that do not involve the likelihood ratio should require similar justification in terms of their impact on inferential disclosures.

By comparing P_x and $P_{x'}$, the probability premetric D_{Pr} specifies *how* to measure differences in variations in the output t . Because inference is derived from the information inherent in these variations, D_{Pr} is the standard of privacy protection afforded by a DP specification.

2.4.6 MULTI-PARAMETER BUDGETS

Many existing DP definitions, such as (ε, δ) -DP, have multi-parameter privacy loss budgets. There are two options for incorporating these definitions into our system. Firstly, all but one parameter may be included inside D_{Pr} . This is the approach taken in the previous section when we showed that the probability premetric D_{Pr} corresponding to (ε, δ) -DP is D_{MULT}^δ .

The second option is to generalize the definition of a probability premetric to have a multidimensional codomain $[0, \infty]^d$, and replace the inequality \leq in the Lipschitz condition (2.3) with the pointwise partial order on $[0, \infty]^d$. (Recall that this partial order is given by $x \preceq y$ if $x_i \leq y_i$ for all i .) This is the approach taken in f -DP (Dong et al., 2022), where $D_{\text{Pr}}(P, Q)$ is the tradeoff function $Tr(P, Q) \in [0, 1]^{[0,1]}$. (Recall that the tradeoff function $Tr(P, Q)$ maps $\alpha \in [0, 1]$ to the infimum type II error over all level- α tests of P (the null) versus Q (the alternative).) Under this approach, $(\varepsilon_0, \delta_0)$ -DP would correspond to the probability premetric D_{MULT} which maps a pair $(P, Q) \in \mathcal{P}^2$ to the function $\delta \mapsto D_{\text{MULT}}^\delta(P, Q)$, along with the privacy loss budget $\varepsilon(\cdot) \in [0, \infty]^{[0,1]}$ given by

$$\varepsilon(\delta) = \begin{cases} \varepsilon_0 & \text{if } \delta = \delta_0, \\ \infty & \text{otherwise.} \end{cases}$$

Using a multidimensional probability premetric does not affect our formulation of DP specifications as Lipschitz conditions. For simplicity, we will not consider multidimensional probability premetrics further, although all ideas in this paper can be naturally extended to account for this generalisation.

2.5 POST-PROCESSING AND COMPOSITION

Two common desiderata for a DP flavor are immunity to post-processing and closure under composition. This subsection will express these two properties using the vocabulary of Section 2.4. We will then show that these desiderata are consequences of the choice of D_{Pr} (except in trivial edge cases).

Closure under composition means that the overall privacy loss smoothly increases as more DP outputs are released. For simplicity, we present the case where the increase in privacy loss is linear:

Definition 2.5.1. A DP flavor $(\mathcal{X}, \mathcal{D}, d_{\mathcal{X}}, D_{\text{Pr}})$ is *closed under linear self-composition* if the following statement holds: For all data-release mechanisms

$$T_1 : \mathcal{X} \times \mathcal{U}_1 \rightarrow \mathcal{T}_1,$$

$$T_2 : \mathcal{X} \times \mathcal{U}_2 \rightarrow \mathcal{T}_2,$$

and all privacy loss budgets $\varepsilon_{\mathcal{D}}^{(1)}, \varepsilon_{\mathcal{D}}^{(2)}$, if T_1 and T_2 both satisfy the DP flavor $(\mathcal{X}, \mathcal{D}, d_{\mathcal{X}}, D_{\text{Pr}})$ with budgets $\varepsilon_{\mathcal{D}}^{(1)}$ and $\varepsilon_{\mathcal{D}}^{(2)}$ respectively, then the composition mechanism (formally defined in Appendix A.5),

$$(T_1, T_2) : \mathcal{X} \times \mathcal{U}_1 \times \mathcal{U}_2 \rightarrow \mathcal{T}_1 \times \mathcal{T}_2,$$

$$(x, (u_1, u_2)) \mapsto (T(x, u_1), T(x, u_2)),$$

also satisfies $(\mathcal{X}, \mathcal{D}, d_{\mathcal{X}}, D_{\text{Pr}})$, with budget $\varepsilon_{\mathcal{D}}^{(1)} + \varepsilon_{\mathcal{D}}^{(2)}$.

Often a DP flavor is closed under the composition of data-release mechanisms T_1 and T_2 only when the mechanisms' seeds U_1, U_2 are independent (Bailie and Drechsler, 2024). When the composition (T_1, T_2) satisfies $(\varepsilon_{\mathcal{D}}^{(1)} + \varepsilon_{\mathcal{D}}^{(2)})$ -DP $(\mathcal{X}, \mathcal{D}, d_{\mathcal{X}}, D_{\text{Pr}})$ for T_1 and T_2 with independent seeds, we say that the DP flavor $(\mathcal{X}, \mathcal{D}, d_{\mathcal{X}}, D_{\text{Pr}})$ is *closed under linear, fresh self-composition*. (For a discussion of why non-fresh composi-

tion is important, see Appendix A.5.)

Immunity to post-processing means that any transformation (i.e. post-processing) of a DP output is also DP, with the same privacy guarantee:

Definition 2.5.2. A DP flavor $(\mathcal{X}, \mathcal{D}, d_{\mathcal{X}}, D_{\text{Pr}})$ is *immune to post-processing* if the following statement holds: For all data-release mechanisms $T : \mathcal{X} \times \mathcal{U} \rightarrow \mathcal{T}$, all functions $f : \mathcal{T} \rightarrow \mathcal{T}'$ and all privacy loss budgets $\varepsilon_{\mathcal{D}}$, if T satisfies $\varepsilon_{\mathcal{D}}\text{-DP}(\mathcal{X}, \mathcal{D}, d_{\mathcal{X}}, D_{\text{Pr}})$ then the post-processing mechanism $f \circ T : \mathcal{X} \times \mathcal{U} \rightarrow \mathcal{T}'$ (formally defined in Appendix A.5) also satisfies $\varepsilon_{\mathcal{D}}\text{-DP}(\mathcal{X}, \mathcal{D}, d_{\mathcal{X}}, D_{\text{Pr}})$.

Post-processing and fresh composition are properties of the output premetric D_{Pr} only. That is, (except for trivial edge cases) whether or not a DP flavor $(\mathcal{X}, \mathcal{D}, d_{\mathcal{X}}, D_{\text{Pr}})$ is immune to post-processing or closed under fresh composition depends only on D_{Pr} , as illustrated by the following two propositions.

Recall that $\mathcal{P}_{(\Omega, \mathcal{F})}$ is the set of all probability measures defined on the measurable space (Ω, \mathcal{F}) .

Proposition 2.5.3. Fix a probability premetric D_{Pr} . The DP flavor $(\mathcal{X}, \mathcal{D}, d_{\mathcal{X}}, D_{\text{Pr}})$ is closed under linear, fresh self-composition for all choices of \mathcal{X}, \mathcal{D} and $d_{\mathcal{X}}$ if and only if, for all $\mathbf{P}, \mathbf{Q} \in \mathcal{P}_{(\Omega, \mathcal{F})}$ and all $\mathbf{P}', \mathbf{Q}' \in \mathcal{P}_{(\Omega', \mathcal{F}')}$,

$$D_{\text{Pr}}(\mathbf{P} \times \mathbf{P}', \mathbf{Q} \times \mathbf{Q}') \leq D_{\text{Pr}}(\mathbf{P}, \mathbf{Q}) + D_{\text{Pr}}(\mathbf{P}', \mathbf{Q}'). \quad (2.10)$$

Proposition 2.5.4. Fix a probability premetric D_{Pr} . The DP flavor $(\mathcal{X}, \mathcal{D}, d_{\mathcal{X}}, D_{\text{Pr}})$ is immune to post-processing for all choices of \mathcal{X}, \mathcal{D} and $d_{\mathcal{X}}$ if and only if, for all $\mathbf{P}, \mathbf{Q} \in \mathcal{P}_{(\Omega, \mathcal{F})}$, and all measurable $f : (\Omega, \mathcal{F}) \rightarrow (\Omega', \mathcal{F}')$,

$$D_{\text{Pr}}[\mathbf{P}, \mathbf{Q}] \geq D_{\text{Pr}}[f_{\star}(\mathbf{P}), f_{\star}(\mathbf{Q})], \quad (2.11)$$

where $f_{\star}(\mathbf{P})$ is the push-forward probability $f_{\star}(\mathbf{P})(S) = \mathbf{P}[f^{-1}(S)]$.

When f is a randomized function, (2.11) is the data-processing inequality for D_{Pr} (Cover and Thomas, 2005). Immunity to post-processing by a random f is implied by non-random post-processing (Definition 2.5.2) by first composing T with the random seed of f , which has zero privacy loss since it is constant in x (see Appendix A.5).

Blackwell’s theorem establishes a connection between post-processing and the tradeoff function Tr , which is defined as follows: For $P, Q \in \mathcal{P}(\Omega, \mathcal{F})$, the tradeoff $Tr(P, Q)(\alpha)$ is the supremal power of a level- α test between P (the null) and Q (the alternative). A recent result of Su (2024) shows that (2.11) implies D_{Pr} is a function of Tr . We extend this result by proving the converse, thereby establishing another characterization of post-processing:

Theorem 2.5.5. *Fix a probability premetric D_{Pr} . The DP flavor $(\mathcal{X}, \mathcal{D}, d_{\mathcal{X}}, D_{\text{Pr}})$ is immune to (randomized) post-processing for all choices of \mathcal{X}, \mathcal{D} and $d_{\mathcal{X}}$ if and only if $D_{\text{Pr}} = \lambda \circ Tr$ for some non-decreasing function λ .*

It is straightforward to verify that $D_{\text{Pr}} = D_{\text{MULT}}$ satisfies (2.11) and (2.10). Hence, the DP flavor $(\mathcal{X}, \mathcal{D}, d_{\mathcal{X}}, D_{\text{MULT}})$ is always immune to post-processing and closed under linear, fresh self-composition. The DP flavor $(\mathcal{X}, \mathcal{D}, d_{\mathcal{X}}, D_{\text{NoR}})$ is also immune to post-processing and closed under fresh self-composition. However, instead of linear composition $\varepsilon_{\mathcal{D}}^{(1)} + \varepsilon_{\mathcal{D}}^{(2)}$, the privacy loss of (T_1, T_2) is bounded by $\sqrt{\rho_1^2 + \rho_2^2}$ when T_i satisfies ρ_i -DP $(\mathcal{X}, \mathcal{D}, d_{\mathcal{X}}, D_{\text{NoR}})$ (assuming T_1 and T_2 have independent seeds).

We leave to other work discussion on the composition of multi-parameter flavors – such as $(\mathcal{X}, \mathcal{D}, d_{\mathcal{X}}, D_{\text{MULT}}^{\delta})$, which measures its privacy loss budget in terms of ε and δ (Dwork et al., 2010b; Kairouz et al., 2017; Steinke, 2022) – and the composition of multiple DP flavors (e.g. what flavor does $T = (T_1, T_2)$ satisfy for mechanisms T_1 and T_2 which satisfy different DP flavors?). This second topic is important for

quantifying the total privacy loss across multiple DP data releases in the likely scenario that these releases do not all share the same DP flavor. (For example, different preprocessing procedures may be used for the different data release, so that the DP flavors have different values for their domain \mathcal{X} .) In this scenario, self-composition does not apply and so new results on the composition of multiple flavors will be needed in order to assess an individual's aggregate privacy loss.

3

Invariant-Preserving Deployments of Differential Privacy for the US Decennial Census¹

3.1 MOTIVATIONS AND CONTRIBUTIONS

3.1.1 DATA PRIVACY WITH INVARIANT CONSTRAINTS

IN 2018, THE UNITED STATES CENSUS BUREAU (USCB) announced an overhaul of its disclosure avoidance system (DAS) (Abowd, 2018). The DAS for the 1990, 2000 and 2010 US Decennial Censuses primarily consisted of a *data swapping* method (Dalenius and Reiss, 1982; Fienberg and McIntyre, 2004), which permuted the geographical data of a randomly selected subset of households (McKenna, 2018). The protection provided by this statistical disclosure control (SDC) method has traditionally been justified with

¹Based on work coauthored with Ruobin Gong and Xiao-Li Meng.

intuitive arguments. In contrast, the DAS for the 2020 Census would be redesigned from the ground up with the primary goal of supplying a mathematical guarantee of protection. Moreover, this guarantee, the USCB decided (Abowd, 2018), must be some type of *differential privacy* (DP) (Dwork et al., 2006b) – a large family of technical standards (Desfontaines and Pejó, 2022) which characterize the ‘privacy’ of an SDC method in terms of its sensitivity to counterfactual changes in its input data.

However, there were other priorities for the 2020 Census, some of which complicated a straightforward adoption of DP. In particular, state population counts are legislatively required to be published exactly as counted, whereas, DP – at least as originally defined in Dwork et al. (2006b) – requires that these counts be infused with random noise. The USCB’s TopDown Algorithm (TDA) addresses this conflict by first adding DP-calibrated noise to all of the 2020 Census data and then removing this noise from a set of key statistics, called *invariants*, via a complex optimization procedure (Abowd et al., 2022a). These invariants include not only the state population totals but also the counts of housing units at the lowest level of Census geography, amongst other statistics (see Table 3.4). (More generally, invariants refer to any summaries of the confidential data that are released without modification.)

A complete and rigorous assessment of TDA’s mathematical guarantee of protection must address the entire procedure, including both the noise infusion in the first step and the noise removal due to the invariant constraints in the second step. A guarantee for the first step is easy to determine, because it follows an established DP method for adding noise to the confidential counts (Canonne et al., 2022). However, the second step, as the Census Bureau’s own assessment makes clear (Ashmead et al., 2019), is particularly challenging to analyze because it does not immediately align with standard formulations of DP, including those which the 2020 DAS invoked (zero-concentrated DP), referenced (approximate DP), or at some

point considered (pure DP) (Abowd et al., 2022a).

The issue is that these formulations of DP do not permit consideration of the multiple counterfactual data universes induced by TDA's invariants. Yet, as we will see, to spell out these universes is key to articulating TDA's actual guarantee of SDC. This same point also applies to any data swapping algorithm, which by definition keeps invariant all counts that are unaffected by its swapping operation. However, these invariants are inevitably much more numerous than the TDA's – an important observation when comparing data swapping with the TDA because, as the number of invariants increases, their impact ranges from negligible to completely nullifying any supposed guarantee of protection. Nevertheless, conceptually and mathematically, all invariants can be handled in a unified way, making it possible to compare different invariant-preserving SDC methods within the same theoretical system.

3.1.2 A SYSTEM OF DP SPECIFICATIONS

Intuitively speaking, the impact of invariants on SDC is similar to conditioning in statistical inference, that is, constraining the possible states of the confidential data by known or assumed information. Indeed, the procedure for infusing invariants into a DP formulation parallels the process of disintegrating a probability into a collection of conditional probabilities. The overall mathematical notion of probability (or of DP) remains the same; the difference is that now there are multiple probabilities (or DP formulations) – one for each possible value of the conditioning random variable (or invariant) – each living on a restricted space.

Even setting aside the adjustments required for invariants, the plethora of existing DP formulations differ across several other dimensions (Desfontaines and Pej6, 2022). As such, DP can vary widely in form and spirit (Dwork et al., 2019), making it difficult to 1) understand what it means for an SDC method to be DP and 2) objectively compare different DP deployments in a systematic way – two tasks which are

central to the goals of this paper. Our own confusion regarding this state of affairs motivated us to explicate a system of *DP specifications*, which we describe in a companion article titled “A Refreshment Stirred, Not Shaken (I): Five Building Blocks of Differential Privacy” (Bailie et al., 2025b). (We will refer to this article as Part I hereafter.)

The phrase, “a refreshment stirred, not shaken,” is intended to emphasize that the system described in Part I is not new, but simply a synthesis of existing literature. Indeed, this system is in essence the formalization of three principles which we believe are widely accepted in the DP community. The first principle states that a DP formulation is a technical standard which requires the rate of change, or ‘derivative,’ of an SDC method to be controlled (hence the epithet ‘differential’). The second principle asserts that the rate of change of an SDC method is defined as the change in the *probability distribution* of the method’s output, per unit change in its input data. And the third principle observes that different DP formulations correspond to different choices for how and where to measure these changes, as well as how much to control the associated rate of change. We call these choices the *building blocks* of DP. Part I identifies five such building blocks:

- The *Domain* \mathcal{X} : a set of datasets x ;
- The *Multiverse* $\mathcal{D} \subset 2^{\mathcal{X}}$: a set of universes $\mathcal{D} \subset \mathcal{X}$;
- The *Input Premetric* $d_{\mathcal{X}}$: a dissimilarity measure on \mathcal{X} ;
- The *Output Premetric* $D_{\mathcal{P}}$: a dissimilarity measure between probability distributions; and
- The *Privacy-Loss Budget* $\varepsilon_{\mathcal{D}}$: a function $\mathcal{D} \rightarrow [0, \infty]$.

A thesis of the system developed in Part I is that formulating DP requires instantiating choices for each of these five building blocks. A collection of such choices is called a *DP specification*, while choices for the first four building blocks define a *DP flavor*.

Definition 3.1.1 (Definition 2.4.4). A *data-release mechanism* (aka SDC method) $T : \mathcal{X} \times \mathcal{U} \rightarrow \mathcal{T}$ satisfies the DP specification $\varepsilon_{\mathcal{D}}\text{-DP}(\mathcal{X}, \mathcal{D}, d_{\mathcal{X}}, D_{\text{Pr}})$ if, for all universes $\mathcal{D} \in \mathcal{D}$ and all $x, x' \in \mathcal{D}$,

$$D_{\text{Pr}} \left[\mathbb{P}(T(x, U) \in \cdot), \mathbb{P}(T(x', U) \in \cdot) \right] \leq \varepsilon_{\mathcal{D}} d_{\mathcal{X}}(x, x'), \quad (3.1)$$

where \mathbb{P} is the probability distribution induced by the random seed $U \in \mathcal{U}$, taking x or x' as fixed.

From the perspective provided by the system of DP specifications, there are two possible ways that invariants can naturally be integrated with DP. Firstly, one can set $d_{\mathcal{X}}(x, x') = \infty$ if x and x' disagree on the invariants. (This shows that the formulations of DP which use the Hamming distance d_{Ham} have as invariant the dataset size because $d_{\text{Ham}}(x, x') = \infty$ whenever x and x' have a different number of records.) The second approach for encoding invariants into a DP specification is through the multiverse \mathcal{D} . Specifically, for invariant statistics $c : \mathcal{X} \rightarrow \mathbb{R}^k$, define the *invariant-induced universe function*

$$\mathcal{D}_c(x) = \{x' \in \mathcal{X} : c(x) = c(x')\}, \quad (3.2)$$

and the *invariant-induced multiverse* $\mathcal{D}_c = \{\mathcal{D}_c(x) : x \in \mathcal{X}\}$. Either of these two approaches has the same result: it ensures that the Lipschitz condition (3.1) only bounds the dissimilarity between $\mathbb{P}(T(x, U) \in \cdot)$ and $\mathbb{P}(T(x', U) \in \cdot)$ when x and x' agree on the invariants. Intuitively, this is the same as conditioning on the invariants, except that a-priori we do not know the realized value of the invariants and hence the DP specification must account for all possible values through the multiverse \mathcal{D}_c , rather than conditioning on any one particular universe $\mathcal{D}_c(x)$. As a concrete example, a DP specification for the TDA should

guarantee SDC regardless of what the actual value of the US total population count turns out to be. As such, requiring the Lipschitz condition (3.1) to hold for any particular universe would be insufficient, since each universe corresponds to a specific total population count, along with specific values for the TDA’s other invariants. Ensuring the Lipschitz condition (3.1) holds for all possible population counts is a key rationale for the concept of the multiverse \mathcal{D} , over and above that of a single universe \mathcal{D} .

3.1.3 PAPER CONTRIBUTIONS AND ORGANIZATION

Section 3.2 introduces data swapping and determines the invariants induced by it. It formally defines the Permutation Swapping Algorithm (PSA), which is a data swapping method with similarities to the 2010 Census DAS. We prove that the PSA satisfies a DP specification (Theorem 3.2.4), and we show how its privacy loss is determined by the swap rate and the maximal size of the swapping strata. The PSA’s DP specification is stated formally in Subsection 3.2.3, after we first define some necessary notation, but intuitively it can be thought of as ‘ ϵ -DP (Dwork et al., 2006b) subject to the invariants induced by the PSA.’ While this means the PSA’s specification differs from conventional formulations of DP, this specification nevertheless provides a mathematical description of the SDC provided by the PSA. As such, this result adds to a growing body of literature showing that – even though they were designed without DP in mind – traditional SDC methods can still be fruitfully analyzed from the perspective of DP (Rinott et al., 2018; Bailie and Chien, 2019; Chien and Sadeghi, 2024; Neunhoffer et al., 2024).

In Section 3.3, we move to the 2020 Census and subject zero-concentrated DP (zCDP) (Bun and Steinke, 2016) to the TDA’s invariants, thereby deriving a DP specification for the TDA. Theorem 3.3.1 proves that the TDA satisfies this DP specification, while also showing that this specification is tight in the sense that the TDA can only satisfy DP subject to its invariants.

Broadening our scope, Section 3.4 compiles DP specifications for all the primary 2020 Census data products, which we aggregate into a single specification covering the 2020 Census as a whole. This specification describes the SDC protection afforded to the 2020 Census data across all major 2020 releases. Naturally, this begs comparison to the counterfactual scenario in which the PSA was used for protecting the 2020 Census. As such, we also supply a DP specification for this counterfactual scenario.

Other contributions of this paper include: an application of the PSA to the 1940 Census (Subsection 3.2.4); an estimate of the DP specification associated with the 2010 Census under the assumption that the PSA was used as the 2010 DAS (Subsection 3.2.5); a summary of the mechanics of the TDA, focusing on aspects which are salient to SDC (Section 3.3); a determination of the *protection* (or *privacy*) *units* for the 2020 Census as ‘post-imputation persons’ with a discussion of why this matters (Subsection 3.4.4); and an exposition of the 2010 DAS through a comprehensive review of publicly-available information (Appendix B.6), along with a comparison between the 2010 DAS and the PSA (Subsection B.6.1) and a discussion of ways the PSA could be modified to further align with the 2010 DAS while still preserving its DP flavor (Subsection B.6.2). Background on data swapping and other related work are provided in Appendices B.1 and B.2 respectively.

3.2 A DP ANALYSIS OF DATA SWAPPING

3.2.1 DATA SWAPPING

Given a dataset x , partition its set of variables V into two non-empty subsets: the *swapping variables* V_{Swap} and the *holding variables* V_{Hold} . A data swapping algorithm randomly selects some records of x and interchanges the values of their swapping variables V_{Swap} . (The values of their holding variables V_{Hold} remain the same.) This creates a new dataset consisting of individual records whose V_{Hold} values are as originally

observed and whose V_{Swap} values are possibly different. The exact procedure for selecting records and interchanging their V_{Swap} values varies between different data swapping methods.

Sometimes, swapping is restricted to records which share the same values on a subset of the holding variables V_{Hold} , called the *matching variables* V_{Match} . Also referred to as the *swap key* (McKenna, 2018; Abowd and Hawes, 2023), the matching variables are often important characteristics of the data population, as swapping records with different V_{Match} values is prohibited. Whenever V_{Match} is nonempty, records are partitioned into strata according to their V_{Match} values and data swapping is repeated independently within each stratum.

Example 3.2.1. This example is a simplification of the disclosure avoidance system (DAS) for the 2010 US Decennial Census. Represent the 2010 Census data as a list of household records, whose variables include all the household’s characteristics, as well as the questionnaire responses from each individual associated with that household. The matching variables V_{Match} (i.e., swap key) include both the number of voting age persons and the total number of persons in the household. V_{Match} also includes a geographic variable V_g (see US Census Bureau (2021e)), either the Census tract, county or state of the household. (The exact choice of V_g has never been made public by the USCB.) V_{Swap} are the geographic variables nested underneath V_g . For example if V_g is the county, then V_{Swap} is the block and tract of the household. All other variables belong to V_{Hold} – in particular, the household and person characteristics. One can imagine the 2010 DAS as digging up pairs of houses of the same size in the same geographic area and swapping their locations but not changing the houses and their occupants. In the 2010 DAS, each household is assigned a risk score based on the USCB’s assessment of how unique the household is within its neighbourhood. These risk scores are used to compute each household’s probability of being swapped. Every

(non-imputed) household has a non-zero swap probability. Selected households are then swapped with one of their neighbours. (See Appendix B.6 for a detailed description of the 2010 DAS and references for this information.)

3.2.2 WHAT INVARIANTS DOES SWAPPING PRESERVE?

Swapping is, very loosely, a synthetic data generation mechanism. Given a dataset x as input, swapping produces a ‘privacy enhanced’ version Z of x . Both x and Z contain the same variables as well as the same number of records. Hence, the invariants of swapping are determined by examining what swapping does, and does not, change in the data.

Consider the dataset x as a matrix whose rows correspond to the records of x and whose columns correspond to the variables V of x . Without loss of generality, the holding variables are ordered before the swapping variables so that x can be partitioned as $[x_{\text{Hold}}, x_{\text{Swap}}]$. A swapping algorithm randomly selects a permutation σ of the rows of x and interchanges the rows of the matrix x_{Swap} according to σ . This operation yields x_{Swap}^σ , whose i th row is given by the $\sigma(i)$ -th row of x_{Swap} . This defines the swapped dataset Z as the matrix $[x_{\text{Hold}}, x_{\text{Swap}}^\sigma]$, and the swapping mechanism releases as its output the fully-saturated contingency table generated by Z .

One can see that after swapping, any statistic generated by only the matrix x_{Hold} is invariant. Moreover, since V_{Match} is identical among swapped records, any statistic generated by only x_{Match} and x_{Swap} is also preserved by swapping. Only statistics that depend nontrivially on both variables V_{Swap} and $V_{\text{Hold}} \setminus V_{\text{Match}}$ can be altered by swapping.

Proposition 1. *Suppose that $V_{\text{Hold}} \setminus V_{\text{Match}}$ and V_{Swap} are non-empty. Then, without loss of generality, we may assume that each of V_{Match} , V_{Swap} and $V_{\text{Hold}} \setminus V_{\text{Match}}$ are univariate. Denote a value of the matching*

variable V_{Match} by m . Similarly, let h and s be values of $V_{\text{Hold}} \setminus V_{\text{Match}}$ and V_{Swap} respectively.

Disregarding the ordering of records, the dataset x can be represented as a 3-dimensional contingency table $H(x) = [n_{mbs}^x]$ of counts in each combination of possible values for m , h and s . (We will omit the superscript x when it is clear from the context.) In general, interior cell counts n_{mbs} are not preserved under swapping and neither are the margins $n_{\cdot bs} = \sum_m n_{mbs}$. But swapping does keep $n_{m \cdot s} = \sum_b n_{mbs}$ and $n_{mb \cdot} = \sum_s n_{mbs}$ invariant.

Proof. First we justify why we can assume that V_{Match} , V_{Swap} and $V_{\text{Hold}} \setminus V_{\text{Match}}$ are univariate (i.e. that these variable sets are singletons). If V_{Match} is empty, replace it with a set consisting of a new variable taking the same value on every record. And if either of V_{Match} , V_{Swap} or $V_{\text{Hold}} \setminus V_{\text{Match}}$ has more than one variable, then cross-classify these variables into a single variable. Neither of these two operations will change the behavior of a swapping method, so we may use them to ensure V_{Match} , V_{Swap} and $V_{\text{Hold}} \setminus V_{\text{Match}}$ are univariate.

Since every permutation σ can be written as the composition of swaps (i.e. 2-cycles), it suffices to show that all possible swaps preserve $n_{m \cdot s}$ and $n_{mb \cdot}$ but not necessarily $n_{\cdot bs}$. A swap pairs a record a in categories mbs with a record b in $mb's'$. It moves a to $mb's$ and b to $mb's'$. The matching category m is the same in a and b by construction. Unless $m = m'$ or $s = s'$, after the swap n_{mbs} and $n_{mb's'}$ decrease by one, and $n_{mb's}$ and $n_{mb's'}$ increase by one. Hence, $n_{m \cdot s}$ and $n_{mb \cdot}$ remain unchanged but $n_{\cdot bs}$ changes whenever $h \neq b'$ and $s \neq s'$. □

Example 3.2.1 (continued). In the 2010 US Census DAS, the number of adults, children and households in each block are invariant. (This is the $n_{m \cdot s}$ margin.) The counts of all the person and household characteristics inside each V_g are also invariant. (This is the $n_{mb \cdot}$ margin.) For example, if V_g is the county, then

the aggregate characteristics at the county level remain unchanged by swapping, but these aggregates at the block and tract level are perturbed.

Definition 3.2.2. Under the setup of Proposition 1, define the *swapping invariants* $c_{\text{Swap}}(x)$ for a given choice of V_{Match} , V_{Swap} and V_{Hold} as the vector of all margins $n_{mb\cdot}$ and $n_{m\cdot s}$, for all possible values of m , b and s . For example, if V_{Match} , $V_{\text{Hold}} \setminus V_{\text{Match}}$ and V_{Swap} take values in $\{1, \dots, \mathcal{M}\}$, $\{1, \dots, \mathcal{H}\}$ and $\{1, \dots, \mathcal{S}\}$ respectively, then

$$c_{\text{Swap}}(x) = \begin{bmatrix} n_{11\cdot} & n_{12\cdot} & \cdots & n_{\mathcal{M}\mathcal{H}\cdot} & n_{1\cdot 1} & n_{1\cdot 2} & \cdots & n_{\mathcal{M}\cdot \mathcal{S}} \end{bmatrix}^T.$$

As the following example illustrates, we do not have complete flexibility in choosing the invariants of swapping.

Example 3.2.3. In the 2020 TDA, there are three invariants: 1) the number of people in each state; 2) the number of housing units in each block; and 3) the count of each type of occupied group quarters (e.g. residence halls, nursing facilities, prisons) in each block (US Census Bureau, 2021d). We cannot design a swapping algorithm which preserves these – and only these – invariants. In other words, the 2020 US Census invariants do not correspond to any swapping invariants c_{Swap} , regardless of the choice of V_{Match} , V_{Swap} and V_{Hold} . Why? Swapping always preserves the one-dimensional marginals: $n_{m\cdot\cdot}$, $n_{\cdot b\cdot}$ and $n_{\cdot\cdot s}$; but the 2020 US Census DAS does not. For example, the number of 25-34 year old people in the US is not invariant under the 2020 TDA, but it must necessarily be invariant under any swapping algorithm.

3.2.3 PERMUTATION SWAPPING SATISFIES ε -DP SUBJECT TO ITS INVARIANTS

In this section, we design a specific data swapping algorithm – called the *Permutation Swapping Algorithm* (PSA) to distinguish it from other data swapping methods – which satisfies the DP flavor $(\mathcal{X}, \mathcal{D}_{c_{\text{Swap}}})$,

$d_{\text{HamS}}^r, D_{\text{MULT}})$. Here \mathcal{X} denotes any set of datasets which all have the same common set of variables, and d_{Ham}^r denotes the Hamming distance (formally defined in (A.3)) at the resolution r of the PSA’s swapping procedure. For example, if the PSA swapped records which correspond to individual persons, then the input premetric of the PSA’s DP flavor would be the Hamming distance d_{HamS}^p on person-records. Alternatively, the PSA could swap household-records, in which case its input premetric would be the Hamming distance d_{HamS}^{hh} at the resolution of households. (This distinction will become important when we compare the PSA with the TDA.) The output premetric of the PSA’s DP flavor is the multiplicative distance, which underlies pure ε -DP (Dwork et al., 2006b) and is defined as:

$$D_{\text{MULT}}(P, Q) = \sup_{E \in \mathcal{F}} \left| \ln \frac{P(E)}{Q(E)} \right|, \quad (3.3)$$

for two probabilities P and Q on the measurable space \mathcal{T} with σ -algebra \mathcal{F} .

While the PSA was not used in 2010, a specific instantiation of it does reflect the essential features of the 2010 DAS’s data swapping algorithm (Subsection 3.2.5). However, certain aspects of the PSA were made with the specific goal of satisfying DP. For example, a swapping method cannot satisfy $(\mathcal{X}, \mathcal{D}_{\text{cSwap}}, d_{\text{HamS}}^r, D_{\text{MULT}})$ if the number of swaps it makes is fixed. (To be clear, based on the available public information, we do not believe the 2010 DAS fixes the number of swaps, although it does appear to control this number to some degree.) To see this, suppose that a possible output dataset z differs from $x \in \mathcal{D}_0$ by m swaps and from $x' \in \mathcal{D}_0$ by $m + 1$ swaps. If the swapping methods allows a maximum of m swaps, then z has non-zero probability given x as input but zero probability given x' , thereby violating $\varepsilon_{\mathcal{D}}\text{-DP}(\mathcal{X}, \mathcal{D}_{\text{cSwap}}, d_{\text{HamS}}^r, D_{\text{MULT}})$ for any finite $\varepsilon_{\mathcal{D}_0}$. More generally, a necessary condition for a swapping method to satisfy $\varepsilon_{\mathcal{D}}\text{-DP}(\mathcal{X}, \mathcal{D}_{\text{cSwap}}, d_{\text{HamS}}^r, D_{\text{MULT}})$ for finite $\varepsilon_{\mathcal{D}_0}$ is that, given input $x \in \mathcal{D}_0$, any dataset $x' \in \mathcal{D}_0$ has a non-zero probability of being outputted (up to reordering of the rows of x').

To ensure this condition, rather than swapping rows of x_{Swap} in the same matching category m , the PSA instead randomly permutes these rows, a type of data swapping method introduced in [DePersio et al. \(2012\)](#). Since we do not want to permute every row of x_{Swap} , rows are randomly selected, independently with probability p , and only these selected rows are shuffled. Or, more accurately, after selecting rows of x_{Swap} with matching value m , the PSA samples uniformly at random a permutation $\sigma_m : \{1, \dots, n_{m..}\} \rightarrow \{1, \dots, n_{m..}\}$ which fixes nonselected rows (i.e. $\sigma_m(i) = i$ for all nonselected i), and deranges selected rows (i.e. $\sigma_m(i) \neq i$ for all selected i). This process is repeated for all values of m so that the final dataset, after all permutations have been applied, is given by $Z = [x_{\text{Hold}}, x_{\text{Swap}}^\sigma]$, where σ is defined by $\sigma(i) = \sigma_m(i)$ for record i with matching category m . In the case that only one record was selected, there are no possible σ_m and so records are re-selected. Hence, the probability that a record with matching category m is swapped is $p \sum_{j=1}^{n_{m..}-1} \binom{n_{m..}-1}{j} p^j (1-p)^{n_{m..}-1-j}$. When $n_{m..} \gg 1$, the expected fraction of records which will have their swapping variables interchanged is approximately p . For this reason, we call p the swap rate.

Pseudocode for the PSA is provided in Algorithm 3.2.1. The output is a fully-saturated contingency table $C(Z) = [n_{mbs}^Z]$ (i.e. a 3-way tensor) computed on the swapped dataset Z . When $V_{\text{Match}}, V_{\text{Hold}} \setminus V_{\text{Match}}$ and V_{Swap} all take a finite number of values, $C(Z) = [n_{mbs}^Z]$ is a collection of \mathcal{M} matrices $C_m(Z) = [n_{mbs}^Z]$, for $m = 1, \dots, \mathcal{M}$, each of which has dimension $\mathcal{H} \times \mathcal{S}$. This contingency table $C(Z)$ fully determines Z up to re-ordering of the rows of Z .

Theorem 3.2.4. *Suppose \mathcal{X} is such that every dataset $x \in \mathcal{X}$ shares the same common set V of variables which is partitioned into V_{Swap} and V_{Hold} . Let $V_{\text{Match}} \subset V_{\text{Hold}}$ be the (possibly empty) set of matching variables and*

$$b = \max\{0, n_{m..} \mid \text{there are two records with different values in matching stratum } m\}.$$

Suppose that the PSA (Algorithm 3.2.1) permutes records of resolution r . Then it satisfies $\varepsilon_{\mathcal{D}}\text{-DP}(\mathcal{X}, \mathcal{D}_{\text{cSwap}}, d_{\text{HamS}}^r, D_{\text{MULT}})$ with

$$\varepsilon_{\mathcal{D}} = \begin{cases} 0 & \text{if } b = 0, \\ \ln(b+1) - \ln o & \text{if } 0 < p \leq \frac{\sqrt{b+1}}{\sqrt{b+1}+1} \text{ and } b > 0, \\ \ln o & \text{if } \frac{\sqrt{b+1}}{\sqrt{b+1}+1} \leq p < 1 \text{ and } b > 0, \\ \infty & \text{if } p \in \{0, 1\} \text{ and } b > 0, \end{cases} \quad (3.4)$$

where $o = p/(1-p)$.

It is worth noting that the monotonic increase of $\varepsilon_{\mathcal{D}}$ with b may seem counter intuitive, until one realizes that the privacy loss budget quantified in Theorem 3.2.4 does not include the loss due to the invariants themselves. In other words, the more invariants the PSA imposes – which tends to lead to smaller b – the less information there is left for the PSA to protect, and hence it is easier to achieve smaller $\varepsilon_{\mathcal{D}}$. This phenomenon is not unique to the PSA, but reflects the fundamentally *relative* nature of DP. See Section 3.5 and Part III (Bailie et al., 2025d) for more discussions, especially regarding how this relative nature of DP provides a perverse way of achieving seemingly low privacy loss budget while increasing disclosure risk.

A proof of Theorem 3.2.4 is presented in Appendix B.3. Here we give a broad sketch for the case $0 < p \leq 0.5$ and $b > 0$. Because $\sqrt{b+1}/(\sqrt{b+1}+1) > 0.5$, we need to show, for fixed datasets x, x' and z in the same universe $\mathcal{D} \in \mathcal{D}_{\text{cSwap}}$, that the budget $\varepsilon_{\mathcal{D}} = \ln(b+1) - \ln o$ satisfies the inequality

$$\mathbb{P}[C([x_{\text{Hold}}, x_{\text{cSwap}}^{\mathcal{D}}]) = C(z)] \leq \exp(k\varepsilon_{\mathcal{D}})\mathbb{P}[C([x'_{\text{Hold}}, x_{\text{cSwap}}^{\mathcal{D}'}]) = C(z)], \quad (3.5)$$

where $k = d_{\text{HamS}}^r(x, x')$. The probabilities in (3.5) are over the random sampling of the permutations σ

Algorithm 3.2.1: The Permutation Swapping Algorithm (PSA)

Input: A dataset $x \in \mathcal{X}$ whose variable set V which is partitioned into V_{Hold} and V_{Swap} , and a set $V_{\text{Match}} \subset V_{\text{Hold}}$ of matching variables which define the matching strata.

```

1: Set  $Z \leftarrow x$ 
2: for all matching strata  $m$  do
3:   if  $n_{m..} = 0$  or  $n_{m..} = 1$  then
4:     continue
5:   end if
6:   for record  $i$  in stratum  $m$  do
7:     Select  $i$  with probability  $p$ 
8:   end for
9:   if 0 records selected then
10:    continue
11:  else if exactly 1 record selected then
12:    Deselect all records
13:    go to line 6
14:  end if
15:  Sample uniformly at random a permutation  $\sigma_m$  which fixes the unselected records and deranges the selected records
16:  /* Permute the swapping variables according to  $\sigma_m$ : */
17:   $Z \leftarrow [Z_{\text{Hold}}, Z_{\text{Swap}}^{\sigma_m}]$ 
18:  Deselect all records
19: end for
20: return fully-saturated contingency table  $C(Z)$ 
```

and σ' in Algorithm 3.2.1. We can show that there exists a derangement ρ of k records such that $C(x) = C([\mathcal{X}'_{\text{Hold}}, x_{\text{Swap}}^{\rho}])$ (Lemma B.3.4). (A derangement is a permutation which does not fix any rows.) Moreover, there is a bijection between the possible σ and σ' given by $\sigma' = \sigma \circ \rho$. Hence, if k_σ is the number of records deranged by σ , we have

$$k_\sigma - k \leq k_{\sigma'} \leq k_\sigma + k. \quad (3.6)$$

For such pairs of possible σ and σ' , the ratio $P(\sigma)/P(\sigma')$ can be bounded in terms of $o^{k_\sigma - k_{\sigma'}}$ and the ratio between the number of derangements of size $k_{\sigma'}$ and of size k_σ . For $o \leq 1$, this can in turn be bounded by

$(b+1)^k o^{-k}$ using the inequality (3.6). Hence $\varepsilon_{\mathcal{D}} = \ln(b+1) - \ln o$ does indeed satisfy (3.5).

In Appendix B.4, we prove that the privacy-loss budget $\varepsilon_{\mathcal{D}}$ for the PSA given in Theorem 3.2.4 is tight in the weak sense that under some mild assumptions the difference between the right and left sides of the inequality (3.5) is arbitrarily close to zero for some choice of x, x' and z .

Remark 3.2.5. Since $n_{j..}$ is an invariant, $n_{j..}^x = n_{j..}^{x'}$ for all x and x' in the same universe. Thus, b is a function of \mathcal{D} and hence so is the privacy-loss budget $\varepsilon_{\mathcal{D}}$ given in (3.4). In the context of the PSA, we will use ε to denote the value of $\varepsilon_{\mathcal{D}_{\text{cSwap}}(x^*)}$ under the universe $\mathcal{D}_{\text{cSwap}}(x^*)$ corresponding to the realized data x^* . We will also report the PSA's privacy loss budget in terms of this value ε and omit the values of $\varepsilon_{\mathcal{D}}$ for other universes \mathcal{D} . Even though it is a function of the realized data x^* , the value of ε can still be publicly reported under the PSA's DP specification without additional privacy loss (Corollary A.2.2).

3.2.4 A NUMERICAL DEMONSTRATION: THE 1940 CENSUS FULL COUNT DATA

We demonstrate the PSA using the 1940 US Decennial Census full count data, obtained from IPUMS USA Ancestry Full Count Database (Ruggles et al., 2021). For the 1940 Census, the smallest geography level is county, hence swapping is performed among household units across counties within each state, where each household's county indicator is set to be V_{Swap} . The matching variables (or swap key) V_{Match} are the number of persons per household and the household's state. Our analysis is focused on the ownership status of household dwellings, an indicator variable taking value of either owned (including on loan) or rented. This is our $V_{\text{Hold}} \setminus V_{\text{Match}}$. The invariants c_{Swap} induced by this swapping scheme include 1) the total number of owned versus rented dwellings at each of the household sizes at the state level, and 2) the total number of dwellings at each of the household sizes at the county level. In our notation, these are the $n_{m.s}$'s and the $n_{mb.}$'s, respectively.

county	owned	rented	total	owned (swapped)	rented (swapped)	total (swapped)
Barnstable	7461	3825	11286	5907	5379	11286
Berkshire	14736	18417	33153	13770	19383	33153
Bristol	33747	63931	97678	35537	62141	97678
Dukes	1207	534	1741	946	795	1741
Essex	53936	81300	135236	52631	82605	135236
Franklin	7433	6442	13875	6337	7538	13875
Hampden	30597	58166	88763	32267	56496	88763
Hampshire	9427	8630	18057	8145	9912	18057
Middlesex	104144	147687	251831	100372	151459	251831
Nantucket	593	432	1025	471	554	1025
Norfolk	44885	40285	85170	38566	46604	85170
Plymouth	24857	23882	48739	21549	27190	48739
Suffolk	49656	176553	226209	67357	158852	226209
Worcester	53126	78535	131661	51950	79711	131661
total	435805	708619	1144424	435805	708619	1144424

Table 3.1: A comparison of two-way tabulations of dwelling ownership by county based on the 1940 Census full count for the state of Massachusetts (left) and one instantiation of the PSA at $p = 50\%$ (right). Total dwellings per county, as well as total owned versus rented units per state, are invariant. All invariants induced by the PSA are not shown.

We restrict our illustration to the state of Massachusetts. Table 3.1 compares the two-way tabulations of dwelling ownership by county based on the original data and one instantiation of the swapping mechanism using a high swap rate of $p = 50\%$. The row margin of either table is the county-level total dwellings and is invariant due to $n_{\cdot b} = \sum_m n_{mb\cdot}$. The column margin is the total number of owned versus rented dwellings in Massachusetts and is invariant due to $n_{\cdot s} = \sum_m n_{m\cdot s}$.

Table 3.2 supplies the conversion between different swap rates to the privacy loss ε of the PSA. Under the current swapping scheme, the largest stratum size delineated by V_{Match} is $b = 264,331$, consisting of all two-person households of Massachusetts. Therefore by (3.4), we see that a low swap rate of 1% corresponds to $\varepsilon = 17.08$, whereas a high swap rate of 50% corresponds to $\varepsilon = 12.48$. It is worth noting

p	0.01	0.05	0.10	0.50
ε	17.08	15.43	14.68	12.48

Table 3.2: Conversion of (expected) swap rate p to privacy loss ε . Under this swapping scheme, the largest stratum size is $b = 264$, 331, the number of all two-person households of Massachusetts.

that since the invariants c_{swap} are fixed in this analysis, the different values of ε presented in this table can be directly interpreted as SDC guarantees of different quantified strengths. On the other hand, as we alluded to earlier, the privacy losses corresponding to different invariants c_{swap} are not directly comparable – see the discussion in Section 3.5.

We also examine the accuracy of the two-way tabulation as a function of swap rate. Figure 3.1 shows the mean absolute percentage error (MAPE) in the two-way tabulation induced by swapping at different swap rates from 1% to 50%. The variability across runs is small: each boxplot reflects 20 independent runs of the PSA.

Here, the mean absolute percentage error of a swapped table from its true table is defined as the cell-wise average of the ratio between their absolute differences and the true table values. The MAPE in Figure 3.1 is with respect to the contingency table of county by dwelling ownership in Massachusetts and is defined in the notation of Proposition 1 as

$$\frac{1}{\mathcal{HS}} \sum_{b,s} \frac{|n_{\cdot,bs}^x - n_{\cdot,bs}^Z|}{n_{\cdot,bs}^x},$$

where x is the true table, Z is the swapped table, b is the county indicator and s whether the house was rented or owned.

The accuracy assessment we demonstrate here is highly limited. The analysis above assesses only cell-wise departures of the swapped two-way marginal table from its confidential counterpart. It does not capture potential loss of data utility in terms of multivariate relational structures. It is well understood in the

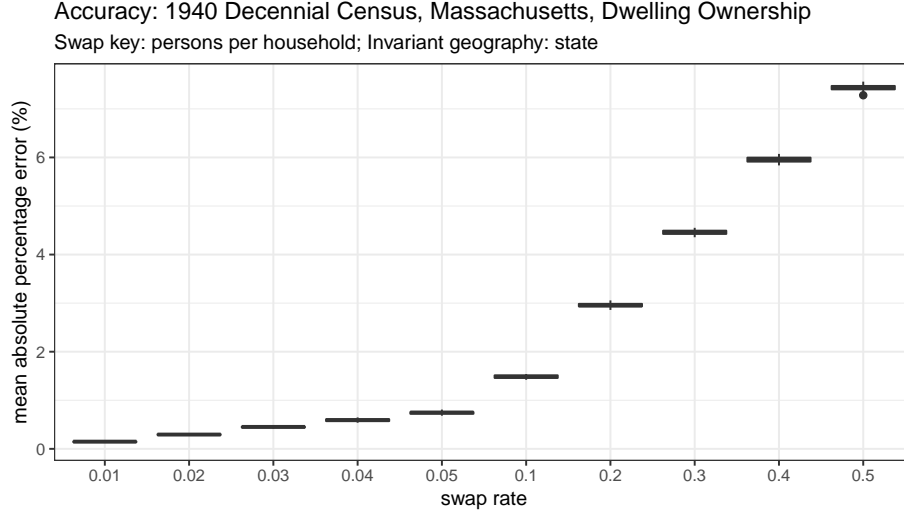


Figure 3.1: Mean absolute percentage error (MAPE) in the two-way tabulation of dwelling ownership by county induced by the PSA applied to the 1940 Census full count data of Massachusetts, at different swap rates from 1% to 50%. Each boxplot reflects 20 independent runs of the PSA at that swap rate.

literature that swapping erodes the correlation between V_{Swap} and $V_{\text{Hold}} \setminus V_{\text{Match}}$; see e.g. [Slavković and Lee \(2010\)](#); [Drechsler and Reiter \(2010\)](#); [Mitra and Reiter \(2006\)](#). For the current example, this means the county-wide characteristics of household dwellings (other than their size) are not preserved, but other multivariate relationships are. While an in-depth investigation into the utility of swapping is out of scope of this paper, we return to the subject of data utility in Part III ([Bailie et al., 2025d](#)) to discuss the implication this work may have on that line of inquiry.

3.2.5 ESTIMATING THE DP SPECIFICATION OF THE 2010 DAS

If we entertain the assumption that the 2010 DAS implemented the PSA, we could obtain a crude sketch of the SDC guarantee afforded to the 2010 Census data. (We examine the validity of this assumption in detail in Appendix B.6.1.) As detailed in Example 3.2.1, the 2010 DAS swapped household records. Therefore, the DP flavor for the 2010 DAS would be $(\mathcal{X}_{\text{CEF}}, \mathcal{D}_{\text{cSwap}}, d_{\text{HamS}}^{bb}, D_{\text{MULT}})$. Here the domain \mathcal{X}_{CEF} is the set

of all possible Census Edited Files, where the term Census Edited File (CEF) refers to the dataset which is inputted into the USCB’s DAS and consists of the Census data after editing and imputation.

The 2010 DAS utilized a swap key which included household size as well as household voting age population and some geography (either tract, county or state). As we are unable to locate 2010 Census data products that allows for the precise calculation of b pertaining to this particular swapping scheme, the swap key we consider here is coarser as it does not accounting for the household count of voting age persons. However, setting V_{Match} to be ‘state \times household size’ would imply $b = 3.65$ million, which serves as an upper bound for the actual b for the 2010 Census. Combined with a purported swap rate p between 2-4% (boyd and Sarathy, 2022) we arrive at (an overestimate of) the nominal ε to be between 18.29 and 19. We emphasize that this value of ε does not necessarily reflect the privacy loss budget of the 2010 DAS, but rather the privacy loss of the PSA when we choose its parameters to reflect what we know about the 2010 DAS.

As is always the case, this privacy loss budget must be interpreted within the context of its DP flavor. Crucially, the DP flavor $(\mathcal{X}_{\text{CEF}}, \mathcal{D}_{\text{c}_{\text{Swap}}}, d_{\text{HamS}}^{bb}, D_{\text{MULT}})$ for the above instantiation of the PSA includes the invariants V_{Hold} and $V_{\text{Swap}} \times V_{\text{Match}}$ (Proposition 1). Under the 2010 parameter choices, these invariants are the counts of households by number of occupants at the block level, and all cross-classifications of non-geographical variables at the state level. The values of ε provided above are modulo any SDC leakage caused by the release of these invariants.

3.3 A DP ANALYSIS OF THE TOPDOWN ALGORITHM

This section provides a DP specification for the TopDown Algorithm (TDA) (US Census Bureau, 2023; Abowd et al., 2022a). The TDA was used to produce the P.L. 94-171 Redistricting Summary File (PL)

(US Census Bureau, 2021a,b) and the Demographic and Housing Characteristics File (DHC) (US Census Bureau, 2023c) for the 2020 Census. Four other products – the Demographic Profile (US Census Bureau, 2023e), the Privacy-Protected Microdata Files (PPMF) (US Census Bureau, 2024b), the Redistricting and DHC Noisy Measurement Files (NMF) (US Census Bureau, 2023h,b) and 118th Congressional District Summary File (US Census Bureau, 2023a) – were also derived during the production of the PL and DHC files. Hence the publication of these four additional data products do not contribute to additional privacy loss, and our SDC guarantees for the PL and DHC files automatically extend to cover the release of all six products.

We prove in Theorem 3.3.1 that the TDA satisfies zero concentrated DP (zCDP) (Bun and Steinke, 2016) subject to its invariants. Specifically, the TDA satisfies the DP flavor $(\mathcal{X}_{\text{CEF}}, \mathcal{D}_{\text{cTDA}}, d_{\text{HamS}}^p, D_{\text{NoR}})$, where D_{NoR} is the normalized Rényi metric (Section 2.4.5) and $\mathcal{D}_{\text{cTDA}}$ is the multiverse induced by the TDA’s invariants: the state population totals; the total number of housing units in each census block; and the count of each type of occupied group quarters in each block. By proving that the TDA cannot satisfy ρ -zCDP (with input premetric d_{HamS}^p) for any finite ρ without conditioning on these invariants, we will also show that the TDA’s DP flavor must have $\mathcal{D}_{\text{cTDA}}$ (or a refinement of $\mathcal{D}_{\text{cTDA}}$) as its multiverse.

The TDA, summarized in Algorithm 3.3.1, was run twice for the 2020 Census – once to produce the PL file and then a second time for the DHC file. It is a two step procedure: The first step (called the “measurement phase” in Abowd et al. (2022a)) produces the NMF $T_p(x_p)$ and $T_b(x_{bh})$. Here x_p and x_{bh} denote the representations of the Census Edited File at the person and household levels respectively. The NMF are privacy-enhanced versions of tabular summaries $Q_p(x_p)$, at the person level, and $Q_b(x_{bh})$, at the household level, respectively. (In this section, we will include group quarters as households for the

purposes of conciseness.) The tabular summaries $Q_p(x_p)$ and $Q_b(x_{bb})$ are different for the PL and DHC files, but, roughly, they are the statistics (without privacy noise) that the US Census Bureau would like to include in each of these files. For example, when releasing the PL file, $Q_p(x_p)$ and $Q_b(x_{bb})$ are the statistics in this file, but aggregated directly from the Census microdata without any privacy protection. (However, to improve accuracy, the USCB adds some additional statistics to $Q_p(x_p)$ and $Q_b(x_{bb})$ which do not appear in the PL file.) Discrete Gaussian noise is added to $Q_p(x_p)$ and $Q_b(x_{bb})$ to produce the NMFs $T_p(x_p)$ and $T_b(x_{bb})$.

		ρ^2	ε (with $\delta = 10^{-10}$)
PL	Household	0.07	2.70
	Person	2.56	17.90
DHC	Household	7.70	34.33
	Person	4.96	26.34
Total		15.29	52.83

Table 3.3: The privacy loss budgets of the mechanisms T_p (person) and T_b (household) used in the first step of the TDA to produce the 2020 Census Redistricting Data (PL 94-171) Summary File (PL) and the Demographic and Housing Characteristics File (DHC). Source: [US Census Bureau \(2023i\)](#). Note here for each row, the value of ε is computed using the conversion $\varepsilon = \rho^2 + 2\rho\sqrt{-\ln \delta}$ given in [Bun and Steinke \(2016\)](#) and adopted by the USCB. (Hence the aggregate loss of 52.83 is not the sum of the individual ε 's.) We follow the USCB's choice of $\delta = 10^{-10}$.

In the second step (called the “estimation phase” in [Abowd et al. \(2022a\)](#)), the PPMF Z_p and Z_b are produced by solving a complex optimisation problem. (The PPMF is also called the Microdata Detail File by [Abowd et al. \(2022a\)](#).) The PPMF Z_p and Z_b agree with the Census Edited File x_p, x_{bb} on the invariants c_{TDA} . In addition, the PPMF Z_p and Z_b for the DHC are consistent with related statistics in the PL file ([US Census Bureau, 2023n](#)). To enforce this consistency, the PL file P is passed as input into the TDA when producing the DHC and a constraint $H(Z_p, Z_b) = P$ is added to the optimization problem. (The input P is not used by the TDA in producing the PL file.)

The PL and DHC files are tabulations of the PPMF datasets Z_p and Z_b . In addition to the PL and DHC files, the USCB released the NMF $T_p(x_p)$ and $T_b(x_{bb})$ produced for the PL and DHC files (US Census Bureau, 2023r) and the PPMF Z_p and Z_b for the DHC file (US Census Bureau, 2023j). The Demographic Profile and the 118th Congressional District Summary File are retabulations of the DHC file (US Census Bureau, 2023d,a).

Algorithm 3.3.1: Overview of the TopDown Algorithm (Abowd et al., 2022a), focusing on aspects salient to SDC.

Input:

A CEF $x \in \mathcal{X}_{\text{CEF}}$ with representations x_p and x_{bb} at the person and household levels respectively
 Person and household queries Q_p and Q_b
 Privacy noise scales D_p and D_b
 Constraints c_{TDA}^+ (including invariants c_{TDA} , edit constraints and structural zeroes)
 (Optional) previously released statistics P along with an aggregation function H which specifies the relationship between P and the Privacy-Protected Microdata Files Z_p and Z_b

1: Step 1: Noise Infusion

2: Sample discrete Gaussian noise (Canonne et al., 2022):

3: $W_p \sim \mathcal{N}_{\mathbb{Z}}(0, D_p)$

4: $W_b \sim \mathcal{N}_{\mathbb{Z}}(0, D_b)$

5: Compute Noisy Measurement Files:

6: $T_p(x_p) \leftarrow Q_p(x_p) + W_p$

7: $T_b(x_{bb}) \leftarrow Q_b(x_{bb}) + W_b$

8: Step 2: Post-Processing

9: Compute Privacy-Protected Microdata Files Z_p and Z_b as a solution to the optimization problem:

10: Minimize loss between $[T_p(x_p), T_b(x_{bb})]$ and $[Q_p(Z_p), Q_b(Z_b)]$

11: subject to constraints $c_{\text{TDA}}^+(Z_p, Z_b) = c_{\text{TDA}}^+(x_p, x_{bb})$ and $H(Z_p, Z_b) = P$.

Output:

Privacy-Protected Microdata Files Z_p and Z_b ;

Noisy Measurement Files $T_p(x_p)$ and $T_b(x_{bb})$ at the person and household levels.

Theorem 3.3.1. *The TDA satisfies the DP specification $\rho\text{-DP}(\mathcal{X}_{\text{CEF}}, \mathcal{D}_{c_{\text{TDA}}}, d_{\text{HamS}}^p, D_{\text{NoR}})$ with privacy-*

loss budget $\rho^2 = 2.63$ for the PL file and $\rho^2 = 15.29$ for the DHC file. (Note that these budgets do not vary with the universe $\mathcal{D} \in \mathcal{D}_{\text{TDA}}$.)

In the opposite direction, let c' be any proper subset of TDA's invariants. Then TDA does not satisfy ρ - $DP(\mathcal{X}_{\text{CEF}}, \mathcal{D}_{c'}, d_{\text{HamS}}^p, D_{\text{NoR}})$ with any finite budget ρ .

A proof of Theorem 3.3.1 is given in Appendix B.5.

Remark 3.3.2. Because the standard parametrization of zCDP's privacy loss budget is equal to the square of our parametrization (see Section 2.4.5), throughout this paper we report zCDP budgets in terms of ρ^2 to maintain consistency with the values reported in existing publications.

3.4 COMPARISONS BETWEEN THE PSA AND THE 2020 DAS

This section compares the DP specification of the PSA with the specification of the DAS used for the 2020 US Decennial Census. Specifically, we examine the PSA's DP specification in the counterfactual situation it was deployed as the 2020 DAS. This provides a comparison between the actual SDC guarantees for the 2020 Census and the hypothetical guarantees that would have been provided by the PSA.

These comparisons are presented in Table 3.4. Explanatory notes to this table are listed in Subsection 3.4.1. Subsection 3.4.2 provides necessary context to the first five rows of Table 3.4 by describing the 2020 DAS and its data products. Subsection 3.4.3 derives the DP specification presented in the final, sixth row of Table 3.4 by applying the PSA to the 2020 Census. Lastly, Subsection 3.4.4 describes the protection units associated with each of the DP specifications given in Table 3.4.

3.4.1 EXPLANATORY NOTES TO TABLE 3.4

1. In addition to invariants, the TopDown Algorithm also enforces that the PPMF Z_p and Z_b satisfy edit constraints and structural zeroes (Abowd et al., 2022a). But, because every possible $x \in \mathcal{X}_{\text{CEF}}$

	D_{Pr}	$d_{\mathcal{X}}$ (Resolution)	Invariants ¹	Privacy Loss Budget ²
TopDown	D_{NoR}	d_{HamS}^p (person)	Population (state) Total housing units (block) Occupied group quarters by type (block)	P.L. 94-171 & DHC: $\rho^2 = 15.29$ ($\varepsilon = 52.83, \delta = 10^{-10}$) See Table 3.3
SafeTab-P			Total housing units (block)	DDHC-A: $\rho^2 = 19.776$
SafeTab-H		d_{HamS}^{bb} (household)		DDHC-B: $\rho^2 = 17.79$
PHSafe			≥ 1 housing unit (block) ³	S-DHC: $\rho^2 = 2.515$
Overall (to date) 2020 DAS ⁴	D_{NoR}	d_{HamS}^p (person)	Same as TopDown	$\rho^2 = 55.371$ ($\varepsilon = 126.78, \delta = 10^{-10}$)
Swapping (PSA)	D_{MULT}	d_{HamS}^{bb} (household)	Varies but much greater than TopDown ⁵	ε between ⁶ 8.42-19.36

Table 3.4: The DP specifications of the TDA (US Census Bureau, 2023l; Abowd et al., 2022a), the SafeTab Algorithms (US Census Bureau, 2023m, 2024e; Tumult Labs, 2022), the PHSafe Algorithm (US Census Bureau, 2024f), and of the hypothetical application of the PSA to the 2020 Decennial Census. For each DP specification, the protection domain is the set \mathcal{X}_{CEF} of all possible CEFs and the multiverse is induced by the listed invariants. d_{HamS}^p and d_{HamS}^{bb} denote the Hamming distance at the resolution of person- and household-records respectively (Subsection 2.4.4); D_{NoR} the normalized Rényi metric (Subsection 2.4.5), which is the output premetric underlying ρ -zCDP (Bun and Steinke, 2016); and D_{MULT} the multiplicative distance (equation (3.3)), which is pure ε -DP’s output premetric (Dwork et al., 2006b).

satisfies these constraints by construction, these requirements need not be included as invariants (Proposition A.2.1).

2. We report zCDP budgets in terms of ρ^2 (rather than ρ) to be consistent with other literature on the 2020 DAS (see Remark 3.3.2). Moreover, following the USCB, we use the formula $\varepsilon = \rho^2 + 2\rho\sqrt{-\ln \delta}$ to convert from ρ -zCDP to (ε, δ) -DP (Bun and Steinke, 2016) with $\delta = 10^{-10}$.
3. The PHSafe has inequality invariants (see equation (2.6)). Specifically, its invariant function $c(x)$ is the vector of indicators for whether each Census block has at least one housing unit or not.
4. This DP specification covers all of the primary 2020 Census data products (US Census Bureau, 2024d) (which are listed in Subsection 3.4.2) but not other data products which are derived from the 2020 Census Edited File, such as the 2020 DAS accuracy metrics (US Census Bureau, 2023g), the Population and Housing Unit Estimates (US Census Bureau, 2023q) and the National Population Projections (US Census Bureau, 2023p). We were unable to locate information on the DP specifications associated with these data products. Nevertheless, as with any data release, they necessarily increase the total privacy-loss budget associated with the 2020 Census. They could also possibly weaken the 2020 Census’s DP flavor by increasing the invariants, weakening the output premetric, or increasing the resolution of the input premetric. Moreover, the USCB may make additional releases in the future, such as the Surname File (US Census Bureau, 2016) or research papers generated with access to Census microdata (Hawes, 2021a). These releases would further weaken the DP specification for the 2020 Census. In comparison, under data swapping, the privacy loss budget and DP flavor covers all data releases (Subsection 3.4.3).
5. Depending on the swap key V_{Match} and the swapping variables V_{Swap} , invariants induced by the PSA are all (multivariate) household characteristics at either the state, county or block group levels, and optionally the household size at the corresponding geography one level lower. See Subsection 3.4.3 for details.
6. The exact privacy loss budget ε depends on the swap rate p and the swap key V_{Match} , with the combination of a higher swap rate and finer geography-household strata giving rise to the lower range and vice versa (Table 3.5).

3.4.2 OVERVIEW OF THE 2020 DAS

The USBC has published three groups of privacy-protected data products for the 2020 Census. Group 1 encompasses the two principal data products of the 2020 Census that we have already discussed, namely

the PL and DHS. (The Demographic Profile is also included in Group 1 but as it is simply a subset of DHC’s tabulations, we do not consider it as a stand-alone product.) As detailed in the previous section, both the PL and the DHC Files were protected using the TDA (US Census Bureau, 2023l; Abowd et al., 2022a).

Group 2 encompasses the Detailed DHC-A (US Census Bureau, 2023f) and Detailed DHC-B files (US Census Bureau, 2024a), respectively protected using the SafeTab-P and SafeTab-H Algorithms (US Census Bureau, 2023m, 2024e; Tumult Labs, 2022) (where “-P” and “-H” stands for “persons” and “households”). It also includes the Supplemental DHC (S-DHC) file (US Census Bureau, 2024c), protected using the PHSafe Algorithm (US Census Bureau, 2024f).

Group 3 contains the additional products derived from the 2020 Census data, most notably the PPMF (US Census Bureau, 2024b), the 118th Congressional District Summary File (US Census Bureau, 2023a), and NMFs (US Census Bureau, 2023h,b). As explained in the previous section, because these data products are derived either from the Group 1 products or the privacy-protected intermediate outputs pertaining to those products, their production does not contribute to the overall 2020 Decennial Census privacy loss budget. As a result, we need not consider these Group 3 products in our analysis. Other Group 3 data releases, such as publications from researchers with access to Census microdata, may be released in the future (Hawes, 2021a). The results presented in Table 3.4 do not account for these releases – or for products derived from 2020 Census data which are not listed in this subsection.

3.4.3 WHAT IF THE 2020 CENSUS USED SWAPPING?

In this subsection we ask the counterfactual question: what if the PSA was applied to the 2020 Decennial Census? In particular, what would the SDC guarantee look like under different choices for the swapping

schemes and swap rates?

Table 3.5 shows the total nominal privacy loss ε that would be achieved by applying PSA to the 2020 Decennial Census for a variety of possible parameter choices. For the purpose of illustration, we stipulate the swapping variable V_{Swap} to be the block, tract, or county membership of each household, and the matching variable V_{Match} to be the geography one level higher than V_{Swap} , either alone or crossed with the household size variable. From the top to bottom rows of Table 3.5, the V_{Swap} levels are ordered according to increasing granularity of geography. Within each level of V_{Swap} , the two V_{Match} levels are nested, in the sense that the swapping scheme represented in the latter row (i.e. crossed with household size) induces a logically stronger and more constrained set of invariants than the former one. These $V_{\text{Match}} \times V_{\text{Swap}}$ level combinations result in largest strata of varying sizes, as can be seen from b ranging from as large as 13.47 million (the total number of households in California) to as small as 4,549 (the total number of 2-person households in a Florida block group).

V_{Match}	V_{Swap}	b	Total ε $p = 5\%$	Total ε $p = 50\%$	Largest Stratum
State	county	13,475,623	19.36	16.42	California
State \times household size	county	3,948,028	18.13	15.19	California, 2-household
County	tract	3,420,628	17.99	15.05	LA County
County \times household size	tract	939,185	16.70	13.75	LA County, 2-household
Block group	block	6,204	11.68	8.73	a CA block group
Block group \times household size	block	4,549	11.37	8.42	a FL block group, 2-household

Table 3.5: The total nominal privacy loss ε for the PSA applied to the 2020 Decennial Census for a variety of V_{Match} , V_{Swap} , and swap rate choices. The column b is the number of households in the largest stratum, obtained from the DHC. (The CA and FL block groups identified in rows 5 and 6 have 2020 Census GEOIDs 060730187001 and 121199114024 respectively.)

This analysis highlights an important, yet perhaps counterintuitive, observation: When the swap rate p is fixed, including more invariants decreases the nominal privacy loss ε of the PSA. As Table 3.5 shows,

when swaps are performed freely across counties in a state, even a high swap rate of 50% renders a nominal ε that is much larger than that pertaining to swaps among households of the same size within a block group at a low swap rate of 5% ($\varepsilon = 16.42$ and 11.37 respectively). If these nominal ε 's are taken at face value, one may be tempted to conclude that swapping schemes with finer invariants should be preferred from a privacy standpoint. Furthermore, one may find it convenient to also recognize that finer invariants are desirable from a data utility standpoint, for the obvious reason that more exact statistics about the confidential are made known. However, as we warned right after presenting Theorem 3.2.4, such a conclusion – that finer invariants should benefit both utility *and* privacy – would be dangerously mistaken, for it overlooks the privacy leakage, in an ordinary sense of the phrase, due to the invariants alone. This highlights the importance of interpreting ε within the context of its DP flavor, and the necessity of treating the invariants as an integral part of the SDC guarantee.

Note that if the PSA were applied to the 2020 Decennial Census, the nominal ε reported in Table 3.5 would be the *total* privacy loss budget across all data products derived from the swapped dataset Z , including the PL, DHC, DDHC and S-DHC files, for both persons and household product types. This is because swapping is performed on the full microdata file, and hence produces a synthetic version of it from which all data products can be generated. Therefore, when comparing the ε values in Table 3.5 with those reported for the 2020 DAS in Table 3.4, it should be understood that the privacy loss for the PSA covers all the 2020 data products. This characteristic of swapping leads to an additional desirable property that is not necessarily enjoyed by mechanisms based on output noise infusion (such as those used in the 2020 DAS): the *logical consistency* between, and within, multiple data products is automatically preserved under swapping without the need for post-processing.

3.4.4 THE PROTECTION UNITS FOR THE 2020 DAS AND FOR THE PSA

As detailed in Section 2.4.4, the protection units (aka privacy units) of a DP specification are the basic entities which are protected under that DP specification. More exactly, a specification's privacy loss budget restricts how much a mechanism can change when a single protection unit's data changes. One might imagine that the protection units of a DP specification correspond to the resolution of its input premetric $d_{\mathcal{X}}$ – and this is true in simplistic examples. Here by ‘the resolution of $d_{\mathcal{X}}$ ’, we mean the size of the change between x and x' when $d_{\mathcal{X}}(x, x') = 1$. For example, if $d_{\mathcal{X}}(x, x') = 1$ whenever x and x' differ on a person-record, then the resolution of $d_{\mathcal{X}}$ is a person. Common resolutions, in order from high to low, are: single transactions or interactions, persons, households and businesses.

However, data preprocessing can create complications, so that the protection units of a specification are not always given by the resolution of $d_{\mathcal{X}}$. In the case of the US Census, an individual respondent's data can be used for multiple records in the CEF $x \in \mathcal{X}_{\text{CEF}}$ because the USCB's imputation procedure replaces missing records with copies of non-missing records. As such, the protection units of $(\mathcal{X}_{\text{CEF}}, \mathcal{D}, d_{\mathcal{X}}, D_{\text{Pr}})$ do not correspond to the resolution of $d_{\mathcal{X}}$. Rather, the protection units of the 2020 DAS are ‘post-imputation persons,’ – i.e. those (fictional) entities with data that is exactly one record in the CEF. Similarly, the PSA's protection units are the ‘post-imputation households’ rather than actual households.

This point is not simply a matter of semantics. From the perspective of a data respondent, the resolution of $d_{\mathcal{X}}$ is not particularly informative in determining the SDC protection provided to them, because the respondent's actual privacy loss – in terms of ρ – is given by the nominal privacy loss multiplied by the number of person-records the respondent contribute to. For example, if the 2020 imputation process duplicates a respondent's record once, then their actual privacy loss is $\rho^2 \geq 221.48$ (or $\varepsilon \geq 364.31$ with

$\delta = 10^{-10}$), rather than $\rho^2 \geq 55.37$ - since doubling ρ quadruples ρ^2 (Bun and Steinke, 2016, Proposition 27). (We write \geq rather than $=$ because the privacy loss will increase due to data releases which we have not accounted for – see note 4 in Subsection 3.4.1.) In general, the conversion from a DP flavor with post-imputation persons as units to a DP flavor with persons as units requires an inflation of the privacy loss budget (ϵ or ρ) by a factor equal to the maximum number of times a record can be duplicated (Proposition 2.4.15). To avoid this complication, we have reported post-imputation budgets in Table 3.4, but we caveat this with the important observation that these budgets correspond to unusual protection units.

3.5 DISCUSSION

This paper continues an existing line of research (Rinott et al., 2018; Bailie and Chien, 2019; Chien and Sadeghi, 2024; Neunhoeffer et al., 2024) examining traditional SDC methods – which are typically regarded as ad-hoc and are motivated by intuitive notions of protection or specific attacker models – under the light of DP. By proving that data swapping can be studied theoretically via the lens of DP, we hope to inspire further formal analyses of other traditional SDC methods. This type of analysis improves our understanding of such methods by supplying mathematical descriptions of the level and substance (or, in our terminology, the intensity and flavor) of the methods’ SDC. Such descriptions are important: they can provide assurance to data providers and custodians that their data is adequately protected; or, conversely, they can reveal inadequate SDC and spur additional protection.

However, it can be challenging to assess whether a given DP specification provides an adequate level of protection. To do so, we must understand how choices for each of the five building block can affect SDC – both individually and in conjunction with choices for the other building blocks. This requires answering a range of difficult socio-technical questions. For instance, taking the other four building blocks as fixed,

what privacy-loss budget (if any) is sufficient for adequate SDC, adequate for whom, and who should decide the adequacy? Also, what is the practical impact of the protection units being post-imputation persons? And, most relevant for this paper, what is the effect of invariants on SDC?

While we know that increasing the invariants strictly weakens the DP specification (Proposition 2.4.11), it is more difficult to determine how they affect an attacker’s ability to make disclosures. [Ashmead et al. \(2019\)](#) have investigated the effect of the 2020 DAS invariants, but there is a need for future work which studies the effect of invariants at the scale of those induced by data swapping. In addition to building technical understanding – and parallel to studies that survey preferences on appropriate settings for the privacy loss budget, protection domain and input and output premetrics – it could be beneficial to gauge public opinion on the acceptability of specific invariants.

Nevertheless, by providing DP specifications for both the PSA and the TDA, we demonstrate the feasibility of mathematically comparing on fair grounds traditional SDC methods with DP-based mechanisms. With these two algorithms as prime examples, the paper points to the possibility of similar comparative analyses between other SDC methods, both those that were explicitly inspired by DP and those that were designed and motivated from non-DP perspectives. By explicating the five building blocks for the PSA and the TDA, we hope to promote nuanced assessments of DP deployments which go beyond discussion of the privacy loss budget ϵ .

We end this paper with discussions on two main results of this paper – the PSA’s privacy loss budget and the comparison between the DP specifications for the 2020 DAS and the PSA, as presented in Table 3.4.

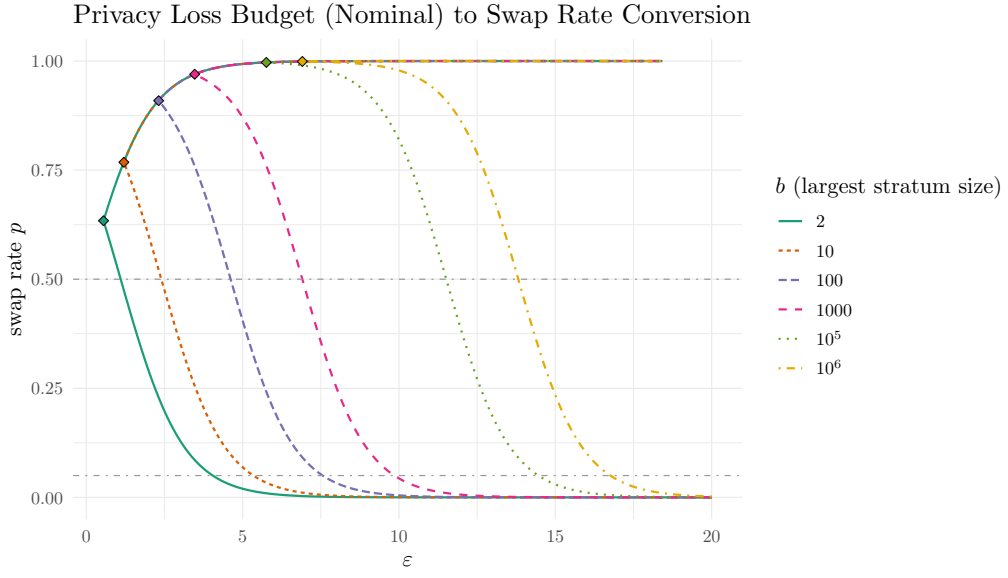


Figure 3.2: Conversion between the nominal privacy loss budget (ϵ) and the swap rate (p) for the PSA. Color and line type encode different values of b , the size of the largest stratum delineated by V_{Match} (from 2 to 1 million). Outlined diamonds indicate the smallest ϵ attainable for each b . Grey dotted horizontal lines correspond to swap rates of 5% and 50% respectively. The ϵ values are nominal in that the privacy guarantee they afford shall be understood in the context of c_{Swap} (and hence the values of ϵ across different values of b are not immediately comparable).

3.5.1 WHAT DOES THE PSA'S BUDGET LOOK LIKE?

Figure 3.2 provides a visual illustration of Theorem 3.2.4, connecting the privacy loss budget ε to the swap rate p , for a number of choices of b . Three observations are worth noting. First, for each b , there exists a smallest ε , call it $\varepsilon^{(b)}$, below which no swap rate $p \in [0, 1]$ can attain. The minimum budget $\varepsilon^{(b)} = \ln(b+1)/2$ is achieved by the swap rate $p^{(b)} = \sqrt{b+1}/(\sqrt{b+1}+1)$. For each b in Figure 3.2, this quantity $\varepsilon^{(b)}$ is marked by an outlined diamond. Importantly, the larger the b , the larger the minimum possible budget $\varepsilon^{(b)}$. For example, when $b = 10$, $\varepsilon^{(b)}$ is 1.20 (attained at $p^{(b)} = 77\%$); whereas when $b = 10^6$, $\varepsilon^{(b)}$ is 6.91 (at $p^{(b)} = 99.9\%$).

That some privacy-loss budgets are not attainable for a fixed b follows from the fact that the ratio $P(z \mid x)/P(z \mid x')$ of probabilities of a swapped dataset z from two different input datasets x and x' depends not just on the swap rate p but also the ratio r of the number of derangements of size $d_{\text{HamS}}^{bb}(x, z)$ to the number of derangements of size $d_{\text{HamS}}^{bb}(x', z)$. (This is because the PSA selects $d_{\text{HamS}}^{bb}(x, z)$ records and then samples derangements of size $d_{\text{HamS}}^{bb}(x, z)$ uniformly at random.) This ratio r is upper bounded by $(b+1)^{d_{\text{HamS}}^{bb}(x, x')}$ which means ε must be at least $\ln(b+1) - \ln p + \ln(1-p)$.

Second, for every b and every budget $\varepsilon > \varepsilon^{(b)}$, two different swap rates can achieve that budget ε , with the higher one often being very close to 100%. For example at $b = 10$, a swap rate of either 35.4% or 95.2% achieves the nominal budget of $\varepsilon = 3$. The mathematical reason behind this is that, for large p (i.e. $p > p^{(b)}$) the ratio r is dominated by the odds $o = p/(1-p)$, in which case $[\ln P(z \mid x) - \ln P(z \mid x')]/d_{\text{HamS}}^{bb}(x, x')$ is maximised when $d_{\text{HamS}}^{bb}(x', z)$ is as small as possible. This results in $\varepsilon = \ln o$, while, as explained in the previous paragraph, $\varepsilon = \ln(b+1) - \ln o$ for $p \leq p^{(b)}$. Since the former ε is monotone increasing in p and the latter monotone decreasing, there are two swap rates p corresponding to any $\varepsilon > \varepsilon^{(b)}$.

This is akin to the behavior of the randomized response mechanism, where a large probability $p_{RR} > 0.5$ of flipping the binary confidential answer inadvertently preserves statistical information, thereby achieving the same budget $\varepsilon = |\ln o_{RR}|$ as $1 - p_{RR}$.

Third and most importantly, we emphasize that the budgets visualized in Figure 3.2 are *nominal* in the sense that the SDC guarantee they afford must be understood with respect to the full context as outlined by the PSA’s DP specification. An aspect of this context is b , the size of the largest stratum of V_{Match} , and as a result, the same value of ε across different b ’s should not be equated to be the same SDC guarantee. Indeed, the ordering of the b curves in the figure suggests a seemingly peculiar fact that, for a larger b , a higher p is needed to achieve the same ε . This apparent contradiction is explained by a point we have repeatedly made: for a fixed dataset, a change in the value of b requires that the swapping invariants, and hence the PSA’s SDC guarantee, also change.

3.5.2 HOW DOES THE 2020 DAS COMPARE WITH SWAPPING?

The DP specifications in Table 3.4 allow for epistemically meaningful comparisons between the 2020 DAS and the counterfactual scenario in which the PSA was used. A side-by-side examination of these specifications’ building blocks elucidates the similarities and differences between the SDC provided by the PSA and the 2020 DAS.

Firstly, both the 2020 DAS and the PSA have the same protection domain: the set of all possible CEFs \mathcal{X}_{CEF} . This means that the PSA and the 2020 DAS protect the data $x \in \mathcal{X}_{CEF}$ as it exists after collection, coding, editing and imputation, rather than as it exists at other stages in its life cycle. As such, it is not the respondents’ data (i.e. their ‘raw’ Census responses) which are directly protected, but rather it is the edited and imputed data (i.e. the CEF) which receives the DP guarantee.

Secondly, because both the 2020 DAS and the PSA have invariants, each of their DP specifications partition the protection domain \mathcal{X}_{CEF} into multiple universes. This operation constrains the scope of the 2020 DAS and the PSA’s SDC protection to datasets which agree on their invariants. Therefore, for the same reasons that the 2020 DAS cannot satisfy the original specification of zCDP given in [Bun and Steinke \(2016\)](#), data swapping cannot satisfy the original pure ε -DP specification of [Dwork et al. \(2006b\)](#). In this sense, both the 2020 DAS and the PSA are DP only in so far as their invariants allow.

To varying degrees, all of the SDC methods used in 2020 have invariants. However, the PSA has many more invariants than any of these methods and, as such, places more restrictions on the scope of protection. Unfortunately, the 2020 Census data products carry a set of invariants that cannot be induced by data swapping. That is, the invariants induced by the TDA do not accord to any choice of V_{Swap} , V_{Match} , and V_{Hold} (as shown in Example 3.2.3). Therefore, we cannot design a swapping algorithm which respects the 2020 invariants, and only those invariants. On the other hand, while swapping almost always has stricter invariants for most variables, it does not necessarily have the TDA’s group quarter invariants. Therefore, the 2020 DAS DP flavor is not strictly stronger than the PSA’s flavor, nor visa versa – although the 2020 DAS places less restrictions on the scope of protection, these are not nested within the restrictions induced by the PSA’s invariants.

Thirdly, the input premetrics for the PSA and the 2020 DAS are both Hamming distances, although with differing resolutions – household-records for the PSA and person-records for the 2020 DAS. This means the protection units are post-imputation households and post-imputation persons respectively. Since the input premetric the yardstick for measuring change in the input data (Subsection 2.4.4), using a lower resolution like household-records provides more protection than a higher resolution like person-

records (all else being equal). That is, a household-level distance is a stronger notion than a person-level distance, since if the record of a single household changes part of its value, the multiple persons residing in a same household may all change their records.

Fourthly, the PSA's output premetric is also stronger than the 2020 DAS's. The PSA uses the multiplicative distance D_{MULT} – as used in pure ϵ -DP (Dwork et al., 2006b) – while the 2020 DAS uses the normalized Rényi metric D_{NoR} – as used in zero-concentrated-DP (zCDP) (Bun and Steinke, 2016). There exist probabilities P and Q with $D_{\text{NoR}}(P, Q)$ arbitrarily small but $D_{\text{MULT}}(P, Q) = \infty$. As such, D_{MULT} ensures a greater level of SDC protection than D_{NoR} (again, assuming that all else is equal).

Fifthly, and finally, the privacy-loss budget of the PSA and of the 2020 DAS are not directly comparable because a budget's 'unit of measurement' is determined by its DP flavor. That is to say, a privacy-loss budget is a nominal measure of SDC protection, which is always relative to – and hence can only be understood within the context of – the four other building blocks. The DP flavors for the PSA and the 2020 DAS are different and so their budgets have different units of measurement. Nonetheless, following the USCB, we can convert the 2020 DAS zCDP budget $\rho^2 = 55.371$ to the approximate DP budget of $\epsilon = 126.78$ with $\delta = 10^{-10}$, which is more comparable with an ϵ -DP budget. Under this crude comparison, the privacy loss of the 2020 DAS is an order of magnitude larger than that of the PSA.

However, even when converting to (ϵ, δ) -DP, the budgets are still not directly comparable because the DP flavors for the PSA and the 2020 DAS also have different invariants and input premetrics. While the 2020 DAS's budget would substantially increase under a household-level input premetric (Appendix B.5), removing even one invariant from the PSA's DP specification would result in a budget $\epsilon = \infty$. This is because, if an invariant statistic is included in the DP flavor, then it does not contribute to the measurement of privacy loss under that flavor. As we have mentioned repeatedly, the PSA has many more invariants and

so many more statistics are removed from consideration when calculating its budget.

Yet, releasing a statistic under a large privacy loss budget is pragmatically equivalent to making that statistic an invariant. Hence in principle it should be possible to effectively tradeoff invariants with large budgets, thus making the comparison between the PSA and the 2020 DAS's budget more tractable. We leave this as an open research question.

4

Can Swapping Be Differentially Private?¹

4.1 WHAT MOTIVATED THIS STIRRED-NOT-SHAKEN TRIO?

SINCE ITS CONCEPTUALIZATION TWO DECADES AGO by [Dwork et al. \(2006b\)](#), differential privacy (DP) has received a tremendous amount of attention in research and practice. Theoretically, it aims to provide a tractable, mathematical framework to quantify and operationalize the evasive concept of privacy within the context of sharing (or releasing) statistical data. Yet, driven by a myriad of constraints (e.g., challenges in establishing theoretical guarantees or in practical implementation), attempts to alter, enhance, and relax the original, so-called pure ϵ -DP definition have led to a plethora of ‘impure’ DP formulations. Today, DP has evolved into an umbrella term encompassing a broad class of technical standards conceptualizing what it means for a data release algorithm to be ‘private’ (see [Desfontaines and Pejó \(2020\)](#) for a dizzying but still partial enumeration of this class), and there now exists a vast and burgeoning body of work supplying

¹Based on work coauthored with Ruobin Gong and Xiao-Li Meng.

algorithms which satisfy these technical standards.

From a practical perspective, the interest in DP, and subsequent explosion of different formulations, is easy to understand. General concerns of privacy breaches have increased substantially as our society adventures deeper into the digital age. Organizations in all sectors and at all levels are compelled to expend effort addressing the issue of data privacy, whether for noble reasons or for fear of liability; yet each entity is faced with its own unique set of concerns, necessitating its own custom solution (Schneider et al., 2025). The adoption of a form of DP by the United States Census Bureau (USCB) for its 2020 Decennial Census of Population and Housing is a shining example of organizational effort—one that has generated much theoretical contemplation and methodological advances, as well as controversies and emotions ranging from excitement to frustration (see the special issue of the *Harvard Data Science Review*, “Differential privacy for the 2020 US Census: Can we make data both private and useful?” Gong et al. (2022)).

Indeed, this trio of articles owes its existence to the bureau’s adoption of DP. Because data privacy has been a central concern for the census since its inception, seeing the USCB’s recent development of the DP-inspired TopDown Algorithm (TDA) (Abowd et al., 2022a) for its 2020 Census, one naturally may wonder in what ways it improves upon their past methods for statistical disclosure control (SDC). In particular, the data swapping strategy used in 2010, just like the TDA, involves injecting artificial randomness into the published census tables. So could it be a form of DP as well, which would then make it easier to compare both methods within a single, unified framework?

Those who adhere to the definition of pure ϵ -DP may immediately declare that any form of swapping cannot satisfy DP because swapping leaves some aggregations of the data unaltered, and hence some statistics—termed its *invariants*—will be released without any artificial noise injected. By the same token, neither can the TDA satisfy DP because it is designed to maintain various total counts (e.g., state

population counts) in order to comply with mandates set by the US Constitution. Yet, such a strict adherence to a narrow perspective not only greatly limits the applicability of DP, but also fundamentally misperceives its essence. (Besides which, DP has accommodated invariants from its very inception: the number of records in the dataset can be released exactly under many DP definitions, including pure ϵ -DP.) After all, DP is not concerned with protecting absolute privacy—no data release method can (Kifer and Machanavajjhala, 2011). Rather, it concerns the protection of information that can be revealed by an individual’s confidential data but is otherwise unavailable. The constitutional mandates reveal information that cannot be protected, and hence any adoption of DP—or any other data protection methodology for that matter—must take that into account.

For the case of swapping, the matter becomes muddier as its invariants are not externally mandated but instead are inherent to its design. One can see immediately the potential complications with, and debates about, different designs and their resulting invariants. Confusions may easily result from comparing SDC methods that are based on different postulates of what is already known or considered not to need protection, akin to the trouble of comparing two distributions when they’re conditioned on by different variables.

Such a problem is only one of many nuanced issues we have had to deal with as we seek precise and contextual answers to the question in this article’s title, as an impetus for a deep study of DP to reveal its statistical essence and to contemplate its practical complications. Consequently, it should come with little surprise that our investigation took considerably more time, and pages, than initially expected. This third part of a trio of papers therefore will first provide an intuitive summary and explanation of both preceding parts (Bailie et al., 2025b,c) before elaborating on their broader and deeper implications, and

discussing the more nuanced or subtle issues that tend to invite misunderstanding, misuse, and misplaced expectations of DP.

Integrating the three parts together, this triptych reflects our triple ambition: firstly, to leverage the merits of DP, including its mathematical assurances and algorithmic transparency, without sidelining the advantages of classical SDC; secondly, to unveil the nuances and potential pitfalls in employing DP as a theoretical yardstick for SDC procedures; and thirdly, to build connections between social and technical conceptualizations of data privacy by outlining real-world considerations behind the five building blocks, as demonstrated by comparing data swapping and the TDA in the context of the US Decennial Censuses.

4.2 HIGHLIGHTS OF PART I: FIVE BUILDING BLOCKS OF DP

A *DP specification*, as explained in Part I, consists of five building blocks, addressing five related, but distinct questions about a *data release mechanism* (also referred to as an SDC method, a data sharing algorithm, or similar). As we will review shortly, the core idea behind any DP definition is that the relative change in the mechanism’s likelihood function with respect to changes in individual data points is deliberately controlled. Indeed, this narrow, mathematical formulation of data privacy in terms of a rate of change highlights DP’s core idea of bounding the derivative of the mechanism, as is alluded to by its nomenclature “differential.”

A technically oriented reader may immediately ask a host of questions. On what space is the likelihood function defined? How is the relative change metricized? What constitutes an “individual data point”? What does control mean? And so on. Indeed these questions are key to formalizing the above idea of DP into mathematical definition, and their many possible answers are reflected by the numerous formulations of DP found in the literature.

A DP specification answers all these questions, thereby providing a complete, self-contained and well-defined standard against which a data release mechanism can be assessed.

The DP specification framework starts with the most basic question: *who is eligible for SDC protection?* which is addressed by specifying the actual, potential or counterfactual datasets that are to be protected. The collection of all such datasets is called the data space, or the *domain*, and is denoted by \mathcal{X} . Because setting down all the possible counterfactual datasets requires situating the actual dataset x in the context of its lifecycle—not just specifying the mathematical structure and schema of x —the domain \mathcal{X} provides the meaning of who x ’s data subjects are and how x represents them.

Based on the actual, confidential dataset x , some output statistics are computed and published via the data-release mechanism, which is a random function of x . (It is worthwhile to emphasize that in this setup, the randomness is not in x , but rather artificially injected into the output statistics to reduce their information content.) From an attacker’s perspective, the confidential dataset x —which always belongs to \mathcal{X} by design—is the unknown ‘parameter’ to be inferred from the output statistics, hence the choice of \mathcal{X} is conceptually analogous to the choice of a parameter space in standard statistical inference.

Explicating \mathcal{X} , however, is only the first step. The next question is *to where does the protection extend?* which can be answered by specifying a *multiverse* \mathcal{D} , which is a collection of the possible data universes \mathcal{D} , be they actual or hypothetical. The need for—and the distinction between—the data multiverse and the actual data universe is well illustrated by the application of DP to the US Decennial Census. Suppose the enumerated US population size N_{US} is 330 million. Then, in order to comply with the constitutional mandate that the state population totals must be released exactly as enumerated, any hypothetical census dataset that does not yield $N_{US} = 330,000,000$ will not be within the scope of protection, since any attacker can easily rule out such data sets. In this instance, if N_{US} is the only aggregation that must be dis-

closed as it is, then the corresponding data universe is simply all datasets in \mathcal{X} such that the corresponding $N_{US} = 330,000,000$.

However, it would be rather unwise to design a mechanism that works only when N_{US} is exactly 330 million. The enumerated US total population count varies from census to census, and indeed it varies within each census due to corrective adjustments (e.g., for reducing the impact of under-counting). While it is essential for any data release mechanism to respect constitutional mandates—i.e., to disclose the total state enumerations—the actual value of the disclosed N_{US} is rather accidental. A sensible design should work regardless of the value of N_{US} , at least for values within a reasonable range (e.g., one may argue that it is insensible to require a mechanism to work for implausible values, such as $N_{US} = 330$ billion). A data multiverse, therefore collects all data universe within each of which the count N_{US} is a constant, so that all datasets in \mathcal{X} are protected—but only within the scope of their respective universes.

Next is the question, *what is the granularity of protection?* which can be a source of confusion, as well as an opportunity for manipulation in capable but malicious hands. Recalling that a DP formulation is a bound on a mechanism’s rate of change, the granularity of protection is understood by considering the entities whose data is counterfactually altered when measuring this rate of change. These entities are termed the *protection units* (or, elsewhere in the literature, the privacy units). Individual persons or business entities are common choices for the protection units, but they are not the only ones. For example, for the SDC protection of electronic communications, the unit may be defined as a single message sent by a person, rather than the sender herself. As an individual may send many messages a day, such a fine-grained protection unit allows a social media platform to declare a nominal level of DP protection which appears impressively high, even though the actual risks to the sender remain exponentially large.

Mathematically, the granularity of protection is formally conceptualized via a premetric $d_{\mathcal{X}}(x, x')$, which

is a measure of the difference (or ‘distance’) between two datasets x and x' in \mathcal{X} . Furthermore, the protection units correspond conceptually to unit differences in $d_{\mathcal{X}}$ —i.e. the differences between *neighboring* (or adjacent) datasets, which are pairs of counterfactual datasets $x, x' \in \mathcal{X}$ with $d_{\mathcal{X}}(x, x') = 1$. More exactly, a protection unit is what differs during the generating processes of two neighboring datasets. While neighboring datasets could differ by the deletion of one record, or the alteration of a single attribute, to properly define the protection units requires an appreciation of what a unit difference in $d_{\mathcal{X}}$ actually represents in the real world. This in turn necessitates placing x within the context of its data pipeline, highlighting once again the importance of the domain \mathcal{X} ’s contextual definition, over and above a purely mathematical or technical description of it.

As we have repeatedly emphasized, DP quantifies SDC protection as the rate of change in the variations of a data release mechanism’s output. For each specification of DP, this rate of change is calculated with respect to the specification’s premetric $d_{\mathcal{X}}$ on the data space \mathcal{X} , but *how is the change in output variations measured?* A premetric D_{Pr} is also used for this, except in this case $D_{\text{Pr}}(P_x, P_{x'})$ is a measure of the difference between probability distributions P_x and $P_{x'}$. Here, P_x denotes the likelihood function—i.e. the probability distribution of the released statistics, as a function of the confidential data x , where the randomness in P_x is introduced solely by the data-release mechanism. This is a sensible approach to SDC: as all statistical information is created by variations in the data, by limiting the relative changes in the output distributions, we limit the changes in variations due to the change from x to x' as measured by $d_{\mathcal{X}}(x, x')$.

This brings us to the fifth and final question of a DP specification, *how much protection is afforded*, which, for each data universe $\mathcal{D} \in \mathcal{D}$, is answered by the quantity $\varepsilon_{\mathcal{D}}$. The value of $\varepsilon_{\mathcal{D}}$ is a bound on the rate of change (as measured by $d_{\mathcal{X}}$ and D_{Pr}) within the universe \mathcal{D} . Smaller values of $\varepsilon_{\mathcal{D}}$ are more restrictive and

hence supply higher levels of protection. As explained in Part I, we propose the term *protection loss budget* for $\varepsilon_{\mathcal{D}}$ instead of the commonly-used “privacy loss budget” (although we maintain the same abbreviation, PLB) because we believe that the former more accurately captures the narrow question DP is addressing (Seeman and Susser, 2023). Indeed, amid an environment of myriad and varied privacy concerns, our choice of terminology—SDC and protection in place of privacy—reflects our desire to avoid running the risk of misrepresenting DP as a solution to the broad gamut of data privacy issues (Nissenbaum, 2010; Bailie and Gong, 2023b).

We allow ε to vary between universes so that different protection loss for different universes, which may be desirable when, for utility reasons or otherwise, some entities or attributes should receive less protection.

As shown in Part I, most common DP definitions, including the classic *pure* ε -DP, *approximate* (ε, δ) -DP, and *zero-concentrated* ρ -DP (zCDP), are all special cases of the general formulation:

$$\frac{D_{\text{Pr}}(\mathbf{P}_x, \mathbf{P}_{x'})}{d_{\mathcal{X}}(x, x')} \leq \varepsilon, \quad \text{for all } x, x' \in \mathcal{D}, \text{ and for all } \mathcal{D} \in \mathcal{D}, \quad (4.1)$$

or more generally,

$$D_{\text{Pr}}(\mathbf{P}_x, \mathbf{P}_{x'}) \leq \varepsilon d_{\mathcal{X}}(x, x'), \quad \text{for all } x, x' \in \mathcal{D}, \text{ and for all } \mathcal{D} \in \mathcal{D}, \quad (4.2)$$

but with different choices for \mathcal{X} , \mathcal{D} , D_{Pr} , $d_{\mathcal{X}}$ and ε . We provide two expressions here because the first one resembles the familiar notation of taking derivative, and hence the term “differential privacy”; while the second shows that mathematically a DP specification is simply a Lipschitz continuity condition on \mathbf{P}_x as a function of the input data x . (Informally speaking, Lipschitz continuity is simply a generalization of differentiation.)

Taken together, these five answers form the building blocks of a *differential privacy specification*:

1. The protection domain (*who* is eligible for protection?), as defined by the set \mathcal{X} .
2. The scope of protection (*where* does the protection extend to?), as instantiated by the multiverse \mathcal{D} , which is a collection of universes $\mathcal{D} \subset \mathcal{X}$.
3. The protection unit (*what* is the granularity of protection?), as conceptualized by the input divergence $d_{\mathcal{X}}$ on the domain \mathcal{X} .
4. The standard of protection (*how* to measure change in output variations?), as captured by the output divergence D_{Pr} on the released data's possible probability distributions.
5. The intensity of protection (*how much* protection is afforded?), as quantified by the protection-loss budget $\varepsilon_{\mathcal{D}}$ for each data universe \mathcal{D} (where smaller budgets correspond to a higher intensity of protection).

In the current literature and practice, PLB has been treated largely as the sole measure of the strength of a privacy guarantee, as is often the case in balancing the privacy-utility tradeoff (see e.g. [Abowd and Schmutte, 2016](#)). However, the five-building-block framework reveals that $\varepsilon_{\mathcal{D}}$ can be meaningfully defined only after the *DP flavor*, that is, the collection of the first four building blocks, has been declared. To make an analogy, different DP flavors are to the PLB as different sovereigns are to their currencies. Exchange rates exist wherever trade relationships exist. Nevertheless, at the end of the day it is not currencies, but purchasing power, that matters. If today Japan announces that starting from tomorrow their currency would be denominated by “centiyen” which is equal to 100 yens, it would change the US-Japan exchange rate by 100 fold, but it would not change how many Macbooks each Japanese person could afford. Currencies are arbitrary proxies to purchasing power which are determined by economic-political features of a sovereign that are hard to measure, just like PLBs are arbitrary proxies to that ephemeral notion of “privacy” that resists definition.

Lessons learned from studying the swapping algorithm, as presented in the following section, show how one may spend apparently less PLB, all the while without adding more noise to the privacy-protected

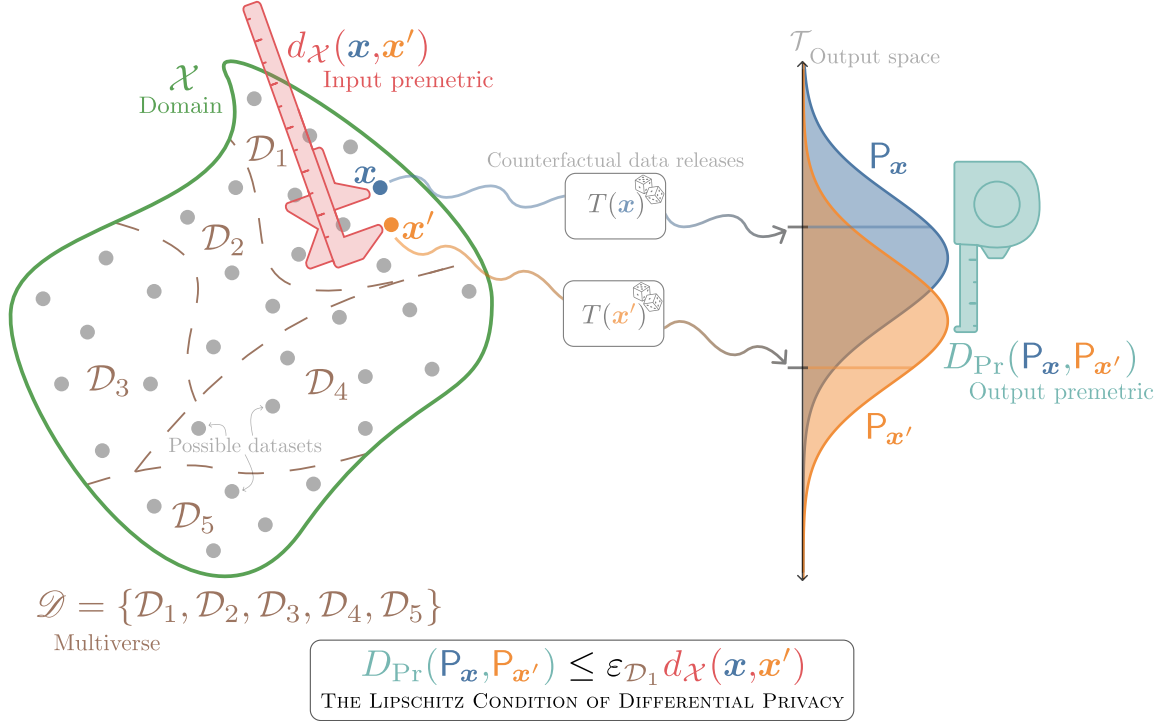


Figure 4.1: Schematic of a differential privacy specification $\varepsilon_{\mathcal{D}}\text{-DP}(\mathcal{X}, \mathcal{D}, d_{\mathcal{X}}, D_{\text{Pr}})$. The *domain* \mathcal{X} is the set of all possible datasets (be they actual, potential or counterfactual). We denote two arbitrary datasets by x and x' ; other possible datasets are depicted by gray circles. The *multiverse* $\mathcal{D} = \{\mathcal{D}_1, \mathcal{D}_2, \mathcal{D}_3, \mathcal{D}_4, \mathcal{D}_5\}$ is a collection of sets of datasets – these sets are called *universes*. (In this schematic, \mathcal{D} partitions the domain \mathcal{X} , as would happen when \mathcal{D} encodes invariants. In general, this need not be the case. In fact, often the universes may be overlapping.) A data release mechanism T transforms a dataset x to a random output $T(x)$, which is a draw from the probability distribution P_x . Intuitively, differential privacy requires that similar datasets x and x' have similar *output distributions* P_x and $P_{x'}$. This is formalized by the Lipschitz condition $D_{\text{Pr}}(P_x, P_{x'}) \leq \varepsilon_{\mathcal{D}_1} d_{\mathcal{X}}(x, x')$, which states that the ‘distance’ $D_{\text{Pr}}(P_x, P_{x'})$ between the output distributions is at most a constant multiple $\varepsilon_{\mathcal{D}_1}$ of the ‘distance’ $d_{\mathcal{X}}(x, x')$ between the corresponding input datasets. Here, similarity (or ‘distance’) between datasets is measured by the DP specification’s *input premetric* $d_{\mathcal{X}}$, visualized above as a caliper, and similarity between probability distributions of the output under different inputs is measured by the DP specification’s *output premetric* D_{Pr} (the tape measure). For simplicity, we depict the output space \mathcal{T} as one dimensional, although in practice it is frequently a high-dimensional space, or even a union of many different probability spaces (as is the case for local DP). (The PLB above, $\varepsilon_{\mathcal{D}_1}$, has the subscript \mathcal{D}_1 because the Lipschitz condition is applied to the datasets x and x' , which are members of the universe \mathcal{D}_1 , and because the PLB is allowed to vary between universes, potentially taking five different values, $\varepsilon_{\mathcal{D}_1}, \varepsilon_{\mathcal{D}_2}, \dots, \varepsilon_{\mathcal{D}_5}$.)

data product, because one can manipulate other building blocks. These are not merely mathematical possibilities or pathological cases, but rather consequences of taking the value of $\varepsilon_{\mathcal{D}}$ nominally and out of context, leading us down a slippery slope.

4.3 HOW TO REDUCE ‘PRIVACY LOSS’ WITHOUT ADDING MORE NOISE: A PERVERSE GUIDE

This section discusses some of the ways to ‘cheat,’ i.e. to reduce the PLB without adding more noise. The mathematical feasibility of this can be explained rather simply. In order to alter one component of a differential privacy specification—in this case the PLB—one can maneuver the other components (\mathcal{X} , \mathcal{D} , $d_{\mathcal{X}}$ or D_{Pr}), or parameters of the data release mechanism to that end. Some such maneuvers may be valid, especially if the data custodian is transparent about the resulting DP specification, but others can be used to corruptly promote a high nominal level of protection (small PLB), while hiding shortcomings in the other building blocks. Needless to say, our intention is not to encourage unscrupulous behavior, but rather to expose the inherent weaknesses that are open for exploitation in what might appear to be an objective and mathematically absolute framework for privacy protection. These warnings may be particularly relevant to commercial implementations of DP, where conflicts of interest are commonplace (see e.g. [Waldman, 2021](#)). For example, when the data custodian and data user are the same entity (such as a tech company collecting data about their customers), it can easily be tempted to cut some differentially-private corners—or engage in some *privacy theatrics* ([Smart et al., 2022](#))—due to a desire to improve data utility while maintaining *prima facie* privacy protection to assuage its data contributors.

We start with the first ingredient of the DP flavor, the protection domain, \mathcal{X} . The less to protect, the less protection budget needs to be spent. Therefore, choosing or interpreting \mathcal{X} more restrictively may

lead to a smaller PLB, even without actually adding more noise to the data. This potentially perverse incentive should receive more scrutiny, because in most practice, the choice of \mathcal{X} is a conscious one on the data custodian’s part, and is indispensable to the interpretation of the DP specification. The domain \mathcal{X} frames the variables that are contained in the dataset and thus depends upon the socio-cultural sensitivity of these variables (Nissenbaum, 2010). More generally, every choice of \mathcal{X} encodes a data conceptualization (Leonelli, 2019)—a representation *by* the data *of* the individuals who contributed their information. Viewing the data release mechanism as a constituent phase in a data life cycle, \mathcal{X} specifies the starting point of that cycle. The impact of this choice permeates through other phases in the cycle, notably before data privatization including coding, cleaning, imputation, (sub-)sampling, and so on (Meng, 2021). Restrictions imposed by each of these steps may impact \mathcal{X} before the privatization step, thus affecting the PLB. Take the concrete example of topcoding and clamping, often performed as a data pre-processing step (see e.g. Kamath et al., 2023). By rounding or projection, the operation forces data to take value in an \mathcal{X} that has a bounded range, effectively reducing the sensitivity of any subsequent data release mechanism, allowing for a smaller declared PLB for the same magnitude of noise introduced.

The second way to apparently spend less PLB without adding more noise is to change the multiverse \mathcal{D} , the second ingredient of the DP flavor. Piling on more invariants, i.e., summaries of data that will be published exactly by the data-release mechanism, is one of such examples, because it creates a more shattered data multiverse \mathcal{D} . (A multiverse \mathcal{D} is a shattering (or refinement) of another multiverse \mathcal{D}' if every $\mathcal{D} \in \mathcal{D}$ is a subset of some $\mathcal{D}' \in \mathcal{D}'$.) For those who appreciate DP as a framework for protecting the *relative* privacy or information, this possibility is rather obvious. The more one discloses via the invariants, the less information left in the data that require protection, and hence a smaller PLB is incurred. Take

swapping as a concrete case: as shown in Part II and briefly recapped in Subsection 4.4.1, the PLB ε of the Permutation Swapping Algorithm is determined by the swap rate p and the largest stratum size b . To decrease the nominal value of ε , one can either increase p (up to a point) or decrease b . When the dataset has a fixed size, the simplest way to decrease b is to define the stratifying variables at a finer resolution, resulting in smaller strata within which swapping is confined. As illustrated in Part II, the various choices of the stratifying variables at different levels of geography, with or without crossing with the household size variable, result in b ranging from as small as 11.7 thousand to as large as 13.7 million, and a nominal ε from 12.31 to 19.38 (respectively) at $p = 5\%$.

A third way to achieve a nominal reduction of the PLB is to redefine the protection units—as captured by the third ingredient of DP flavor, $d_{\mathcal{X}}$ —at a finer granularity. With all else being equal, a DP specification with coarser protection units packs more weight in its PLB; it offers a stronger protection guarantee at the same nominal budget than a specification with on finer protection units. In the opposite direction, when a more expansive protection unit is supplanted by a narrower one, the input premetric $d_{\mathcal{X}}$ becomes inflated in that a unit of change in the former sense may amount to multiple units of change in the latter sense, “watering down” the PLB by the same amount.

This maneuver has been recognized, and to some extent utilized, by the literature in privacy mechanism design for complex data structures. For example, the choice of neighbors is particularly important for network data – are neighbors defined by removing a node or an edge from the network, that is, are privacy units edges or nodes (Raskhodnikova and Smith, 2016)? For business databases, does a company constitute a unit, or should units be employees, or both (Haney et al., 2017; Schmutte, 2016; He et al., 2014)? Or should they be the company’s transactions? Similarly for large personal databases in commer-

cial settings, should an individual constitute a unit, or should each of their interactions with the platform (such as a post or a ‘like’) be privacy units, or should units be the set of a user’s interactions within a given time period (e.g. a single day) (Kenthapadi and Tran, 2018; Messing et al., 2020b; Desfontaines, 2023)? Finally, when publishing social statistics, do households deserve privacy protection above and beyond the protection afforded to their individual members (Machanavajjhala, 2022)?

A fourth way to gain nominal PLB out of thin air is to artificially introduce an output divergence D_{Pr} that systematically assesses two distributions to be closer. Technically speaking, the relaxation from ε -DP to (ε, δ) -approximate DP (ADP) (Dwork et al., 2006a) can be understood as a maneuver of this type. (This is not to say that ε, δ -ADP is never a valid choice—in certain situations the gains to data utility may legitimately outweigh the loss to SDC of adopting (ε, δ) -DP. However, often there are other choices which allow for the same gains in utility while requiring that a data release mechanism ‘fail gracefully’ rather than allowing a non-zero probability of ‘catastrophic failure’ (see e.g. Near and Abuah, 2025, Chapter 7).) This can be seen by writing out the multiplicative divergence for (ε, δ) -approximate DP, as in Part I:

$$D_{MULT}^{\delta}(P, Q) = \sup_{S \in \mathcal{F}} \left\{ \ln \frac{[P(S) - \delta]^+}{Q(S)}, \ln \frac{[Q(S) - \delta]^+}{P(S)}, 0 \right\}, \quad (4.3)$$

where $[x]^+ = \max\{x, 0\}$, P and Q are two probability measures on the same output space of the data-release mechanism T , and \mathcal{F} is a collection of events of T that are of interests and that permit logically coherent probabilistic assignment and operation, technically known as σ -algebra. But clearly, for any $\delta > 0$, $D_{MULT}^{\delta}(P, Q) < \sup_{S \in \mathcal{F}} |\ln P(S) - \ln Q(S)|$, which is the output premetric for pure ε -DP. Hence, we have reduced the PLB without changing the actual data release mechanism.

Because the PLB is used to bound the worst-case rate of change, as seen in (4.1), any strategy of reducing the extremeness of the worst-case will permit a smaller PLB without injecting any additional noise. This

opens doors for manipulation or misuse. Specifically, we can reduce the σ -algebra \mathcal{F} , by pretending that we are interested in fewer events for the outcome T . This can be seen clearly in (4.3), where the right-hand side is the largest possible value over all events in \mathcal{F} (denoted by $S \in \mathcal{F}$). If we reduce \mathcal{F} to a sub- σ -algebra, then this extreme value will deflate, affording us with a smaller PLB.

The concept of subspace differential privacy (Gao et al., 2022) reflects a maneuver of this type, for it requires control over the output divergence for every set in the Borel σ -algebra generated by a linear subspace of the ambient space. Coarsening the σ -algebra associated with the output space signifies a weaker standard against which the data release mechanism is held, and does not compel the mechanism to be non-measurable with respect to another richer σ -algebra. Therefore, for subspace differential privacy, the mechanism could still take values in the ambient space. This is also different from the requirement of invariants, which operates on the input space rather than the output space of the data release mechanism T .

To emphasize further the importance of understanding the vulnerability and subtlety of DP, consider the case where we use a constant data release mechanism, that is $T(x) = C$, where C does not change with x . Clearly such a data release mechanism has zero PLB because it is completely insensitive to any manipulation of the input data x . But what if the data curator chooses C to be the same value as the very data or query we try to protect? Surely that means $\text{PLB}=\infty$, since the actual query or data is disclosed. Whereas this may appear to be a pathological case, it carries a critical message: the design of the data release mechanism cannot be permitted to depend on the confidential data set itself. This is the very reason that for 2020 Census, the US Census Bureau used the original data from 2010 Census, in its construction of the 2020 TopDown Algorithm, instead of from the 2020 Census itself (Abowd and Hawes, 2023).

What has been discussed thus far does not nearly exhaust all perverse means to game the PLB. How-

ever, our message is the opposite of perverse: the privacy loss budget is contextual in nature. Its context in the narrowest sense encompasses the choice of parameter in a privacy specification and a data release mechanism. These aspects of a DP guarantee deserve proper recognition.

4.4 HIGHLIGHTS OF PART II: THE US CENSUS’S EVOLVING DATA PROTECTION

The Decennial Census of Population and Housing is a critical piece of US infrastructure. It determines the apportionment of seats in the House of Representatives; it is relied upon for allocating trillions of dollars in federal funding each year; and it informs the decision-making of businesses, urban planners and hospitals, amongst many others (Villa Ross, 2023; Reamer, 2019; National Research Council, 1995). Safeguarding such an important data source requires robust SDC, a task that the US Census Bureau has long taken seriously. In fact, the bureau has been managing the risk of indirect disclosure in the Decennial Census from 1940 onward (US Census Bureau, 2019b). In the 1990 Census, protections were strengthened and a new SDC method was introduced: *data swapping* (McKenna, 2018; Dalenius and Reiss, 1982; Fienberg and McIntyre, 2004).

Data swapping (or record swapping) is a general concept that encompasses a broad class of algorithms. These algorithms select a set of records and then shuffle the values of certain variables among these records. We call the variables whose values are shuffled the *swapping variables*; all other variables are called the *holding variables*. Usually, records are partitioned into groups (or strata) according to the values they take on a subset of the holding variables we call the *matching variables*; and records are only shuffled within their matching group.

The primary SDC methods used in the 1990, 2000 and 2010 Censuses were forms of data swapping, the

full technical details of which have not been made public due to confidentiality concerns. We know however that the bureau swapped entire households, rather than shuffling person-level data between households; that the swapping variable was geographic (e.g. block group, tract, or county); and that the matching variables included broader levels of geographies (i.e. tract, county or state) as well as the household's total counts of adults and children. Furthermore, unique or unusual households that the bureau believed had higher disclosure risk had a higher chance of being swapped.

In the 2010s, spurred by an increasing awareness of privacy risks in statistical products (Dinur and Nissim, 2003), the US Census Bureau conducted a *reconstruction attack* on the 2010 Census. Using published tables, and publicly-available information on the relationships between these tables, they were able to determine with a high degree of confidence much of the underlying post-swapped microdata which produced these tables (Abowd et al., 2023). This led to a revolution at the bureau – the 2020 Census would not be protected by using data swapping, as was the case for the previous three decades, but rather by brand new SDC methods which were explicitly designed to satisfy DP (Abowd, 2018). Yet, does the USCB's official adoption of DP, on its own, truly represent a sea change in how it protects the census? What if, in fact, the census was already protected in 2010 by a DP method – or at least by a method which is very similar to a DP one?

4.4.1 A DP GUARANTEE FOR THE 2010 CENSUS?

While data swapping was not originally invented with DP in mind, it may still be possible for it to satisfy some DP specification. However, it is effectively a method for adding noise *only* to the relationship between the swapping and holding variables within each matching group. As such, the marginal distributions of these three sets of variables are *invariant* under swapping, as is the joint distribution of the swapping and

matching variables. Data swapping can therefore only be DP subject to the invariants it induces – i.e. it can satisfy a DP specification only if the specification’s data multiverse respects its invariants. This limiting of the scope of protection greatly reduces the SDC guarantee provided by DP, as we discuss extensively in the following section.

Moreover, some of the technical implementation details of the 2010 swapping procedure preclude it from satisfying a pure ϵ -DP specification (i.e. a specification whose output premetric D_{Pr} is the same as the one used in the original DP specification of [Dwork et al. \(2006b\)](#)). Nevertheless, the *Permutation Swapping Algorithm* (PSA) – which keeps to the spirit of the 2010 procedure, if not the exact implementation² – does satisfy pure ϵ -DP subject to the invariants it induces.

The PSA is very simple to describe: It selects records independently with probability p (the ‘swap rate’) and then permutes the values of those records’ swapping variables. Incidentally, this idea (under the name *n*-Cycle swapping) was under active investigation by the USCB up until the bureau redirected its research efforts towards DP ([McKenna and Haubach, 2019](#)).

The following statement is an informal version of the main result of Part II, which proves that the PSA satisfies a DP specification.

Theorem II.1 (informally). *Subject to the invariants induced by it, the PSA is ϵ -differentially private, with*

$$\epsilon \leq \begin{cases} \ln(b+1) - \ln o & \text{if } 0 < p \leq 0.5, \\ \max \{ \ln o, \ln(b+1) - \ln o \} & \text{if } 0.5 < p < 1, \end{cases}$$

where $o = p/(1-p)$, and b is the size of the largest matching group.

To contextualize the protection loss budget ϵ in Theorem II.1, we briefly describe in non-technical

²We point the interested reader to Part II for a detailed comparison between the PSA and the 2010 swapping procedure.

terms the first four building blocks of the PSA’s DP specification. Firstly, the protection domain \mathcal{X} is any set of datasets which all share the same variables. Secondly, what “subject to the invariants” means is that among all possible datasets in \mathcal{X} , the only ones that carry the protection guarantee are the ones that share the same invariant values with the actualized, confidential dataset. This is because other datasets will be excluded from consideration by any competent attacker due to the fact that their invariant values are different to those published. Mathematically, the invariants partition the data space \mathcal{X} into data universes, with all datasets in each universe sharing the same invariant values. By subjecting DP to the PSA’s invariants, we mean that DP’s Lipschitz condition (2.1)) is restricted to datasets in the same universe. Hence, comparisons between datasets with different values on the invariants are excluded from consideration.

Thirdly, the granularity of the PSA’s DP specification is equal to the resolution of the PSA’s swaps. For example, if the PSA swaps person records, then the granularity of protection is person records. More technically, we mean that the PSA’s input premetric $d_{\mathcal{X}}$ is the Hamming distance on person records. Thus, if a number n of person records are changed, the distribution of the PSA’s output will change by at most εn units, as measured by the standard of protection for the PSA’s DP specification (assuming that the changes to these records do not result in a change in the invariants). (Fourthly) this standard of protection is captured by the maximum likelihood ratio – that is, the maximum value, over all possible outputs t , of the relative likelihood of observing output t , under input dataset x as compared to under the some alternative input x' .

In order to get an (approximate) description of the 2010 Census’s SDC protection, we instantiate the PSA’s parameters to align as closely as possible with what we know about the 2010 swapping procedure. This gives us a concrete DP specification for the 2010 Census (unlike the above specification, which gener-

ically applies to all instantiations of the PSA), under the counterfactual scenario that this census was protected by the PSA. We emphasize that this does not provide a DP guarantee for the actual 2010 Census since such a guarantee must reckon with the exact implementation details of the 2010 SDC methods, not the PSA. However, because we believe the PSA can closely parallel the 2010 swapping procedure by appropriately choosing its implementation parameters, the resulting DP specification is nevertheless a useful perspective on the protection provided to the 2010 Census.

Our choices for the PSA’s implementation parameters are as follows. To integrate the PSA into the 2010 Census data pipeline, we place it after all imputation and editing processes. This means the PSA’s protection domain is the set \mathcal{X}_{CEF} of all possible *Census Edited Files* – i.e. all hypothetical outputs resulting from the first stages of the Census data pipeline through to the imputation and editing processes. This has important implications on what data is actually being protected by the PSA: the edited and imputed records, not the ‘raw’ Census responses. Moreover, because \mathcal{X}_{CEF} determines what it means to counterfactually alter data, it also has implications on what the protection unit in 2010 was. Indeed, mirroring the 2010 swapping procedure, we set the PSA to swap household-level records, so that its input premetric $d_{\mathcal{X}}$ is the Hamming distance on household records; yet this does not imply the 2010 protection units were households – because a change in a single record of the Census Edited File does not always correspond to a single household changing their Census responses. Instead, the 2010 protection units are ‘post-imputation households’ – imaginary entities which can alter their own records in the Census Edited File freely without affecting other, imputed records. (See Part II for an explanation of how the PLB must be inflated in order to have households as the protection unit.)

We set the PSA’s matching variables to be the household’s state and size, and its swapping variables to

be the household’s county. This results in a multiverse \mathcal{D}_{2010} where all statistics at the state and national levels are invariant, as well as the counts of households by size at the block level. Finally, we set the swap rate p to be 2-4%, as [boyd and Sarathy \(2022\)](#) states was used in 2010. This results in a PLB ε between 18.29 and 19.

Since the 2010 swapping procedure also included the number of voting-aged people as a matching variable, 18.29-19 is an upper bound for 2010’s approximate PLB. (We cannot compute the PLB when the number of adults is invariant because the necessary statistic—the value of b in Theorem II.1—is not publicly available.) However, this also implies that \mathcal{D}_{2010} gives a lower bound on the invariants: in addition to the statistics reported above, the block-level breakdown of households by the number of adult occupants (and hence also by the number of children occupants) were invariant.

4.4.2 THE DP GUARANTEE OF THE 2020 CENSUS

Having established that the 2010 Census may be analyzed from the perspective of DP, it is fruitful to compare it with the DP specification of the 2020 Census. We will start this comparison by examining each of the five DP building blocks in turn. As in 2010, the 2020 disclosure avoidance system (DAS) also took the Census Edited File as input. Hence the protection domain remains constant across the two census; in both cases it is the set of all possible Census Edited Files \mathcal{X}_{CEF} . Secondly, the granularity of protection in 2020 was person-level records. All other components being equal, this would imply weaker protection than in 2010, which protected household-level records; but, as we will see, the three remaining components are not equal.

Most importantly, the 2020 DAS has far fewer invariants than was the case in 2010. The 2020 invariants were carefully considered and minimized to those required by operational and constitutional mandates.

These invariants are the state populations as well as the counts at the block level of housing units and of each type of occupied group quarters. As we will discuss in Section 4.6, initial analysis suggests that these invariants have minimal effect on SDC; the same cannot be said about 2010’s invariants (Abowd et al., 2023).

The standard of protection used in 2020 is what we call the *normalized Rényi metric*; this is the choice of output premetric corresponding to zero-concentrated DP (zCDP) (Bun and Steinke, 2016). Under these settings for the first four components, the 2020 Census’s PLB is given by $\rho^2 = 55.371$. (We follow the standard convention of using ρ to denote the PLB in the case where the standard of protection corresponds to zCDP.) Additionally, we may translate from zCDP to (ε, δ) -DP and thereby also express the 2020 PLB by $\varepsilon = 126.78$ with $\delta = 10^{-10}$. (To be clear, in this translation across DP specifications, the first three building blocks stay the same, while the output premetric changes from the normalized Rényi metric to the δ -approximate multiplicative divergence.)

It is worth noting that the above DP specification only assesses the disclosure risk associated with the primary 2020 Census products (namely, the P.L. 94-171 Redistricting Summary File (US Census Bureau, 2021a,b), the Demographic and Housing Characteristics (DHC) File (US Census Bureau, 2023c), the Detailed DHC-A and -B Files (US Census Bureau, 2023f, 2024a), the Supplemental DHC (US Census Bureau, 2024c) and related auxiliary products—see Part II). As of the time of writing, there have already been additional releases, and there will be future releases which rely on the 2020 Census data (e.g., the annual Population and Housing Unit Estimates and the National Population Projections (US Census Bureau, 2023q,p)). While we have not been able to obtain information on their SDC, these releases will at a minimum increase the 2020 PLB. They may also possibly weaken the other components of the 2020

DP specification. This would not happen for data swapping; the 2010 DP specification from the previous subsection would cover all Census products because they were all generated from the post-swapped microdata (see Subsection 4.5.2).

Beyond the core details of the 2020 DP specification explained above, several other aspects merit attention. Firstly, as for 2010, setting the protection domain to be \mathcal{X}_{CEF} in 2020 has important implications: The 2020 Census does not have “end-to-end” DP protection (c.f. [Hu et al., 2024](#)); its protection units are ‘post-imputation persons’ and as such does not provide protection to individuals’ Census responses directly. Secondly, a more nuanced perspective on the 2020 DAS would examine its *per-attribute* PLBs ([Ashmead et al., 2019](#)). A per-attribute analysis considers a DP specification in which only one variable (i.e. attribute) is allowed to vary within each data universe. This allows for a more fine-grained assessment of SDC, rather than assuming the worst-case possibility of complete dependence between variables when composing the per-attribute budgets into a single total budget. Apart from the following two observations, we leave this important discussion to future work: The per-attribute budgets are much smaller than the overall 2020 PLB. And a per-attribute analysis is not applicable to data swapping since its DP specification does not rely on the composition theorem of DP.

4.5 WHAT DOES IT MEAN IF SWAPPING IS DIFFERENTIALLY PRIVATE?

4.5.1 DIFFERENTIAL PRIVACY BESTOWS ADVANTAGES

The PSA’s DP specification gives a precise, mathematical formulation of the SDC it provides. It delimits the information the PSA does not protect (its invariants) and the extent to which it protects the remaining information. It describes the granularity at which attackers are limited in learning about aspects of the confidential data that are not disclosed by the invariants alone, and the standard against which this learning

is measured. In short, the PSA's DP specification answers the 'who', 'where', 'what', 'how' and 'how much' questions of SDC – a precursor for determining whether the PSA is appropriate for a given data release, and, if so, for choosing its implementation parameters.

DP is not just a descriptive framework. It also provides a calculus for reasoning about how protection loss accumulates across multiple data products, a tool which is becoming increasingly more valuable as national statistical offices (NSOs) diversify their offerings (Kitchin, 2015). Moreover, adopting a DP flavor as the yardstick for measuring SDC permits complete transparency of the data release mechanism, since DP guarantees do not degrade with the attacker's *knowledge of the mechanism* (in contrast, degradation can occur with the attacker's knowledge of the relationships among the data subjects). For example, even when armed with the complete knowledge of the PSA's implementation details (including the values of all its parameters, such as the swap rate or swap key, but excluding, of course, the value of its random seed), it is still impossible for an attacker to thwart the SDC protections, as measured by its DP specification. While transparency does not assist attackers in breaking DP's protection guarantees, it is important to legitimate data users as an essential prerequisite for valid statistical analysis of privacy-protected data (Gong, 2022b). Indeed, by allowing quantitative analysts such as statisticians and social scientists to correct for the statistical errors induced by SDC protection, transparency increases data utility and supports robust research findings.

As Section 4.6 will discuss in more detail, the claim that the details of a DP-compliant method can be disseminated at no cost to privacy rests on two assumptions. Barring this complication, transparency – as provided by recasting SDC techniques as DP – will be a major development because the details and implementation parameters of SDC algorithms have traditionally been kept secret. For example, the currently

available public documentation on the 2010 DAS is deliberately deficient, stymieing researchers’ ability to appropriately account for the noise it injects into Census data (Kenny et al., 2024). Our work provides the necessary framework to justify the publication of a comprehensive description of the 2010 DAS swapping procedure, without further degrading its protection beyond the loss attributable to the invariants (which have, by and large, already been made public; see e.g. Abowd and Hawes (2023)). Assuming that the knowledge of what published 2010 statistics are invariant is not itself a disclosure risk, the transparent knowledge of the 2010 DAS will not only provide crucial retrospective insight into its quality (and the quality of other data products subject to similar SDC protection protocols), but will also further an open discourse regarding the optimal SDC standards of official statistical agencies, including design parameters for the disclosure avoidance system of future data products.

Swapping mechanisms in particular have received criticism since they have been shown theoretically to introduce bias into the published data (e.g. Drechsler and Reiter, 2010). But the level and nature of this bias depends on the particular swapping algorithm used and its implementation parameters. Only with transparency of the 2010 swapping algorithm—as enabled by a formal privacy analysis—can the extent of this bias be quantified. This would provide belated yet crucial insight into the quality of the past Decennial Census data treated with swapping, above and beyond what the current theoretical understanding can provide.

4.5.2 A DIFFERENTIALLY PRIVATE METHOD WITH SOME OF THE BENEFITS OF TRADITIONAL SDC?

Traditional SDC techniques also have their own value, with data swapping in particular enjoying advantages that most DP methods do not. For example, swapping maintains *facial validity*—the 2010 Census outputs all look ‘reasonable’ in the sense that there are no negative or fractional counts, nor are there im-

plausibly large or small reported values. More generally, the 2010 publications pass the sanity checks an observant reader might make, which is useful for building trust in the census among the general public. From the opposite perspective, a lack of facial validity is an important concern for statistical agencies like the USCB. It presents an issue for data users who are confused and disinclined to use seemingly-erroneous data; it erodes the public image of the agency; and it hampers efforts to improve differential response rates among disadvantaged communities (boyd and Sarathy, 2022; Drechsler, 2023; Oberski and Kreuter, 2020). (Related, but distinct, to facial validity is the concept of *face privacy*, which is the requirement that the data release mechanism produce output which appears to the casual observer to offer privacy protection (Hod and Canetti, 2025). In contrast, facial validity requires that the output is a plausible representation of the real world, even to a data user who is unaware that artificial noise was added for SDC protection.)

Logical consistency is another advantageous property of data swapping, as it ensures all statistics produced from the 2010 Census align with one another. For example, in any contingency table released in 2010, the sum of cells in a single row or column always matches the marginal total. Likewise, reported values for the same count remain consistent across different publications. Additionally, the 2010 Census outputs respect the structural zeroes and edit constraints present in the underlying confidential microdata. In contrast, many of the 2020 Census outputs will not be consistent across publications and even within the same publication, row- and column-sums will not match the reported totals. (However, the outputs produced by the TDA—the PL and DHS files—are logically consistent, although as we will see, this comes at a cost.)

Like facial validity, logical consistency is important for users and advocates of census data (boyd and Sarathy, 2022; Hotz and Salvo, 2022; Ruggles et al., 2019). Yet, many DP methods do not maintain fa-

cial validity nor logical consistency, and others, such as the TopDown Algorithm, cannot satisfy these properties without partially destroying statistical transparency.³ This is because the majority of DP methods rely on optimization-based post-processing to restore facial validity and logical consistency (e.g. Barak et al., 2007; Hay et al., 2010). Optimization-based post-processing can be algorithmically transparent but in most cases it destroys the statistical transparency of the resulting two-step privacy mechanism – a crucial requirement for principled statistical analysis (Gong, 2022b). The recent proposal by Dharangutte et al. (2023) does away the need for post-processing when the noise infusion is additive. However, it relies on MCMC sampling and hence is non-trivial to implement for large-scale data products. In contrast, swapping achieves facial validity and logical consistency automatically without the need for additional computation.

Furthermore, data swapping—like other traditional SDC techniques but unlike many DP methods (such as those used in 2020)—is easy to communicate and understand at a high level by a broad, non-technical audience. This is important for building trust and maintaining the buy-in of data providers, custodians and other stakeholders. Swapping is also easily implementable and amenable to the types of data collected by government agencies, as evidenced by its use in the US, the UK and the EU (McKenna, 2018; Office for National Statistics, 2023; de Vries et al., 2023).

Finally, as a pre-tabular perturbation method, swapping also has the advantage of producing a ‘synthetic’ dataset that serves as the source for all census publications. This simplifies the data release process as all outputs are derived from this ‘post-swapped’ data without requiring additional SDC treatment. This

³A data-release mechanism T is statistically (or probabilistically) transparent if the conditional probability distributions $P_x(T \in \cdot)$ are public knowledge (Gong, 2022b). Statistical transparency is distinct to algorithmic transparency, which requires that the source code of the mechanism T is disseminated. For all practical purposes, statistical transparency is a stronger requirement than algorithmic transparency since T ’s source code may be so complex that it is practically impossible to derive the conditional distribution it induces.

also explains why the 2010 publications maintain logical consistency; because no further noise is introduced, all outputs are consistent with the post-swapped data and hence also with each other. Moreover, this approach ensures that releasing additional data products does not degrade the PSA's DP specification. As long as all products are based on the same post-swapped microdata, they are all covered under this specification by DP's post-processing theorem. In this way, data swapping allows the statistical agency to publish a single DP specification which encompasses all existing and future publications. As mentioned in Subsection 4.4.2, this stands in contrast to the 2020 Census. There, each publication has its own DP specification, and to understand the SDC provided to the census data as a whole, one must aggregate these DP specifications into a single comprehensive one – a process which must be repeated with each new publication. And, because every release introduces additional disclosure risk, the overall 2020 DP specification weakens each time, in comparison to the single, upfront DP specification associated with data swapping.

4.6 INVARIANTS, TRANSPARENCY AND DATA UTILITY

A concrete benefit of the new perspective we provide is that it sheds light on debates concerning swapping. In what follows, we review and provide our comments on three current discussions. We will argue that most of these contentious issues are tangential to the fundamental nature of swapping and DP noise infusion as mechanisms for data privacy protection. Our work provides a level playing field that allows for a much needed, informed, and fair comparison.

4.6.1 UNDERSTANDING THE IMPACT OF INVARIANTS ON DISCLOSURE RISK.

A major criticism of the swapping method implemented in the 2010 Census is that it induces too many invariants. One salient consequence of the plurality of invariants is that the permissible values for the confidential data are severely constrained. As a result, it can be impossible to simultaneously maintain

a low degree of data disruption and control the risk of identification via swapping. The most damning source of identification risk pertains to the population uniques which, if exist, can be directly revealed as logical consequences of the invariants. As an extreme example, any swap-key stratum with only duplicate records would result in an exact reconstruction of that stratum. As [Abowd and Hawes \(2023\)](#) discuss, the invariants in the 2010 Census swapped data elevate disclosure risk, because 1) total and voting age populations at the block level constitute information at very fine granularity, and 2) the existence of a high fraction of unique persons within blocks (57%) further facilitates reidentification via record linkage.

The bureau carried out a suite of simulated reconstruction attacks against the 2010 Census and observed high rates of reidentification ([Abowd et al., 2023](#)). Generally speaking, reconstruction attacks work by collating many aggregate statistics about the confidential (unknown) microdata and then constructing a database which agrees with these statistics. This database is a plausible guess for the confidential microdata, since it generates identical statistics to the ones generated by the microdata. The larger the number of such statistics and the more accurate they are, the more heavily they constrain the possible configurations of the reconstructed database, and hence the more likely this reconstruction is to agree with the true confidential microdata. As a result, it is easier to create reconstructed databases with a high chance of leading to the reidentification of units via linking to external data sources. The experiments further suggest that the rate of swapping must be significantly increased to achieve what can be deemed as an acceptable level of protection for the population uniques ([Abowd and Hawes, 2023](#)). It is from these observations the bureau concluded an urgent need to revamp swapping-based SDC. However, the question remains open: in what ways does a specific set of invariants impact the disclosure risk of the resulting data product, and the effective privacy guarantee it can afford?

Before we remark on how the above conjectures and empirical evidence may implicate the relationship between invariants and disclosure risk, two things are worth noting at the outset. First, the degree of vulnerability of a privacy-protected data product against a class of reconstruction attacks is a measure of its *absolute disclosure risk* (Duncan and Lambert, 1986; Reiter, 2005), defined as the extent of certainty with which an agent can make inferences about the confidential information from the data product. It is well understood that unless strong assumptions about the agent’s prior knowledge are made, differential privacy does not directly translate into any quantifiable degree of control over the absolute disclosure risk; see e.g. Dwork (2006); McClure and Reiter (2012); Kenny et al. (2021); Hotz and Salvo (2020). Therefore, the success of reconstruction attacks against a data release mechanism, regardless of the privacy guarantee they bear (or not bear), should be taken as indirect evidence if it is to be contrasted with the result against a differentially private mechanism characterized by its design parameters.

Second, invariants are not a proprietary consequence of swapping. Whether the data custodian implements swapping or another privacy protection mechanism, to maintain some invariants in the data product is unavoidable. The production settings of the TopDown Algorithm employed invariants as specified in Subsection 4.4.2. Notably, the Census Bureau arrived at this final list of invariants through an iterative process. For example, among the bureau’s previously considered invariants are block-level population invariants; see Ashmead et al. (2019); Kifer (2019). The often iterative process of determining invariants illustrates that it is often a part of many privacy mechanism designs and parameter choices.

Notwithstanding the caveats, it is worthwhile to inquire, to the extent possible, about the impact of invariants on disclosure risk through the lens of DP. Such an inquiry can be challenging within the classic differential privacy paradigm, because invariants are not captured by the privacy loss budget, the sole

measure of privacy guarantee. By contrast, the DP specification that we lay out in this work is more dexterous. Specifically, Definition 2.4.2 reveals that invariants are captured by the data multiverse \mathcal{D} . Under this specification, an analysis of the impact of invariants on disclosure risk amounts to a five-dimensional comparison between alternative privacy specifications that differ on \mathcal{D} and potentially on other elements as well. A comprehensive description of the five-way dynamics remains open for future research, though investigations with a restricted scope can already be informative if concrete and feasible alternatives are contrasted. For example, it can be shown that reconstruction attacks can be increasingly successful if applied to DP data when more invariants are imposed on them (Protivash et al., 2022). Our analysis of Section 3.4.3 also indicates that the granularity of invariants has a larger numerical impact on the privacy loss budget ϵ , more so than the swap rate for a given set of invariants, suggesting that a reduction of the invariants may have a larger impact compared to an increase of swap rate.

4.6.2 MITIGATING THE IMPACT OF INVARIANTS ON DISCLOSURE RISK.

Because of the omnipresence of invariants and their potential adverse impact on disclosure risk, the modern data custodian are entitled to methodologies that allow for the specification of invariants in a flexible and precise manner, in order to design a tailored solution that balances privacy and accuracy targets. To this end, swapping – instantiated either as in the previous Decennial Censuses or as in this work – does not suffice. In addition to comparative investigations discussed previously that may instruct the trade-off of invariants with other dimensions of the DP specification, several tangible remedies may directly mitigate the impact of hard invariants and are worthy of exploration.

As part of its comparative analysis between swapping and TopDown, the Census Bureau considered methods to override the hard invariants due to swapping. One such method is *probabilistic unit matching*.

Instead of using V_{Match} to form hard strata that confine swapping, allow, with a small probability, for units across different strata to be swapped. The probability could be inversely proportional to some distance metric on the strata. As a demonstration, take the 1940 Census full count example from Section 3.2.4, where V_{Match} is the state indicator and size of the household and V_{Swap} is the county indicator. Suppose for some $\alpha > 0$, a household chosen for swapping would have a $(1 - \alpha)\%$ chance of being swapped with another household of the same size, but an $\alpha\%$ chance of being swapped with a differently-sized household. Doing so retains the county-wide household counts as invariant, but the county-wide total populations are no longer invariant.

Another pair of approaches to remove invariants in swapping is *pre-swap perturbation* and *post-swap perturbation*. As their names suggest, the former infuses noise into the confidential record prior to applying swapping (Hawes and Rodríguez, 2021, p. 23), whereas the latter perturbs an intermediate data product after applying swapping. Notably, data swapping followed by tabular perturbation is a common SDC strategy, as it is the approach taken by the Office of National Statistics (ONS) for the protection of the 2021 UK Census (Office for National Statistics, 2023). There, the cell key method (CKM) is employed to perturb the cells of contingency tables after targeted recording swapping has been applied to the underlying microdata (Fraser and Wooton, 2005; Thompson et al., 2013; Marley and Leaver, 2011). Notably, the CKM procedure has been analyzed through the lens of DP (Rinott et al., 2018; Bailie and Chien, 2019; Chien and Sadeghi, 2024). In addition to its use at the ONS, applying swapping and then CKM perturbation is also recommended by Eurostat’s *Centre of Excellence on Statistical Disclosure Control* for EU censuses (Glessing and Schulte Nordholt, 2017).

We leave to future work the investigation of the full theoretical guarantees of probabilistic matching

as well as pre-swap and post-swap perturbation. Note that compared to classic swapping alone, all of the above procedures induce strictly more auxiliary randomness into the data product. Therefore, it would be reasonable to expect the resulting algorithms to enjoy DP guarantees while supplying fewer and more flexible choices of invariants. One particularly salient question for this line of research is to determine the DP specification for the sequential composition of two mechanisms, when both mechanisms satisfy (possibly different) DP specifications.

4.6.3 DATA UTILITY UNDER TRANSPARENT PRIVACY

The argument that the details of DP methods can be disseminated at no cost to privacy rests on two underlying assumptions. The first assumption, which we will discuss further later in this section, is that the epistemic uncertainty of a SDC method is not a legitimate form of privacy protection. The second assumption, which we will discuss in more detail later, is that it is safe to disseminate the method’s DP specification, a non-trivial judgement when there are invariants involved. Barring these complications, transparency – as provided by recasting SDC techniques as DP – will be a major development because the details and implementation parameters of SDC algorithms have traditionally been kept secret. For example, the currently-available public documentation on the 2010 DAS is deliberately deficient, stymieing researchers’ ability to appropriately account for the noise it injects into Census data (Kenny et al., 2024). Yet our work suggests that publishing a complete description of the 2010 DAS swapping procedure (including its source code) will not degrade its privacy protection beyond the decrease associated with knowing the 2010 invariants (which have, by and large, already been made public). Transparency in this case would provide belated yet crucial insight into the quality of existing data products which were protected using the USCB’s swapping methods and would contribute to resolving the on-going debate between the 2010 and the 2020 DAS.

Publishing the USCB’s swapping methods will also help to dispel the “statistical illusions” (boyd and Sarathy, 2022) associated with US censuses before 2020. It is easier to argue that census data contain errors due to privacy protection when one can point precisely to how these errors were introduced and quantify their distribution exactly. Transparency enables concrete statements like, ‘the expected error due to privacy protection is X%’, in place of vague expressions such as, ‘these counts may have some error as their contributing households could have been swapped’. Raising awareness of the census data’s long-existing privacy errors by publicly documenting the DAS methods is a step towards “shifting the statistical imaginary to account for uncertainty” (boyd and Sarathy, 2022) and thereby improving the USCB’s legitimacy.

Lastly, transparency counters a common principle in traditional SDC: privacy through obscurity. This principle, which states that a data custodian should not fully reveal the implementation details of their SDC method, is based on the rationale that these details could be used by an attacker to unpick the method’s privacy protection (see Slavković and Seeman (2023) and references therein). Therefore, SDC’s privacy protections are not solely due to the aleatoric uncertainty introduced by random noise injection, but also due to the epistemic (aka structural) uncertainty in the attacker’s knowledge of the protection method. There is protection provided directly by the SDC method and then – according to this line of reasoning – there is protection provided by plausible deniability in exactly how the method was implemented. Epistemic uncertainty is, unfortunately, much harder than aleatoric uncertainty to model and reason about (see, for example, the extensive literature on imprecise probabilities (Shafer, 1976; Walley, 1991; Augustin et al., 2014)). Consequently, the privacy protection provided by the epistemic uncertainty of an SDC method is difficult to describe with mathematical privacy guarantees and, as far as the authors are aware, has not been systemically studied. Nevertheless, the principle of privacy through obscurity is

frequently invoked by NSOs in reference to their SDC methods (McKenna, 2018; UK Statistics Authority, 2021; Chipperfield et al., 2016). On the other hand, DP follows a long tradition in cryptography, dismissing epistemic uncertainty as a brittle form of protection which depends on a watertight security system to safeguard a privacy method’s implementation details. Since epistemic uncertainty’s protection can change widely with an attacker’s background knowledge and assumptions, DP’s privacy guarantees are derived solely from a mechanism’s aleatoric uncertainty, justifying DP’s tenet of ‘transparency for free’.

Our last point of discussion re-emphasizes another important motivation for this work, which was only mentioned briefly earlier. By casting swapping as DP, we can theoretically allow its algorithmic specification to be made public. As the main SDC method for the Decennial Census of the previous three decades, a peek into the technical specification of swapping can bring tremendous utility to data users and privacy researchers alike.

Data users who conduct statistical modeling with official data products criticize swapping because it negatively affects the quality of downstream data analyses. It has been well understood in the literature that swapping inflicts the most utility damage to the relationships between swapping and holding variables. Mitra and Reiter (2006) and Drechsler and Reiter (2010) demonstrate that even low swap rates (e.g. 5%) can substantially reduce the effective coverage of confidence intervals for the regression coefficient between such variables.

We surmise that the deterioration in coverage is in part due to performing a naïve regression analysis on processed data, without accounting for the privacy mechanism itself. As Gong (2022b) demonstrates, performing naïve regression analyses on data protected via DP noise-infusion results in similar types of coverage deterioration, and further that this deterioration can be restored once the privatization process is statistically modeled (at the expense of wider, though valid, intervals). However, the analyst cannot pos-

sibly be blamed for performing the naïve analyses on swapped data when the swapping procedure is not public. Unfortunately, swapping by tradition has not been a transparent SDC technique. The explicit statement of swapping’s privacy guarantees provides theoretical justification to publish the implementation details of the swapping procedure. This would allow the swapping mechanism to be appropriately accounted for via statistical modeling.

Note that the justification for transparency relies on the privacy guarantee being public. In the case of swapping (or for any DP specification whose data multiverse partitions the data space), publishing the privacy guarantee necessitates the release of its invariants. However, as we repeatedly emphasize throughout this work, there is danger of privacy leakage associated with the knowledge of invariants in and of themselves. For example, a plurality of invariants supply the adversary with confidence in their efforts to reconstruct the microdata and reidentify individual records. Therefore, careful deliberation has to be practised in weighing the cost of making public the invariants against the benefits of algorithmic transparency this allows.

PART II

STATISTICAL INFERENCE UNDER PRIVACY CONSTRAINTS

This page intentionally left blank.

5

General Inferential Limits Under Differential and Pufferfish Privacy^I

5.1 INTRODUCTION

THE WORLD TODAY IS WITNESSING an explosive growth of large-scale datasets containing personal information. Demographic and economic surveys, biomedical studies and massive online service platforms facilitate understanding of human biological functions and socio-behavioural environments. At the same time, they pose the risk of exposing confidential information about data contributors. Breaches of privacy can happen counter-intuitively and without malice. For example, [Homer et al. \(2008\)](#) demonstrated that even coarsely aggregated SNP (single-nucleotide polymorphisms ([Kim and Misra, 2007](#))) data from genome-wide association studies (GWAS) can still reliably reveal individual participants. This unsettling revelation led to the decision by the U.S. National Institute of Health to remove aggregate SNP data from

^IBased on work coauthored with Ruobin Gong.

open-access databases (Yu, 2013). This incident, and similar occurrences across science, government and industry Narayanan and Shmatikov (2008); Dwork et al. (2017); Culnane et al. (2019); Sweeney (2000), have attracted public attention and sparked debate about privacy-preserving data curation and dissemination.

Commensurate with the increasing risk of privacy breaches, the recent decades have also seen rapid advances in formal approaches to statistical disclosure limitation (SDL). These methodologies supply a solid mathematical foundation for endeavors that enhance confidentiality protection without undue sacrifice to data quality. Notably, *differential privacy* (DP) (Dwork et al., 2006b; Bun and Steinke, 2016; Kifer and Machanavajjhala, 2014) puts forth a class of rigorous and practical standards for assessing the level of privacy provided by a data release. Many large IT companies, including Google (Erlingsson et al., 2014), Apple (Apple Inc., 2017), and Microsoft (Ding et al., 2017), have been early adopters of DP. More recently, the U.S. Census Bureau deployed DP to protect the data publications of the 2020 Decennial Census (Abowd et al., 2022b). The U.S. Internal Revenue Service is also exploring differentially private synthetic data methods for the publication of individual tax data (Bowen et al., 2022). These decisions by statistical agencies and corporations showcase the growing popularity of DP among major data curators.

Innovations in privacy protection methods have prompted quantitative researchers to confront a new reality, as existing modes, practices and expectations of data access are subject to renewal. We highlight two points of tension in this development. First, DP promises *transparency*, in the sense that the design details about the protection method can be made public without degrading DP’s mathematical assessment of the level of privacy protection. Transparency is one of the advantages of DP over traditional SDL methods since it supports valid statistical inference by providing the analyst with the ability to model the privacy

noise. However, this promise often falls short in practice, leaving the statistician with tied hands [Gong \(2022b\)](#). Second, following the high-profile adoption of DP by the U.S. Census Bureau, a debate ensued concerning its interpretation, or its *semantics*, as well as its reconciliation with other notions of statistical disclosure risk; see e.g. [Kenny et al. \(2021\)](#); [Hotz et al. \(2022\)](#); [Kifer et al. \(2022\)](#); [Jarmin et al. \(2023\)](#); [boyd and Sarathy \(2022\)](#); [Francis \(2022\)](#); [Muralidhar and Domingo-Ferrer \(2023\)](#); [Garfinkel \(2023\)](#); [Sánchez et al. \(2023\)](#); [Keller and Abowd \(2023\)](#). These issues motivate theoretical investigations that may shed light on the pragmatic translation between rigorous privacy standards and usable statistical advice.

The current work takes multiple steps toward the resolution of these debates by examining DP via the lens of imprecise probabilities (IP). Our focus is restricted to two important flavors of DP: 1) the classic notion of pure ϵ -differential privacy (ϵ -DP) [Dwork et al. \(2006b\)](#), and 2) Pufferfish privacy [Kifer and Machanavajjhala \(2014\)](#), a conceptually-distinct variant of ϵ -DP that is showing semantic promise (see e.g. [Jarmin et al. \(2023\)](#)). We begin by describing ϵ -DP as a Lipschitz continuity condition (Section 5.2). This description enables the interpretation of ϵ -DP as an *interval of measures* ([DeRobertis and Hartigan, 1981](#)) induced by the data-release mechanism (Section 5.3). From here, we derive some implications of this interpretation on the problem of statistical inference using privacy-protected data releases. These results concern the probability model of the observable privatised data (Section 5.4), as well as frequentist hypothesis testing (Section 5.5) and Bayesian posterior inference (Section 5.6) using these data. Next we turn to address Pufferfish privacy (Section 5.7) – showing that it too can be described as a Lipschitz continuity condition – and discuss its semantic interpretation as limits to frequentist and Bayesian inferences in an analogous manner (Section 5.8). Further, we link Pufferfish to another IP object: the density ratio neighbourhood (Theorem 5.8.4). The results in Sections 5.4-5.8 establish bounds on key inferential ob-

jects while having general validity under mild assumptions about the data model, the privacy mechanism, and (when applicable) the analyst’s prior. Section 5.9 demonstrates that these results are optimal in the sense that the bounds we obtain are pointwise tight. Section 5.10 concludes the paper with a discussion.

Throughout this work, we demonstrate that various objects from the imprecise probability literature naturally arise when studying differential privacy. Specifically, definitions of DP often invoke *distortion models*: neighbourhoods of precise probabilities defined as closed balls with respect to some metric – or, more correctly, some *distorting function* [Montes et al. \(2020a,b\)](#). Moreover, the choice of the distortion model (partially) determines the flavor of the resulting privacy guarantee [Bailie et al. \(2025b\)](#). In the following sections, we will outline the appropriate distortion models formulations as they arise.

DP objects are naturally amenable to IP analysis. Indeed, the rich vocabulary of IP can help to articulate the properties of a DP object in a precise yet general manner. Within the current literature on DP, the study of data privacy protection using IP is a nascent endeavour. [Komarova and Nekipelov \(2020\)](#) examines the issue of partial identification in inference from privacy-protected data, where in certain situations the identification set can be described with a belief function. In [Li et al. \(2022\)](#), the authors formulate local differential privacy definitions for belief functions, a proposal that amounts to a set-valued SDL mechanism whose probability distribution is given by the mass function associated with a belief function. [Liu et al. \(2023\)](#) examines constraints on DP mechanisms in terms of belief revision and updating. On the matter of using IP to explore mathematical formalisations of data privacy, we will return and remark on a few concrete potential future directions in the discussion (Section 5.10).

5.2 PURE ε -DIFFERENTIAL PRIVACY

Define the data universe \mathcal{X} as the set of all theoretically-possible observable datasets. Let d be a metric on \mathcal{X} .² Given confidential data $x \in \mathcal{X}$, consider releasing some (potentially randomised) summary statistic $T \in \mathcal{T}$ of x . To formalise this, equip the set \mathcal{T} with a σ -algebra \mathcal{F} and define a *data-release mechanism* as a function $M : \mathcal{X} \times [0, 1] \rightarrow \mathcal{T}$ whose inputs are the confidential data x and a random *seed* $U \in [0, 1]$, and whose output is the summary statistic T . (We require that $M(x, \cdot)$ is $(\mathcal{B}[0, 1], \mathcal{F})$ -measurable for each $x \in \mathcal{X}$, where $\mathcal{B}[0, 1]$ denotes the Borel σ -algebra on $[0, 1]$.)

A distribution on the seed U induces a probability on the summary statistic $T = M(x, U)$. Without loss of generality, we may take $U \sim \text{Unif}[0, 1]$. Denote by P_x the probability measure of $M(x, U)$ induced by U , taking x as fixed:

$$P_x(M(x, U) \in S) = \lambda(\{u \in [0, 1] : M(x, u) \in S\}), \quad (5.1)$$

where λ is the Lebesgue measure on $[0, 1]$.

The realised value of the seed U and the observed dataset x are assumed to remain secret, while all other details of M (including the distribution of U) may – and should – be made public (Gong, 2022b). An attacker is tasked with inferring x (or some summary $h(x)$ of x) based on observing a draw $T = M(x, U) \sim P_x$. This set-up is analogous to fiducial inference Hannig et al. (2016), with x taking the role of the parameters, T the data, and M the data-generating process.

Pure ε -DP is a Lipschitz condition on M :

Definition 5.2.1. Given a data universe \mathcal{X} equipped with a metric d , a data-release mechanism $M : \mathcal{X} \times$

²Throughout this work, we allow metrics to have codomain $[0, \infty]$ rather than the more standard $[0, \infty)$. We precisely define a metric in Definition C.3.1 of Appendix C.3.

$[0, 1] \rightarrow \mathcal{T}$ satisfies (*pure*) ε -differential privacy if, for all $x, x' \in \mathcal{X}$,

$$D_{\text{MULT}}(P_x, P_{x'}) \leq \varepsilon d(x, x'), \quad (5.2)$$

where

$$D_{\text{MULT}}(\mu, \nu) = \sup_{S \in \mathcal{F}} \left| \ln \frac{\mu(S)}{\nu(S)} \right|,$$

is the *multiplicative distance*³ between measures μ, ν on $(\mathcal{T}, \mathcal{F})$.

The smallest Lipschitz constant $\varepsilon \geq 0$ which satisfies (5.2) is called the *privacy loss* associated with releasing T . Larger ε intuitively corresponds to less privacy; smaller ε gives stronger privacy protection. A tenet of DP (in contrast with many other statistical disclosure risk frameworks) is that dependence of $M(x, U)$ on x implies non-negligible privacy loss $\varepsilon > 0$: Since $D_{\text{MULT}}(\mu, \nu) = 0$ if and only if $\mu = \nu$, complete privacy ($\varepsilon = 0$) is only possible by releasing pure noise – or, more exactly, by releasing $T \sim P_x$ where P_x is a function of x only through its *connected component* $[x] = \{x' \in \mathcal{X} \mid d(x, x') < \infty\}$ (see Definition 5.3.2 below). (This statement is formalised in Proposition 15.)

In the ideal case, the data custodian decides upon a maximum value of ε which is acceptable when considering the sensitivity of the data x and the privacy protection they deserve. The data custodian then designs a data-release mechanism which satisfies ε -DP, for this chosen value of ε . From this perspective, the maximum acceptable value of ε is called the *privacy loss budget*.

³In defining D_{MULT} we set $0/0 = \infty/\infty = 1$; and on the RHS of (5.2), we set $0 \times \infty = \infty$. On the space of probability measures, D_{MULT} is *strongly equivalent* (Definition C.3.2 in Appendix C.3) to the *density ratio metric* d_{DR} (Wasserman, 1992) (Definition 5.8.3 below). Namely,

$$D_{\text{MULT}}(P, Q) \leq d_{\text{DR}}(P, Q) \leq 2D_{\text{MULT}}(P, Q), \quad (5.3)$$

for *probability* measures P, Q on $(\mathcal{T}, \mathcal{F})$, so that ε -DP can be defined with d_{DR} in place of D_{MULT} , up to rescaling of ε . Equation (5.3) is proven in Proposition 13.

Two common choices of the metric d on \mathcal{X} are:

A) the Hamming distance

$$d_{\text{Ham}}(x, x') = \begin{cases} \sum_{i=1}^n 1_{x_i \neq x'_i} & \text{if } |x| = |x'| = n, \\ \infty & \text{otherwise,} \end{cases}$$

where the data $x = (x_1, x_2, \dots, x_n)$ are vectors and $|x|$ is the size of x ; and

B) the symmetric difference metric

$$d_{\Delta}(x, x') = |x \setminus x'| + |x' \setminus x|,$$

where the data $x, x' \in \mathcal{X}$ are multisets and $x \setminus x'$ is the (multi-)set difference.⁴

Equation (5.2) with d the Hamming distance is referred to as *bounded* DP and with the symmetric difference as *unbounded* DP.

The intuition behind differential privacy considers each record x_i in the data x as representing a single distinct individual. A distance $d(x, x') = 1$ then implies that x and x' differ according to the change in behaviour of a single individual – a change in the individual’s response, for the Hamming distance; or a change in whether the individual responds or not, for the symmetric difference metric. Specifically, ε -DP implies that a single individual can change the summary statistic $T = \mathcal{M}(x, U)$ by at most ε , where “change” is interpreted probabilistically in terms of the multiplicative distance.

Under the mild assumption that d is a graph distance with unit edges (Assumption 5.3.3, given below in Section 5.3), the converse implication also holds. That is, ε -DP is equivalent to the Lipschitz condition (5.2) holding when $d(x, x') = 1$. (This follows by the triangle inequality; for details see the proof of Theorem 5.3.5.) From herein, we restrict our attention to the set of such metrics, which includes d_{Ham} and d_{Δ} .

⁴We formally define a multiset S to be a non-negative-integer-valued function c_S , where $c_S(a)$ is the number of times the element a appears in S . The multiset difference $S \setminus S'$ is defined as the function $c_{S \setminus S'}(a) = \max\{0, c_S(a) - c_{S'}(a)\}$ and the multiset cardinality is defined as $|S| = \sum_a c_S(a)$.

Since DP controls the change in t due to perturbations in the data x , it can naturally be understood as a robustness property (Dwork and Lei, 2009; Avella-Medina, 2020, 2021; Asi et al., 2023; Hopkins et al., 2023). Measuring the “change in t ” by the multiplicative distance D_{MULT} – in place of more-familiar metrics typically seen in the robustness literature, such as the Kolmogorov distance or the total variation distance – is motivated by the strong notion of privacy as *indistinguishability*. The formulations of both pure ε -DP and Pufferfish as intervals of measure (which we describe in later sections) make this motivation clear; we therefore postpone further discussion of indistinguishability to Remarks 5.3.7 and 5.8.2.

This link to robustness hints at the connection between DP and IP. As we will soon see in Section 5.3, the multiplicative distance D_{MULT} is a distorting function Montes et al. (2020a), and consequently pure ε -DP can be characterised in terms of a distortion model, a point we expand on later in Remark 5.3.6. Indeed, D_{MULT} satisfies many of the common desiderata for distorting functions: it is positive definite, symmetric and quasi-convex, and it satisfies the triangle inequality. Additionally, $D_{\text{MULT}}(\mu, \nu)$ is continuous with respect to the supremum norm if and only if \mathcal{T} has finite cardinality and μ and ν have support \mathcal{T} . See Appendix C.4 for details on these desiderata.

Example 5.2.2 (Laplace mechanism (Dwork et al., 2006b)). Consider the problem of releasing a sanitised version of a deterministic summary statistic $q : \mathcal{X} \rightarrow \mathbb{R}^k$. (The terms ‘sanitised,’ ‘privatised,’ ‘privacy-protected,’ and ‘privacy-preserving’ are synonymous in the DP literature.) The Laplace mechanism adds noise with standard deviation proportional to the *global ℓ_1 -sensitivity* of q :

$$\Delta(q) = \sup_{\substack{x, x' \in \mathcal{X} \\ d(x, x')=1}} \|q(x) - q(x')\|_1. \quad (5.4)$$

Specifically, define $M(x, L) = q(x) + bL$, where $b = \frac{\Delta(q)}{\varepsilon}$ and L (the seed) is a k -vector of i.i.d. Laplace

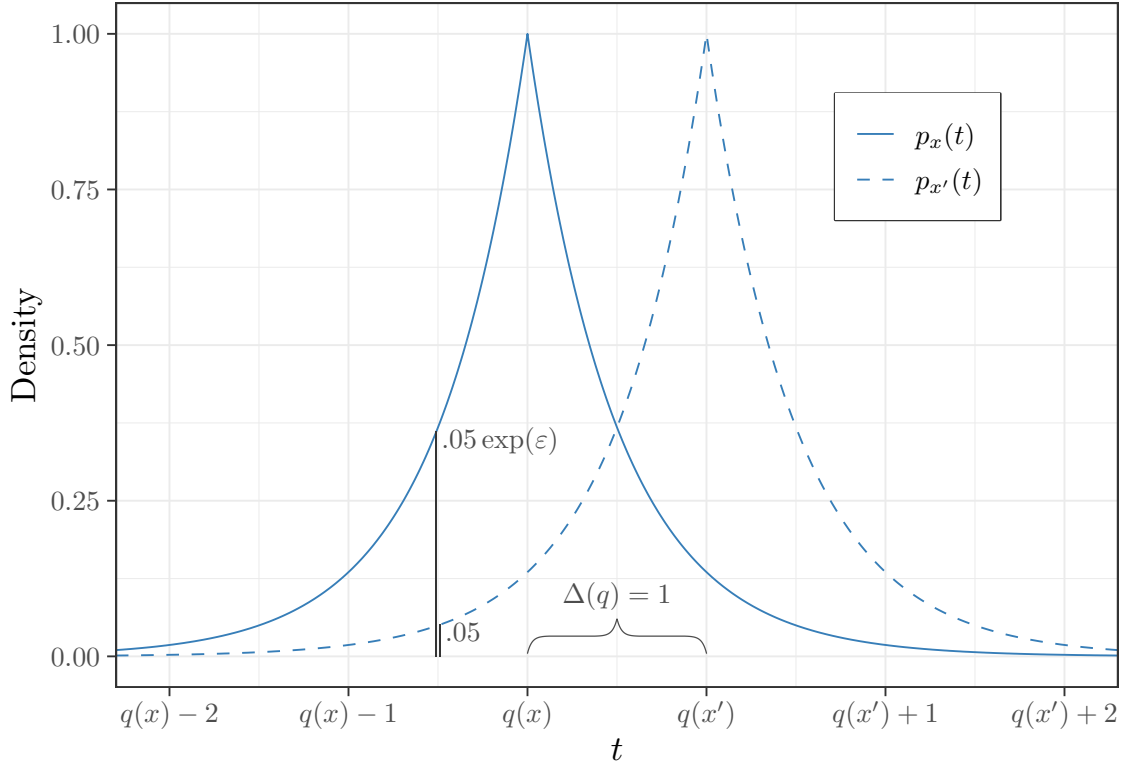


Figure 5.1: An illustration of the Laplace mechanism (Example 5.2.2). Here $p_x(t)$ and $p_{x'}(t)$ are probability densities for the Laplace mechanism's output T . The datasets x and x' are chosen so that $d(x, x') = 1$ and $q(x') - q(x) = \Delta(q) = 1$. The annotations 0.05 and $0.05 \exp(\varepsilon)$ are demonstrative of the property: $p_x(t) = e^\varepsilon p_{x'}(t)$ for all $t \leq q(x)$ and $p_x(t) = e^{-\varepsilon} p_{x'}(t)$ for all $t \geq q(x')$. (In this illustration, $\varepsilon = 2$ and $k = 1$.)

random variables with density $f(z) = 0.5 \exp(-|z|)$. (See Figure 5.1.) When $d(x, x') = 1$,

$$\begin{aligned}
 P_x(S_1 \times \dots \times S_k) &= \prod_{i=1}^k \left[0.5b^{-1} \int_{S_i} \exp\left(-\frac{|z - q_i(x)|}{b}\right) dz \right] \\
 &\geq \exp\left(\frac{-\Delta(q)}{b}\right) \prod_{i=1}^k \left[0.5b^{-1} \int_{S_i} \exp\left(-\frac{|z - q_i(x')|}{b}\right) dz \right] \\
 &= \exp(-\varepsilon) P_{x'}(S_1 \times \dots \times S_k).
 \end{aligned}$$

We will see in Theorem 5.3.5 that this suffices to prove that \mathcal{M} is ε -DP.

Example 5.2.3 (randomised response (Warner, 1965)). Taking $\mathcal{X} = \bigcup_{n \in \mathbb{N}} \{0, 1\}^n$ as the data universe,

the randomised response mechanism M flips each bit x_i with probability $p = (\exp \varepsilon + 1)^{-1}$. That is, given a binary n -vector x as input, M outputs another binary n -vector T with i -th component $T_i = x_i + B_i \bmod 2$ where $B_1, B_2, \dots \stackrel{iid}{\sim} \text{Bernoulli}(p)$. This mechanism is ε -DP when $d = d_{\text{Ham}}$.

Moreover, M conforms with the *local* model of DP (Kasiviswanathan et al., 2011; Duchi et al., 2018), since each data point can be independently infused with noise by the data respondents themselves. (For example, the i -th respondent can flip their own coin B_i and report their noisy answer T_i .) Local DP models are typically used when data are collected by an untrusted entity (such as an IT company), since these models require that the privacy protection is applied to each record before data collection. In the non-interactive setting, this requirement implies that $P_x(T \in \cdot)$ must factor as a product measure $\prod_{i=1}^n P_{x_i}(T_i \in \cdot)$, where $n = |x|$ is the number of records in x . (In contrast, the local *interactive* model of DP has the weaker condition that the distribution of user i 's response T_i cannot depend on x_j for $j \neq i$ (like in the non-interactive setting) but that this distribution can depend on the previous users' responses T_j for $j < i$.) The local privacy model contrasts with the *central* privacy model, under which the raw data x can be aggregated by a central, trusted authority (such as a national statistical office) before privacy protection is applied. (And hence, the probability P_x of a central privacy mechanism need not be factorizable.)

5.3 PURE ε -DIFFERENTIAL PRIVACY AS AN INTERVAL OF MEASURES

We introduce the definition of an *interval of measures* (IoM), due to DeRobertis and Hartigan (1981):

Definition 5.3.1. For measures μ and ν on the measurable space $(\mathcal{T}, \mathcal{F})$, write $\mu \leq \nu$ to denote that $\mu(S) \leq \nu(S)$ for all $S \in \mathcal{F}$.

Given measures L, U on $(\mathcal{T}, \mathcal{F})$ with $L \leq U$, the convex set of measures

$$\mathcal{I}(L, U) = \{\mu \text{ a measure on } (\mathcal{T}, \mathcal{F}) \mid L \leq \mu \leq U\}$$

is called an *interval of measures*. L and U are called the *lower* and *upper measures* of $\mathcal{I}(L, U)$.

Let Ω be the collection of all σ -finite measures on $(\mathcal{T}, \mathcal{F})$ and $\Omega_1 = \{P \in \Omega \mid P(\mathcal{T}) = 1\}$ be the collection of all probability measures on $(\mathcal{T}, \mathcal{F})$. In the vast majority of cases we encounter (and indeed all the practically meaningful ones), the upper measure U of an IoM $\mathcal{I}(L, U)$ is σ -finite and hence $\mathcal{I}(L, U) \subset \Omega$. The restriction $\mathcal{I}_1(L, U) = \mathcal{I}(L, U) \cap \Omega_1$ of an IoM $\mathcal{I}(L, U)$ to its probabilities forms a convex *credal set* [Levi \(1980\)](#).⁵ This set $\mathcal{I}_1(L, U)$ – which has previously been studied in the IP literature when $|\mathcal{T}|$ is finite under the name *probability interval* (PI) [de Campos et al. \(1994\)](#) – is the fundamental object of analysis throughout this paper.

As a direct consequence of Definition 5.3.1, the odds $P(A)/P(B)$ – for any $P \in \mathcal{I}(L, U)$ and any $A, B \in \mathcal{F}$ – are bounded between $L(A)/U(B)$ and $U(A)/L(B)$, whenever these ratios are well-defined. An IoM can also be expressed as a *density bounded class*, which is defined as follows: Fix some $\nu \in \Omega$ and pick ν -densities $l \leq u$. The density bounded class $\mathcal{I}(l, u)$ consists of ν -densities f satisfying $l \leq f \leq u$. (This is equivalent to Definition 5.3.1 when $U \in \Omega$ since every $\mu \in \mathcal{I}(L, U)$ is absolutely continuous with respect to U and so will always have a ν -density when $\nu = U$. See Proposition 14 in Appendix C.5 for more details.) Density bounded class, or the closely-related density ratio classes, are often used as prior neighbourhoods in robust Bayesian analysis due to their attractive properties; see e.g. [Berger \(1990\)](#); [Lavine \(1991a\)](#); [Wasserman \(1992\)](#); [Seidenfeld and Wasserman \(1993\)](#) and especially [Wasserman and Kadane \(1992\)](#). Moreover, IoMs have also been used to represent neighbourhoods of sampling distributions ([Lavine, 1991b](#)). When used in conjunction with prior neighbourhoods they augment Bayesian robustness beyond prior robustness without resorting to trivial posterior bounds. In fact, a neighbourhood of sampling distributions

⁵A credal set is simply a set of probabilities.

must have densities bounded away from zero and infinity – as is the case for a density bounded class $\mathcal{I}(l, u)$ (with $0 < l$ and $u < \infty$), but is not so for other popular neighbourhood models – to ensure that the resulting posterior neighbourhood have non-trivial extrema (Lavine, 1991b, Example 1), a point which is closely connected to the necessity of D_{MULT} for encoding the notion of “privacy as indistinguishability” (see Remark 5.3.7).

Here and elsewhere in this article, the term “density” is used in the broad sense of the Radon-Nikodym derivative $\frac{d\mu}{d\nu}$ of a measure $\mu \in \Omega$ with respect to a dominating measure $\nu \in \Omega$. Among other examples, this usage encompasses both probability density functions (PDFs) of continuous real-valued random variables (where the dominating measure is the Lebesgue measure) and probability mass functions (PMFs) of discrete random variables (where the dominating measure is the counting measure).

Theorem 5.3.5 establishes an equivalence between the ε -DP property of a data-release mechanism M and the interval of measures M induces.

Definition 5.3.2. Two datasets $x, x' \in \mathcal{X}$ are *connected* – or more precisely, *d*-connected – if $d(x, x') < \infty$. In this case, we say that x is a *connection* of x' , and that the probability measures P_x and $P_{x'}$ are *connected*. More generally, $S \subset \mathcal{X}$ is connected if all $x, x' \in S$ are.

The data universe \mathcal{X} is partitioned into *connected components* $[x] = \{x' \in \mathcal{X} \mid d(x, x') < \infty\}$. More generally, for $S \subset \mathcal{X}$, define

$$[S] = \{x \in \mathcal{X} \mid \exists x' \in S \text{ s.t. } d(x, x') < \infty\}.$$

Since the Lipschitz condition (5.2) is vacuous when $d(x, x') = \infty$, DP only constrains a mechanism M to act similarly on connected datasets x, x' ; it makes no (explicit) restrictions between outputs $M(x, U)$ and $M(x', U)$ for unconnected x, x' . That is, there is no privacy guarantee of indistinguishability between

unconnected datasets (although restrictions between outputs for connected x, x' may induce restrictions between outputs for unconnected x, x').

When $d = d_{\text{Ham}}$, any dataset x, x' of different dimension (i.e. x, x' such that $|x| \neq |x'|$) are unconnected. Hence, ε -DP with $d = d_{\text{Ham}}$ does not protect against, for example, an attacker determining $|x|$. Unconnected datasets also arise in the presence of *invariants* (Gong and Meng, 2020; Bailie et al., 2025b).

Assumption 5.3.3. $d(x, x')$ is equal to the length of a shortest path between x and x' in a graph on \mathcal{X} with unit-length edges.

When $d(x, x') > 1$, the Lipschitz condition (5.2) is called *group privacy*. This terminology comes from the following intuition: When each x_i represents an individual, condition (5.2) with $d_{\text{Ham}}(x, x') > 1$ (or $d_{\Delta}(x, x') > 1$) is protecting multiple individuals' privacy simultaneously. We prove in Theorem 5.3.5 that Assumption 5.3.3 and individual-only privacy (i.e. condition (5.2) for x, x' with $d(x, x') = 1$) together imply group privacy.

The following lemma is useful for Theorem 5.3.5 and subsequent discussions.

Lemma 5.3.4. *For any $\mu, \nu \in \Omega$ and $\varepsilon > 0$, we have $\nu \in \mathcal{I}(e^{-\varepsilon}\mu, e^{\varepsilon}\mu)$ if and only if*

$$D_{\text{MULT}}(\mu, \nu) \leq \varepsilon.$$

Hence, for any $\mu, \nu \in \Omega$ and $0 < a \leq 1 \leq b < \infty$,

1. $\nu \in \mathcal{I}(a\mu, b\mu)$ implies $D_{\text{MULT}}(\mu, \nu) \leq \max(-\ln a, \ln b)$; and
2. $D_{\text{MULT}}(\mu, \nu) \leq \min(-\ln a, \ln b)$ implies $\nu \in \mathcal{I}(a\mu, b\mu)$.

The proof of Lemma 5.3.4 is given in Appendix C.6, which also contains all other proofs which have been omitted from the main body of this paper.

Theorem 5.3.5. *Let $M : \mathcal{X} \times [0, 1] \rightarrow \mathcal{T}$ be a data-release mechanism with the seed $U \sim \text{Unif}[0, 1]$ inducing a probability P_x on $M(x, U)$ (where x is taken as fixed).*

For $0 \leq \varepsilon < \infty$, the following statements are equivalent given Assumption 5.3.3:

- I M is ε -differentially private.*
- II $P_{x'}(S) \leq e^\varepsilon P_x(S)$ for all $S \in \mathcal{F}$ and all $x, x' \in \mathcal{X}$ with $d(x, x') = 1$.*
- III For all $\delta \in \mathbb{N}$ and all $x, x' \in \mathcal{X}$ with $d(x, x') = \delta$,*

$$P_{x'} \in \mathcal{I}_1(L_{x, \delta\varepsilon}, U_{x, \delta\varepsilon}),$$

where $L_{x, \delta\varepsilon} = e^{-\delta\varepsilon} P_x$ and $U_{x, \delta\varepsilon} = e^{\delta\varepsilon} P_x$.

- IV For all $x \in \mathcal{X}$ and all measures $\nu \in \Omega$, if P_x has a density p_x with respect to ν , then every d -connected $x' \in [x]$ also has a density $p_{x'}$ (with respect to ν) satisfying*

$$p_{x'}(t) \in p_x(t) \exp[\pm \varepsilon d(x, x')], \quad (5.5)$$

for all $t \in \mathcal{T}$.

In (5.5), the notation $a \in \exp(\pm b)$ is shorthand for

$$\exp(-b) \leq a \leq \exp(b).$$

II is the standard definition of pure ε -DP (Dwork et al., 2006b) and is listed here to justify our novel formulation given in Definition 5.2.1. Without Assumption 5.3.3, group privacy is not implied by II. Hence Assumption 5.3.3 is needed only to extend II to provide group privacy; the equivalences between I, III and IV are automatic. Without Assumption 5.3.3 (which almost always holds in practice, such as for $d = d_{\text{Ham}}$ or d_{Δ}), our definition of ε -DP would be more stringent than the standard formulation.

Proof. (sketch) “I \Leftrightarrow II”: Since d is a graph distance, there is a path $x = x_0, \dots, x_n = x'$ such that $d(x, x') = n$ and $d(x_i, x_{i+1}) = 1$. By the triangle inequality,

$$D_{\text{MULT}}(P_x, P_{x'}) \leq \sum_{i=0}^{n-1} D_{\text{MULT}}(P_{x_i}, P_{x_{i+1}}).$$

Hence ε -DP is equivalent to the Lipschitz condition (5.2) holding only when $d(x, x') = 1$. The equivalence between I and II then follows by an application of Lemma 5.3.4: $e^{-\varepsilon} P_x(S) \leq P_{x'}(S) \leq e^{\varepsilon} P_x(S)$ for all $S \in \mathcal{F}$ if and only if $D_{\text{MULT}}(P_x, P_{x'}) \leq \varepsilon$. “I \Leftrightarrow III” is immediate by Lemma 5.3.4.

“III \Leftrightarrow IV”: The direction \Rightarrow is straightforward since the densities in an interval of measure $\mathcal{I}(L, U)$ are bounded by the densities of L and U . In the other direction \Leftarrow , P_x is always absolutely continuous with respect to itself, hence taking P_x to be the dominating measure ν , we have that (5.5) implies $P_{x'} \in \mathcal{I}_1(L_{x, \delta\varepsilon}, U_{x, \delta\varepsilon})$. \square

Remark 5.3.6. A *distorting function* d_{DIST} can be thought of as a generalised notion of distance between two probabilities or measures. (See Appendix C.4 for a precise definition.) Given a distorting function d_{DIST} and a *distortion parameter* $r > 0$, the *distortion model* on a probability P is the closed ball $B_{d_{\text{DIST}}}^r(P) = \{Q \in \Omega_1 \mid d_{\text{DIST}}(Q, P) \leq r\}$ centred at P with radius r [Montes et al. \(2020a\)](#).

Given a probability P , the symmetric probability interval $\mathcal{I}_1(e^{-\varepsilon}P, e^{\varepsilon}P)$ is a *distortion model* because, by Lemma 5.3.4, $\mathcal{I}(e^{-\varepsilon}P, e^{\varepsilon}P)$ is the closed ε -multiplicative-distance-ball:

$$\mathcal{I}(e^{-\varepsilon}P, e^{\varepsilon}P) = B_{D_{\text{MULT}}}^{\varepsilon}(P) = \{\mu \in \Omega \mid D_{\text{MULT}}(\mu, P) \leq \varepsilon\}.$$

Since the multiplicative distance D_{MULT} is used in defining the ball $B_{D_{\text{MULT}}}^{\varepsilon}(P)$, it serves the role of the *distorting function* of the distortion model $\mathcal{I}_1(e^{-\varepsilon}P, e^{\varepsilon}P)$; and the radius ε is the *distortion parameter*.

It is straightforward to verify that the credal set $\mathcal{I}_1(L, U)$ is convex. In Appendix C.4, we also prove that

it is closed with respect to the supremum norm, when restricting to the setting of [Montes et al. \(2020a\)](#).

That is, $\mathcal{I}_1(L, U) \cap \Omega_1^*$ is a closed subset of $\Omega_1^* = \{P \in \Omega_1 \mid P(\{t\}) > 0 \forall t \in \mathcal{T}\}$ (where closure is respect to the supremum norm), provided that the space \mathcal{T} has finite cardinality.

In general, a PI $\mathcal{I}_1(L, U)$ does not satisfy the definition of a distortion model. In fact, $\mathcal{I}_1(L, U)$ is a distortion model only when $\mathcal{I}(L, U)$ is symmetric in the sense that the lower and upper measures are equidistant from a central (‘nucleus’) probability P . This is the case for the PI $\mathcal{I}_1(e^{-\varepsilon}P_x, e^{\varepsilon}P_x)$ induced by an ε -DP mechanism \mathcal{M} .

As proven in Theorem 5.3.5, pure ε -DP is the requirement that P'_x lies in the neighbourhood $B_{D_{\text{MULT}}}^{\delta\varepsilon}(P_x)$, where $\delta = d(x, x')$. Many of the variants of ε -DP – such as (ε, δ) -DP ([Dwork et al., 2006a](#)), zero-concentrated DP ([Dwork and Rothblum, 2016](#); [Bun and Steinke, 2016](#)) and Rényi DP ([Mironov, 2017](#)) – replace the multiplicative distance D_{MULT} with another distorting function d_{DIST} . Consequently, these variants can also be characterised as distortion models: each of them is the requirement that $P_{x'}$ lies in the neighbourhood $B_{d_{\text{DIST}}}^{\delta\varepsilon}(P_x)$, for the appropriate choice of distorting function d_{DIST} . (See [Baillie et al. \(2025b\)](#) for the choices of d_{DIST} corresponding to (ε, δ) -DP, zero-concentrated DP and Rényi DP, and see [Desfontaines and Pejó \(2020\)](#) for a catalogue of variants of DP.)

Remark 5.3.7. IV of Theorem 5.3.5 is a strong property. It provides an quantification of the “indistinguishability” between data $x, x' \in \mathcal{X}$: if x, x' have densities $p_x, p_{x'}$ satisfying (5.5), then they are indistinguishable at the level ε . (Equation (5.5) is termed ε -indistinguishability in the literature, see e.g. [Dwork et al. \(2006b\)](#); [Dwork and Roth \(2014\)](#); [Vadhan \(2017\)](#)). More fundamentally, IV provides a categorical notion of indistinguishability: It implies that, for an ε -DP mechanism, all connected P_x are mutually absolutely continuous. Further, for all connected $x, x' \in \mathcal{X}$ and all $t \in \mathcal{T}$, either $p_x(t)$ and $p_{x'}(t)$ are both

zero or both non-zero. In intuitive terms, this means that if any x is plausible after observing $T = t$ (i.e. $p_x(t) > 0$) then all its connections $x' \in [x]$ are also plausible. This is a strong notion of privacy: regardless of the output $T = M(x, U)$, it's impossible for an attacker to distinguish between connected x, x' with certainty. In other words, the fiducial distribution for x is never degenerate (assuming that every x has at least one connection).

This notion of privacy is the motivation for D_{MULT} in place of more standard concepts in the robustness literature such as total variation distance or ε -contamination classes. Indeed, this categorical notion of indistinguishability requires that $p_x(t)/p_{x'}(t)$ is bounded away from zero and infinity, which is equivalent to $P_{x'} \in \mathcal{I}(aP_x, bP_x)$ for some $0 < a \leq 1 \leq b < \infty$. Yet Lemma 5.3.4 states that $P_{x'} \in \mathcal{I}(aP_x, bP_x)$ only if $D_{\text{MULT}}(P_x, P_{x'}) \leq \max(-\ln a, \ln b)$. Therefore, using the multiplicative distance D_{MULT} is necessary to encode the idea of privacy as indistinguishability between connected x, x' .

This argument demonstrates that the Lipschitz condition (5.2) with another distorting function d_{DIST} in place of D_{MULT} will not ensure indistinguishability (except in the trivial case where $\alpha d_{\text{DIST}} \geq D_{\text{MULT}}$ for some constant α). This is why common variants of pure ε -DP – such as (ε, δ) -DP (Dwork et al., 2006a), zero-concentrated DP (Dwork and Rothblum, 2016; Bun and Steinke, 2016) and Rényi DP (Mironov, 2017) (which, as described in Remark 5.3.6, all replace D_{MULT} with another distorting function d_{DIST}) – do not guarantee this strong notion of privacy, even though they may be preferred over pure ε -DP for data utility reasons.

The observations of Theorem 5.3.5, specifically the equivalent characterization of ε -DP via intervals of measures established by III and IV, bear important consequences for statistical inference from privacy-protected data. Notably, they impose meaningful bounds on both the probability of the privatised query

and on relevant quantities in the frequentist and Bayesian inference from the privatised queries. These bounds are valid under arbitrary statistical models for the unknown confidential database, assuming only mild conditions on the models' support. The next three sections explore these consequences in detail.

5.4 BOUNDS ON THE PRIVATISED DATA PROBABILITY

Consider the situation of statistical inference, where a data analyst supplies a parametric model $\mathcal{P} = \{P_\theta \mid \theta \in \Theta\}$ of data-generating distributions P_θ . Nature generates data $X \sim P_\theta$ according to some unknown $\theta \in \Theta$. (We use capital X to emphasise that the dataset is now random, whereas in the previous sections, it was considered fixed.) In the typical non-private setting, the data analyst observes X directly. In the private setting, the data analyst only sees the summary statistic $T = \mathcal{M}(X, U) \sim P_X$ outputted from a privacy-preserving data-release mechanism \mathcal{M} . (We now require that the data universe \mathcal{X} is equipped with a σ -algebra \mathcal{G} and that every data-release mechanism \mathcal{M} is $(\mathcal{G} \otimes \mathcal{B}[0, 1], \mathcal{F})$ -measurable, where $\mathcal{B}[0, 1]$ is the Borel σ -algebra on $[0, 1]$.)

The relevant vehicle for inference in the private setting is the marginal probability of the observed data T :

$$P(T \in S \mid \theta) = \int_{\mathcal{X}} P_x(S) dP_\theta(x). \quad (5.6)$$

We call $P(T \in S \mid \theta)$ the *privatised data probability*. (Proposition 17 proves that it is well-defined.) Viewed as a function of θ , $P(T \in S \mid \theta)$ is the *marginal likelihood* of θ . When the data observed by the analyst is privacy-protected, all frequentist procedures compliant with likelihood theory and all Bayesian inference hinge on this function [Berger and Wolpert \(1988\)](#). The crucial role of (5.6) for inference from privacy-protected data was first recognized in the differential privacy literature by [Williams and McSherry](#)

(2010), and has since been utilized extensively to derive likelihood and Bayesian methodologies (e.g. Awan and Slavković, 2018, 2020; Bernstein and Sheldon, 2018, 2019; Gong, 2022a; Ju et al., 2022).

When M is ε -DP and the support $\text{supp}(P_\theta)$ of P_θ is d -connected, the existence of a density $p(t \mid \theta)$ for $P(T \in S \mid \theta)$ is implied by Theorem 5.3.5. The following result proves this density always exists – as long as one restricts to a subspace of \mathcal{T} and assumes that (informally) the support of “ $P(x \mid t_0, \theta)$ ” is d -connected for some given $t_0 \in \mathcal{T}$. Other than this weak assumption, the following results hold for arbitrary data-generating models $\{P_\theta \mid \theta \in \Theta\}$ and ε -DP mechanisms M .

To state this assumption more precisely, define $\text{supp}(x \mid t, \theta)$ as the set of databases $x \in \mathcal{X}$ which could both generate t and be generated by P_θ . That is, $\text{supp}(x \mid t, \theta)$ is informally the intersection of $\text{supp}(P_\theta) \approx \{x \mid p_\theta(x) > 0\}$ and $\{x \mid p_x(t) > 0\} \approx \{x \mid t \in \text{supp}(P_x)\}$. See Appendix C.1 for an exact definition.

Theorem 5.4.1. *Let M be an ε -DP mechanism. Fix some $t_0 \in \mathcal{T}$ and suppose that $\text{supp}(x \mid t_0, \theta)$ is d -connected. Define $\mathcal{T}_0 = \{t \in \mathcal{T} \mid \text{supp}(x \mid t, \theta) \subset \text{supp}(x \mid t_0, \theta)\}$. Then, there exist measures $L_{\theta, \varepsilon}$ and $U_{\theta, \varepsilon}$ on $(\mathcal{T}, \mathcal{F})$ with densities $l_{\theta, \varepsilon}$ and $u_{\theta, \varepsilon}$ satisfying*

$$l_{\theta, \varepsilon}(t) = \text{ess sup}_{x_* \in \text{supp}(x \mid t_0, \theta)} \exp(-\varepsilon d_*) p_{x_*}(t), \quad \text{and} \quad u_{\theta, \varepsilon}(t) = \text{ess inf}_{x_* \in \text{supp}(x \mid t_0, \theta)} \exp(\varepsilon d_*) p_{x_*}(t),$$

for all $t \in \mathcal{T}_0$, where $d_* = \sup_{x \in \text{supp}(x \mid t_0, \theta)} d(x, x_*)$.

Furthermore, the privatised data probability $P(T \in \cdot \mid \theta)$ is bounded by $L_{\theta, \varepsilon}$ and $U_{\theta, \varepsilon}$ on \mathcal{T}_0 :

$$P(T \in \cdot \cap \mathcal{T}_0 \mid \theta) \in \mathcal{I}(L_{\theta, \varepsilon}, U_{\theta, \varepsilon}). \quad (5.7)$$

Proof. (sketch) The existence of a density $p(t \mid \theta)$ for $P(T \in \cdot \cap \mathcal{T}_0 \mid \theta)$ follows from the fact that all P_x with $x \in \text{supp}(x \mid t_0, \theta)$ are mutually absolutely continuous by Theorem 5.3.5. For the upper bound

of (5.7), first observe that

$$\begin{aligned}
p(t \mid \theta) &= \int_{\text{supp}(x \mid t_0, \theta)} p_x(t) dP_\theta(x) \\
&\leq \int_{\text{supp}(x \mid t_0, \theta)} e^{\varepsilon d(x, x_*)} p_{x_*}(t) dP_\theta(x) \\
&\leq e^{\varepsilon d_*} p_{x_*}(t).
\end{aligned}$$

Since the above inequalities hold for all $x_* \in \text{supp}(x \mid t_0, \theta)$, we can take the essential infimum over x_* to obtain the bound $p(t \mid \theta) \leq u_{\theta, \varepsilon}(t)$. The lower bound of (5.7) follows similarly. \square

It becomes apparent in the proof of Theorem 5.4.1 that this result can be generalised in the following way: In defining $\mathcal{T}_0 = \{t \in \mathcal{T} \mid \text{supp}(x \mid t, \theta) \subset \text{supp}(x \mid t_0, \theta)\}$, one may replace $\text{supp}(x \mid t_0, \theta)$ with any measurable S satisfying

$$\text{supp}(x \mid t_0, \theta) \subset S \subset [\text{supp}(x \mid t_0, \theta)].$$

(The notation $[\cdot]$ is defined in Definition 5.3.2.) Theorem 5.4.1 holds with this new \mathcal{T}_0 , provided that $\text{supp}(x \mid t_0, \theta)$ is replaced by S in the definitions of $l_{\theta, \varepsilon}$, $u_{\theta, \varepsilon}$ and d_* . This demonstrates that the density $p(t \mid \theta)$ exists on a larger \mathcal{T}_0 , although the resulting bounds $l_{\theta, \varepsilon}$ and $u_{\theta, \varepsilon}$ on $p(t \mid \theta)$ may be wider.

Theorem 5.4.1 shows that the privatised data probability $P(T \in \cdot \mid \theta)$ is in a probability interval, and that this probability interval is bounded by $L_{\theta, \varepsilon}$ and $U_{\theta, \varepsilon}$ on \mathcal{T}_0 . Broadly speaking, this theorem has two uses. Firstly, t_0 may be taken to be the realised value of T . Then Theorem 5.4.1 can be interpreted as bounds on the marginal likelihood $l(\theta \mid t_0)$ of θ . (For this application, one must make the additional assumption that $\bigcup_{\theta \in \Theta} \text{supp}(x \mid t_0, \theta)$ is d -connected, so that the densities $p(t_0 \mid \theta)$ of the privatised data probability, across the different values of θ , share a common dominating measure. This ensures that the likelihood $l(\theta \mid t_0) = p(t_0 \mid \theta)$ is well-defined as a function of θ .) Secondly, one may be interested in

understanding the privatised data probability within some subspace $S \subset \mathcal{T}$. Although it may not always be possible, if one can find some $t_0 \in \mathcal{T}$ such that $S \subset \mathcal{T}_0$, then Theorem 5.4.1 provides information on what the privatised data probability looks like within the subspace of interest S .

Surprisingly, the interval of measures $\mathcal{I}(L_{\theta,\varepsilon}, U_{\theta,\varepsilon})$ in (5.7) depends on the data-generating distribution P_θ only through $\text{supp}(x \mid t_0, \theta)$. When $\text{supp}(P_\theta)$ is constant, $\mathcal{I}(L_{\theta,\varepsilon}, U_{\theta,\varepsilon})$ is completely free of θ . Alternatively, one may take the essential-infimum of $L_{\theta,\varepsilon}$ over $\theta \in \Theta$ to obtain a bound on $P(T \in \cdot \cap \mathcal{T}_0 \mid \theta)$ which is completely free of θ , although it is likely such a bound will be vacuous.

Theorem 5.4.1 is only practically meaningful when $d_* < \infty$. Typically $\sup_{x,x' \in \mathcal{X}} d(x, x') = \infty$, which one might presume would imply that $d_* = \infty$. But $\text{supp}(P_\theta)$ can be much smaller than the data universe \mathcal{X} when the analyst has prior knowledge of the data X . The analyst is free to restrict $\text{supp}(P_\theta)$ to the set of datasets they deem plausible; the tighter this restriction, the stronger Theorem 5.4.1 is. For example, the analyst may have an upper bound b on the number of records $|X|$. (Provided that $\text{supp}(x \mid t_0, \theta)$ is connected – a weak assumption, as explained in Remark 5.4.2 – this would imply $d_* \leq b$ for typical choices of d .) Moreover, $\text{supp}(x \mid t_0, \theta)$ can be much smaller than \mathcal{X} when t_0 restricts the possible values of X , such as in the presence of invariants (Gong and Meng, 2020; Bailie et al., 2025b). For example in local DP (or bounded DP more generally), the number of records $|t|$ is invariant; this restricts $\text{supp}(x \mid t_0, \theta)$ to data x satisfying $|x| = |t_0|$, which would typically imply $d_* \leq |t_0|$.

Remark 5.4.2. Theorem 5.4.1 only relies on a single assumption which concerns the connectedness of $\text{supp}(x \mid t_0, \theta)$. This assumption is weak. In fact, we can always augment the data-release mechanism \mathcal{M} so that this assumption is satisfied without increasing \mathcal{M} 's privacy loss ε . Specifically, the (deterministic) mechanism $x \mapsto [x]$ (which publishes the connected component $[x]$ of the observed data x) is trivially ε -DP

with $\varepsilon = 0$. Publishing $[x]$ alongside $\mathcal{M}(x, U)$ ensures that $\text{supp}(x \mid t, \theta)$ is always connected, for all t and θ . (This argument is formalised in Proposition 16.)

We illustrate Theorem 5.4.1 with two examples.

Example 5.4.3 (privatised binary sum). Suppose the database $x \in \mathcal{X} = \bigcup_{n=1}^{\infty} \{0, 1\}^n$ consists of n records of binary features, and its sum $q(x) = \sum_{i=1}^n x_i$ is to be queried. Consider sanitising $q(x)$ using the Laplace mechanism defined in Example 5.2.2. For every privacy loss $\varepsilon > 0$ and every database x ,

$$p_x(t) = \frac{\varepsilon}{2\Delta(q)} \exp\left(\frac{\varepsilon|t - q(x)|}{\Delta(q)}\right),$$

where, in this case, the global ℓ_1 -sensitivity $\Delta(q)$ (defined in (5.4)) is one.

The data analyst posits an arbitrary statistical model $X \sim P_\theta$ for $\theta \in \Theta$ with $\text{supp}(P_\theta) \subset \{x \in \mathcal{X} \mid |x| \leq 10\}$, and considers the confidential and unknown database x to be a realization from this model. Since $\text{supp}(P_x) = \mathbb{R}$ for all $x \in \mathcal{X}$, the assumption of Theorem 5.4.1 simplifies to the requirement that $\text{supp}(P_\theta)$ is d -connected. Moreover, $\mathcal{T}_0 = \mathcal{T} = \mathbb{R}$. (Both of these points hold regardless of the choice of t_0 .)

Figure 5.2 displays the lower and upper densities, $l_\varepsilon = \text{ess inf}_{\theta \in \Theta} l_{\theta, \varepsilon}$ and $u_\varepsilon = \text{ess sup}_{\theta \in \Theta} u_{\theta, \varepsilon}$, for the privatised data probability $p(t \mid \theta)$. The analyst upper bounds the number of records $|x|$ by 10, so that $d_* = 10$. The left and right panels display bounds under two different settings of ε . The bounds are tighter and more informative when privacy protection is more stringent ($\varepsilon = 0.1$), and looser as the privacy loss increases ($\varepsilon = 0.25$). Notice that these bounds for $p(t \mid \theta)$ are functions of the value of the privatised query t . In particular, they do not depend on θ nor the form of the posited data model P_θ .

Example 5.4.4 (local ε -DP). Suppose the distribution P_x of the published summary statistic T factors as

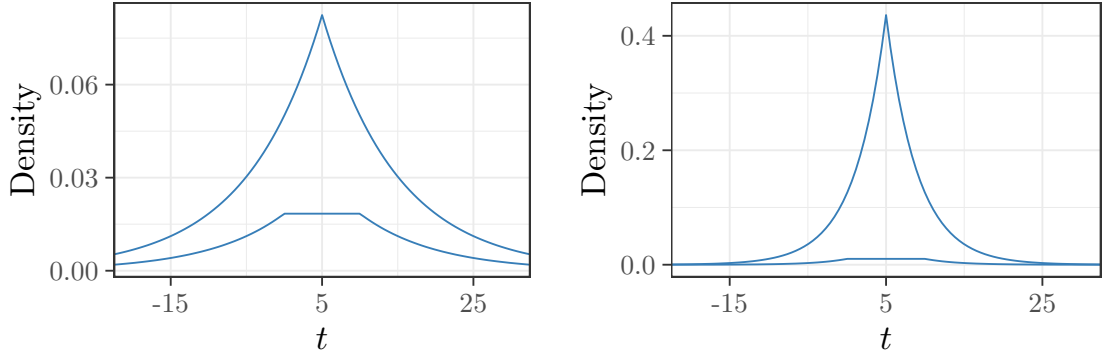


Figure 5.2: Upper and lower bounds for the density $p(t \mid \theta)$ of the privatised binary sum (Example 5.4.3). The privacy loss is $\varepsilon = 0.1$ (left) and $\varepsilon = 0.25$ (right). These bounds depend on the assumed data model P_θ only through the analyst's belief that the number of records $|x|$ is bounded by ten, which means that $\text{supp}(P_\theta) \subset \{x \in \mathcal{X} \mid |x| \leq 10\}$. They are tighter and more informative when the privacy protection is more stringent (i.e. when ε is smaller).

$\prod_{i=1}^n P_{x_i}$, where $n = |x|$. (This always holds under the local, non-interactive model of DP, as we described in Example 5.2.3.) Then $|T| = |X|$ and hence $\text{supp}(P_x) \subset \{t \in \mathcal{T} \mid |t| = |x|\}$.

Most local DP mechanisms satisfy the stricter assumption that $\text{supp}(P_x) = \{t \in \mathcal{T} \mid |t| = |x|\}$. Under this assumption, $\mathcal{T}_0 = \{t \in \mathcal{T} \mid |t| = |t_0|\}$ and, if $d = d_{\text{Ham}}$ (as is typical for local DP), then $d_* \leq |t_0|$ regardless of the choice of x_* . Hence, by Lemma C.6.3 of Appendix C.6 (which is used in proving Theorem 5.4.1), the density of an ε -DP mechanism is bounded by

$$p(t \mid \theta) \in \prod_{i=1}^n p_{x_i}(t_i) \exp(\pm \varepsilon n),$$

for any x and any t with $|x| = |t| = n$. Applying this result to the randomised response mechanism (Example 5.2.3), $\min_{x_i} p_{x_i}(t_i) = (\exp \varepsilon + 1)^{-1}$ and $\max_{x_i} p_{x_i}(t_i) = e^\varepsilon (\exp \varepsilon + 1)^{-1}$, so that

$$\frac{1}{(\exp \varepsilon + 1)^{|t|}} \leq p(t \mid \theta) \leq \frac{\exp(|t|\varepsilon)}{(\exp \varepsilon + 1)^{|t|}}. \quad (5.8)$$

The bounds in (5.8) depend on t only through $|t|$ (the number of records), regardless of the records' values.

Figure 5.3 displays these bounds as a function of $|t|$ for $\varepsilon = 1$. As more records are released (larger $|t|$), both

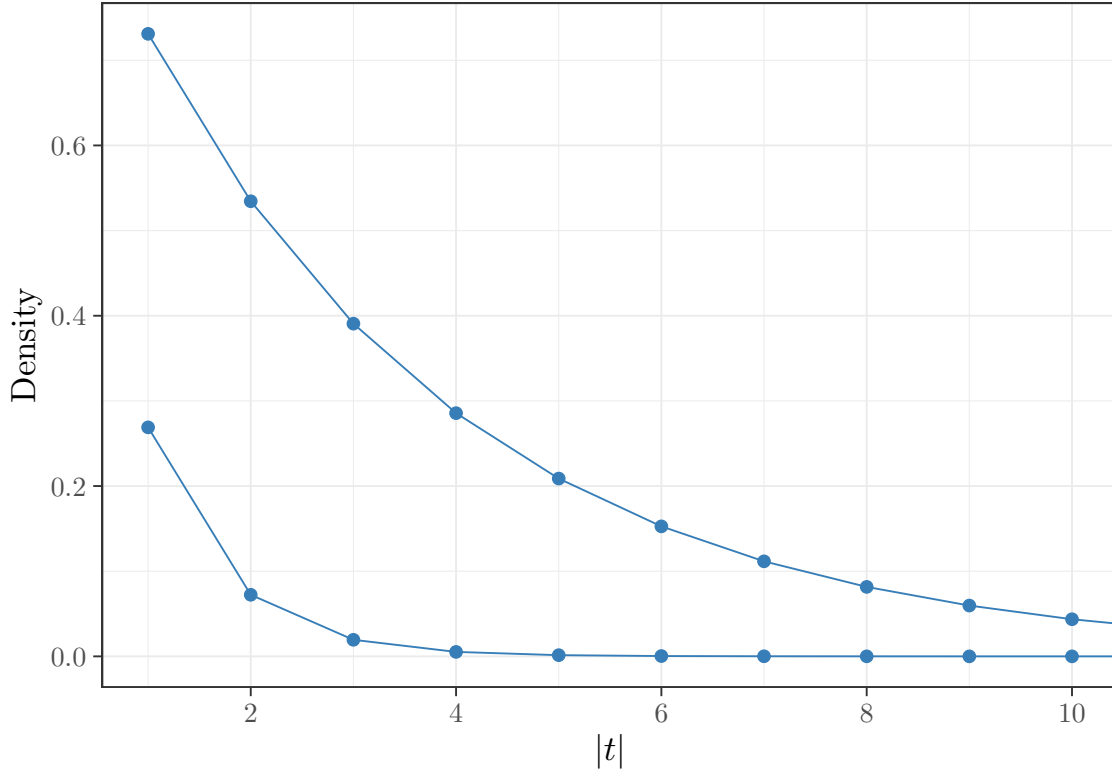


Figure 5.3: Upper and lower density bounds for $p(t \mid \theta)$ under randomised response (Example 5.4.4). The privacy loss is $\varepsilon = 1$. These bounds are a function of t only through $|t|$ (the number of observed records).

bounds tend to zero with a narrowing gap.

5.5 FREQUENTIST PRIVACY-PROTECTED INFERENCE

The interval of measures formulation of ε -DP also shows that Neyman-Pearson hypothesis testing is restricted in the private setting, as demonstrated by the following theorem.

Theorem 5.5.1. *Consider testing $H_0 : \theta = \theta_0$ versus $H_1 : \theta = \theta_1$ for some $\theta_0 \neq \theta_1 \in \Theta$. Let $S_i = \text{supp}(P_{\theta_i})$ and suppose that $S_0 \cup S_1$ is d -connected. In the private setting where the observed data T is the*

output of an ε -DP mechanism, the power of any level- α test is bounded above by $\alpha \exp(d_{**}\varepsilon)$ where

$$d_{**} = \sup_{x \in \mathcal{S}_0, x' \in \mathcal{S}_1} d(x, x').$$

Proof. (sketch) By IV of Theorem 5.3.5,

$$\begin{aligned} \frac{p(t \mid \theta_1)}{p(t \mid \theta_0)} &= \frac{\int_{\mathcal{S}_1} p_x(t) dP_{\theta_1}(x)}{\int_{\mathcal{S}_0} p_{x'}(t) dP_{\theta_0}(x')} \\ &= \int_{\mathcal{S}_1} \left[\int_{\mathcal{S}_0} \frac{p_{x'}(t)}{p_x(t)} dP_{\theta_0}(x') \right]^{-1} dP_{\theta_1}(x) \\ &\in \exp(\pm \varepsilon d_{**}). \end{aligned}$$

Let R be the rejection region of a test with size $P(T \in R \mid \theta_0) \leq \alpha$ and let ν be the dominating measure of the densities $p(t \mid \theta_0)$ and $p(t \mid \theta_1)$. Then

$$\begin{aligned} P(T \in R \mid \theta_1) &= \int_R p(t \mid \theta_1) d\nu(t) \\ &\leq \exp(d_{**}\varepsilon) \int_R p(t \mid \theta_0) d\nu(t) \\ &\leq \alpha \exp(d_{**}\varepsilon). \end{aligned} \tag{5.9}$$

□

Compare Theorem 5.5.1 to the hypothesis test $H_0 : x_{1:m} = y$ versus $H_1 : x_{1:m} = y'$ where $m \leq |x|$. (Here we assume that the datasets $x \in \mathcal{X}$ are vectors of length $|x|$ and consist of records x_1, x_2, \dots, x_n , where $n = |x|$. For $1 \leq i \leq j \leq |x|$, the notation $x_{i:j}$ denotes the sub-vector $(x_i, x_{i+1}, \dots, x_j)$, consisting of the i -th through j -th records of x .) This test models an attacker trying to distinguish the first m records of the database. Wasserman and Zhou (2010) showed that any level- α test of $x_{1:m}$ has power at most $\alpha \exp(\varepsilon m)$ when the records X_i are i.i.d.

If the data analyst restricts S_0 and S_1 to datasets of length m , then typically $d_{**} = m$. Thus, any level- α test on the parameter θ has the same bound $\alpha \exp(\varepsilon m)$ on its power (under an arbitrary data-generating model, not just i.i.d. X_i).

Theorem 5.5.1 strictly generalises the result of Wasserman and Zhou (2010). By taking $\Theta \subset \mathcal{X}$ and setting P_θ as degenerate point masses, we recover the set-up of an attacker’s hypothesis test.⁶ Thus, Theorem 5.5.1 is applicable to both the attacker testing x (as in Wasserman and Zhou (2010)) and the analyst testing θ (with non-degenerate P_θ). This highlights the fundamental tension between data privacy and data utility: bounding an attacker’s power will bound the power of a legitimate analyst. However, another look at Theorem 5.5.1 seems to suggest a possible way to partially resolve this tension under certain circumstances. The data custodian might have the liberty to choose a metric d on \mathcal{X} that ensures the connectedness assumption of Theorem 5.5.1 holds for the hypothesis tests of the typical attacker but not for those of the legitimate analyst. If this happens, the hypothesis test of the attacker – but not of the legitimate analyst – will be constrained by Theorem 5.5.1. Such a choice for d would therefore resolve the tension between privacy and utility as it appears in Theorem 5.5.1. (This is not to suggest that the analyst will be totally unaffected by such privacy protection – any noise infusion can in general decrease the power of their test – but, at least, they will not be affected to the extent suggested by Theorem 5.5.1.)

⁶This ignores one minor technicality: the attacker may take some records as nuisance parameters, which they do not want to test. It is straightforward to generalise Theorem 5.5.1 to this situation. Without loss of generality, suppose $x_{m+1:n}$ are nuisance parameters when testing $x_{1:m}$ against $x'_{1:m}$. By assigning a conditional probability on $x_{m+1:n}$ satisfying $\pi(x_{m+1:n} \mid x_{1:m}) = \pi(x_{m+1:n} \mid x'_{1:m})$, the nuisance parameters can be integrated out in (5.9). This gives the same power bound $\alpha \exp(d_{**}\varepsilon)$, except now with

$$d_{**} = \sup_{x_{m+1:n}} d\left([x_{1:m}, x_{m+1:n}], [x'_{1:m}, x_{m+1:n}]\right),$$

which is typically equal to m , as before.

Corollary 5.5.2. *Under the set-up of Theorem 5.5.1, the power $1 - \beta$ of any size- α test is bounded by the inequalities:*

$$\max(\alpha e^{-d_{**}\varepsilon}, 1 - e^{d_{**}\varepsilon}[1 - \alpha]) \leq 1 - \beta \leq \min(\alpha e^{d_{**}\varepsilon}, 1 - e^{-d_{**}\varepsilon}[1 - \alpha]). \quad (5.10)$$

Proof. (Kifer et al., 2022, Section 6.1) Let R be the rejection region of a test with size $\alpha = P(T \in R \mid \theta_0)$ and power $1 - \beta = P(T \in R \mid \theta_1)$. In the proof of Theorem 5.5.1, we showed that

$$P(T \in R \mid \theta_1) \leq \exp(d_{**}\varepsilon)P(T \in R \mid \theta_0).$$

In the same way, one can show that

$$P(T \in R \mid \theta_1) \geq \exp(-d_{**}\varepsilon)P(T \in R \mid \theta_0),$$

and that

$$\exp(-d_{**}\varepsilon)P(T \notin R \mid \theta_0) \leq P(T \notin R \mid \theta_1) \leq \exp(d_{**}\varepsilon)P(T \notin R \mid \theta_0).$$

Combining these four inequalities gives (5.10). □

5.6 BAYESIAN PRIVACY-PROTECTED INFERENCE

Following the set-up from the previous two sections, we further assume that the analyst is Bayesian and places a (proper) prior π on Θ . This setting can be seen as a Bayesian hierarchical model where the raw, confidential data X acts as latent parameter in the Markov chain $\theta \rightarrow X \rightarrow T$.

We make the following assumption throughout this section.

Assumption 5.6.1. Define

$$\text{supp}(x \mid t) := \bigcup_{\theta \in \text{supp}(\pi)} \text{supp}(x \mid t, \theta).$$

Fix some $t_0 \in \mathcal{T}$. Suppose that (A) $\text{supp}(x \mid t_0)$ is d -connected. Further, assume that (B) the prior π on θ is proper.

By the same reasoning as in Remark 5.4.2, Assumption 5.6.1(A) is weak because it can always be satisfied by augmenting the data-release mechanism \mathcal{M} without additional privacy loss.

Theorem 5.6.2 establishes bounds on the analyst's prior predictive distribution $P(T \in S) = \iint P_x(S) dP_\theta(x) d\pi(\theta)$ for the privatised data T .

Theorem 5.6.2. *Let \mathcal{M} be an ε -DP mechanism. Define $\mathcal{T}_0 = \{t \in \mathcal{T} \mid \text{supp}(x \mid t) \subset \text{supp}(x \mid t_0)\}$. Then, there exist measures L_ε and U_ε on $(\mathcal{T}, \mathcal{F})$ with densities l_ε and u_ε satisfying*

$$l_\varepsilon(t) = \text{ess sup}_{x_* \in \text{supp}(x \mid t_0)} \exp(-\varepsilon d_*) p_{x_*}(t) \quad \text{and} \quad u_\varepsilon(t) = \text{ess inf}_{x_* \in \text{supp}(x \mid t_0)} \exp(\varepsilon d_*) p_{x_*}(t),$$

for all $t \in \mathcal{T}_0$, where $d_* = \sup_{x \in \text{supp}(x \mid t_0)} d(x, x_*)$.

Furthermore, the Bayesian analyst's prior predictive probability $P(T \in \cdot)$ is bounded by L_ε and U_ε on \mathcal{T}_0 :

$$P(T \in \cdot \cap \mathcal{T}_0) \in \mathcal{I}(L_\varepsilon, U_\varepsilon). \quad (5.11)$$

Proof. (sketch) Since $p(t) = \int_{\Theta} p(t \mid \theta) d\pi(\theta)$, Theorem 5.6.2 follows by showing $p(t \mid \theta)$ is bounded by $l_\varepsilon(t)$ and $u_\varepsilon(t)$ for almost all $t \in \mathcal{T}_0$. The proof of this is analogous to (5.7). \square

As for Theorem 5.4.1, one can replace $\text{supp}(x \mid t_0)$ in the definition of \mathcal{T}_0 and in the statement of Theorem 5.6.2 with any measurable $S \subset \mathcal{X}$ which satisfies

$$\text{supp}(x \mid t_0) \subset S \subset [\text{supp}(x \mid t_0)].$$

In this way, one can obtain bounds $l_\varepsilon(t)$ and $u_\varepsilon(t)$ on the prior predictive density $p(t)$ which apply for a larger subspace \mathcal{T}_0 , although these bounds will be wider.

The prior predictive distribution $p(t)$ plays an important role in Bayesian inference and model checking. Before observing the data, $p(t)$ captures the analyst's implied specification on the data-generation process. After observing the data, this quantity assessed at their value is called *model evidence* where low $p(t)$ reveals potential *conflict* between the data and the prior [Evans and Moshonov \(2006\)](#); [Walter and Augustin \(2009\)](#). In addition, it is also the normalizing constant for the posterior distribution $\pi(\theta \mid t)$ and hence is useful for computation.

Theorem 5.6.2 shows that the prior predictive distribution $P(T \in \cdot)$ is in a probability interval, and this probability interval is bounded by L_ε and U_ε on \mathcal{T}_0 .

As an illustration, we can see from Figure 5.2 of Example 5.4.3 that when $\varepsilon = 0.1$, the prior predictive probability of the privatised query is lower-bounded at ≈ 0.02 whenever $0 \leq t \leq 10$, and can never exceed ≈ 0.08 even when $t = 5$. On the other hand, when privacy protection is less stringent ($\varepsilon = 0.5$), the upper bound on the prior predictive probability increases to more than 0.4.

An important observation on Theorem 5.6.2 is the following: While $p(t)$ is a function of both the data model P_θ and the prior π , the density bounds $l_\varepsilon(t)$ and $u_\varepsilon(t)$ are free of both. In this sense, these bounds provide a non-trivial yet almost assumption-free prior predictive model sensitivity analysis. Non-trivial bounds on $p(t)$ are not possible in general; in this case they are a consequence of the data T being ε -DP.

Theorem 5.6.3 provides general bounds limiting the learning of a Bayesian analyst.

Theorem 5.6.3. *Suppose that an ε -DP mechanism M outputs the realisation t_0 . The analyst's posterior probability given t_0 satisfies*

$$\pi(\theta \in S \mid t_0) \in \pi(\theta \in S) \exp(\pm \varepsilon d_{**}), \quad (5.12)$$

for all $S \in \mathcal{F}$, where $d_{**} = \sup_{x, x' \in \text{supp}(x \mid t_0)} d(x, x')$.

Proof. (sketch) As in the proof of Theorem 5.5.1, we can show that

$$\frac{p(t_0 | \theta)}{p(t_0 | \theta')} \in \exp(\pm \varepsilon d_{**}),$$

for all $\theta, \theta' \in \text{supp}(\pi)$. Plugging this into $\pi(\theta | t_0) = \frac{p(t_0 | \theta) \pi(\theta)}{\int_{\Theta} p(t_0 | \theta') d\pi(\theta')}$ gives the result. \square

Theorem 5.6.3 demonstrates that the posterior $\pi(\theta \in \cdot | t_0)$ is in a probability interval which is centred at the prior $\pi(\theta \in \cdot)$ and has radius $\exp(\varepsilon d_{**})$:

$$\pi(\theta \in \cdot | t_0) \in \mathcal{I}_1(L, U),$$

where $L = \pi(\theta \in \cdot) \exp(-\varepsilon d_{**})$ and $U = \pi(\theta \in \cdot) \exp(\varepsilon d_{**})$.

Remark 5.6.4. By following the proof of Theorem 5.6.3, one can observe that $D_{\text{MULT}}(P_x, P_{x'})$ being bounded away from infinity, for all $x, x' \in \text{supp}(x | t_0)$, is a necessary condition for

$$D_{\text{MULT}}[\pi(\theta | t_0), \pi(\theta)] < \infty.$$

(Note that (5.12) is equivalent to $D_{\text{MULT}}[\pi(\theta | t_0), \pi(\theta)] \leq \varepsilon d_{**}$.) Indeed this condition is required for the posterior to be in a non-vacuous probability interval centred at the prior – i.e. for the posterior to be in an probability interval of the form $\pi(\cdot | t_0) \in \mathcal{I}_1(a\pi, b\pi)$ where $0 < a \leq 1 \leq b < \infty$. Hence the use of D_{MULT} in the Lipschitz condition (5.2) is the unique choice (modulo distorting functions d_{DIST} satisfying $\alpha d_{\text{DIST}} \geq D_{\text{MULT}}$ for some constant α) that ensures a bound on the prior-to-posterior of the form (5.12). This is analogous to the fact that D_{MULT} is the unique choice of distorting function that encodes privacy as indistinguishability (see Remark 5.3.7). Furthermore, by similar logic D_{MULT} is also the unique choice of distorting function which enables bounds on hypothesis testing like those in Theorem 5.5.1.

These uniqueness properties are mirrored in the results of Wasserman (1992) and Lavine (1991b) on

the uniqueness of intervals of measures in robust Bayesian inference.

Theorem 5.6.3 contributes to what is called the *prior-to-posterior semantics* of differential privacy (see [Kasiviswanathan and Smith \(2014\)](#); [Dwork et al. \(2006b\)](#); [Duncan and Lambert \(1986\)](#)), in the sense that (5.12) describes the extent to which a Bayesian agent’s posterior about a parameter θ can depart from their prior when learning from an ε -DP data product.⁷ Analogous to the discussion on frequentist attackers at the end of Section 5.5, Theorem 5.6.3 demonstrates the trade-off between restricting a Bayesian attacker while allowing for legitimate Bayesian learning: By setting $\Theta \subset \mathcal{X}$ and P_θ as degenerate point masses, we strictly generalise the result of [Gong and Meng \(2020\)](#) which bounds an attacker’s prior-to-posterior change in a single record x_i .⁸ Hence, we see that Theorem 5.6.3 applies to both the legitimate analyst who is inferring population-level characteristics and the illegitimate attacker who is inferring individual-level information. Restricting the attacker (by decreasing ε) necessarily hurts the analyst; whilst furnishing the analyst (by increasing ε) also assists the attacker. What makes this dilemma tractable is that d_{**} is typically much larger for the analyst than for the attacker because the analyst is interested in population quantities while the attacker is interested in individual records. Hence, Theorem 5.6.3’s bounds

⁷An alternative type of semantics for differential privacy is the *posterior-to-posterior semantics* [Dinur and Nissim \(2003\)](#); [Kasiviswanathan and Smith \(2014\)](#), whose focus is on the extent to which a Bayesian agent’s posterior may vary were it derived from privacy-protected queries based on different (counterfactual) confidential databases. Previous literature in differential privacy predominantly adopted posterior-to-posterior semantics; see e.g. [Kifer et al. \(2022\)](#). However, prior-to-posterior semantics have recently attracted increasing attention as they circumvent counterfactuals and are closely connected with the literature on statistical disclosure risk; see e.g. [Gong and Meng \(2020\)](#); [Hotz et al. \(2022\)](#).

⁸By fixing some $x \in \mathcal{X}$ and setting $\pi(X_{-i} = x_{-i}) = 1$, we get $d_{**} = 1$ and thereby rederive the result from [Gong and Meng \(2020\)](#). (Here x_{-i} denotes the dataset x – which is assumed to be a vector – with the i -th record removed.)

Alternatively, one could set $\pi(\theta) = \pi(x_i \mid x_{-i})$, in which case Theorem 5.6.3 implies that

$$\pi(x_i \mid t, x_{-i}) \in \pi(x_i \mid x_{-i}) \exp(\pm \varepsilon d_{**}),$$

(with $d_{**} = 1$ when, for example, $d = d_{\text{Ham}}$).

on the inference of the analyst are wider than those of the attacker. However, this argument breaks down when the analyst is interested in small subpopulations (such as in small-area estimation) because in these situations there is little light between the attacker's and the analyst's interests, and as such the values of d_{**} associated with the attacker and the analyst will be similar. (This commentary – on why the tradeoff between analysts and attackers is tractable – also applies to Theorem 5.5.1.)

Theorem 5.6.3 is powerful because it holds for arbitrary specifications of the data model P_θ and is applicable to the agent's arbitrary (proper) prior $\pi(\theta)$. So long as d_{**} is finite (see the discussion after Theorem 5.4.1 on why this is not unreasonable), the bounds in (5.12) are non-trivial.

With that said, whenever d_{**} is large, the bounds provided by Theorem 5.6.3 are wide, rendering the results weakly informative at best. Indeed, rather than a pair of wide posterior bounds, the agent would be better off with a precise Bayesian posterior, which is theoretically derivable via the simple relation

$$\pi(\theta \mid t) \propto \pi(\theta)p(t \mid \theta), \quad (5.13)$$

where $p(t \mid \theta)$ can in turn be derived from the convolution of the data model P_θ and the privacy mechanism P_x according to (5.6). In practice, however, direct computation or sampling from (5.13) is not always possible or feasible. Such difficulties arise in situations A) where the privacy mechanism P_x is not fully transparent to the analyst due to its complex dependence on x , whether by design or by post-processing [Gong \(2022b\)](#); B) where the data model P_θ is intractable, such as if defined algorithmically or treated as a black-box; or C) where their convolution (5.6), typically an n -dimensional integral, is intractable. Under any of these situations, the analyst may still rely on Theorem 5.6.3 to obtain bounds on their posterior.

Despite their width, these bounds are optimal whenever ε is the smallest constant satisfying the Lipschitz condition (5.2). Without adding further assumptions on M , P_θ , or π , these bounds cannot be

shrunk. (This also applies to the bounds from Sections 5.4 and 5.5. We prove this in Section 5.9.) Yet they are not necessarily tight at a given θ . This deficiency is an inevitable consequence of our analysis, which replaced the average case, $\int p_x(t) dP_\theta(x)$, with the extremal case, $p_{x_*}(t) \exp(\varepsilon d_*)$. Such an analysis is necessarily loose whenever there is any variation away from the extreme. But the analysis cannot be tightened without making assumptions about the nature of this variation – i.e. by making further assumptions on M , P_θ , or π .

We illustrate the posterior bounds of Theorem 5.6.3 with an example of Bayesian inference for a privatised count.

Example 5.6.5 (privatised single count). Suppose the database consists of a single count record $x \in \mathbb{N}$. We wish to query the value of x after it has been *clamped* to a pre-specified range $[a_0, a_1]$. That is, $q(x) = a_0$ if $x < a_0$, $q(x) = a_1$ if $x > a_1$, and $q(x) = x$ otherwise. In differentially private mechanism design, clamping can be a necessary procedure when the intended query has otherwise unbounded global sensitivity. Under clamping, the sensitivity is reduced to $\Delta(q) = a_1 - a_0$.

The analyst's Bayesian model is

$$\begin{aligned}\theta &\sim \text{Gamma}(\alpha, \beta), \\ x \mid \theta &\sim \text{Pois}(\theta), \\ t \mid x &\sim \text{Lap}(q(x); \varepsilon^{-1} \Delta(q)).\end{aligned}$$

For illustration, set $a_0 = 0$, $a_1 = 6$, $\alpha = 3$, $\beta = 1$. Figure 5.4 depicts in blue solid lines the upper and lower density bounds on the analyst's posterior distribution $p(\theta \mid t)$ as given by Theorem 5.6.3. With $\varepsilon = 1$ and $d_{**} = 1$, they are equal to the $\text{Gamma}(3, 1)$ prior density (blue dashed line), scaled by $\exp(\pm 1)$. Overlaid in grey are Monte Carlo posterior densities $p(\theta \mid t^{(k)})$, $k = 1, \dots, 10$, produced via the exact sam-

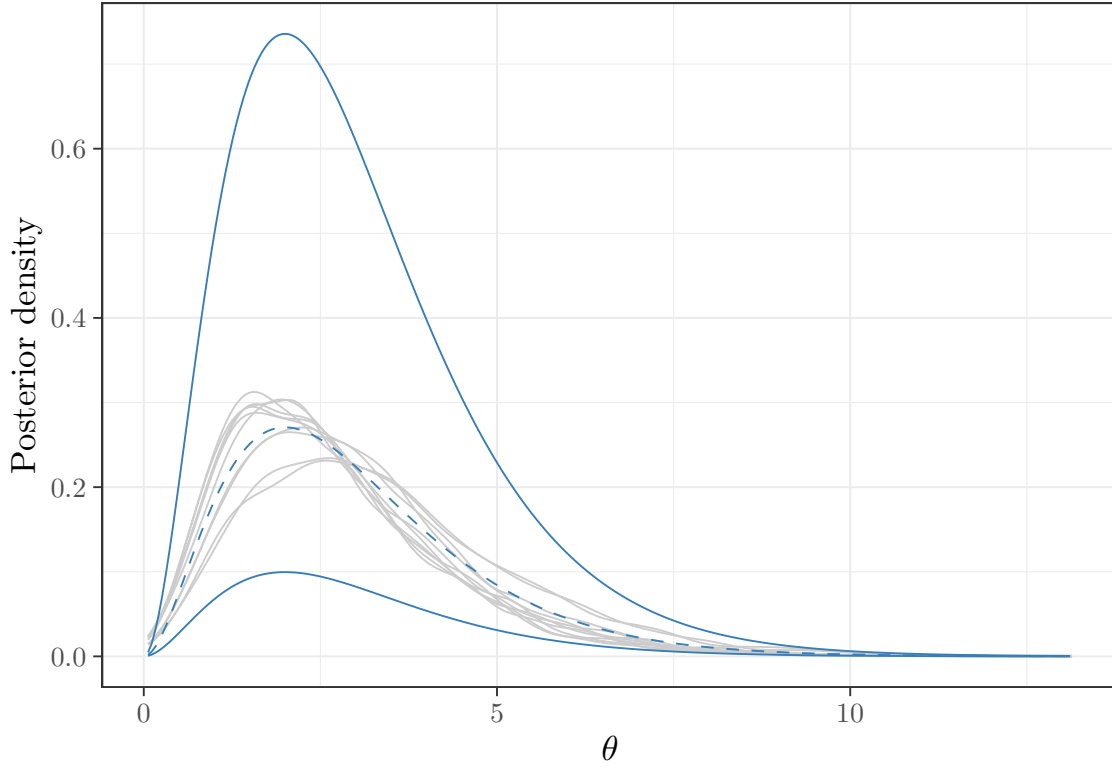


Figure 5.4: Density bounds for the posterior $p(\theta \mid t)$ from a privatised single count (Example 5.6.5). The dashed blue line is the density of the $\text{Gamma}(3, 1)$ distribution, the analyst's prior for θ . In grey are simulation-based posterior densities based on 10 realizations of t from its prior predictive distribution under the Poisson data model (Gong, 2022a). Upper and lower density bounds for the posterior $p(\theta \mid t)$ are in solid blue. The clamping range is $[0, 6]$ and the privacy loss is $\varepsilon = 1$.

pling algorithm proposed by Gong (2022a). Each $t^{(k)}$ is independently simulated from the prior predictive distribution of the above Bayesian model.

Several aspects of Example 5.6.5 are worth noting. First, the posterior density bounds (solid blue) are functions of the analyst's chosen prior $\pi(\theta)$ and the privacy mechanism parameters ε and d_{**} only. They are valid for any data model P_θ that the analyst wishes to employ, including (but not limited to) the Poisson data model that underlie the depicted precise posteriors densities $p(\theta \mid t^{(k)})$ in grey. On the other hand, while these precise posterior densities display moderate variations among each other, they do not depart

much from the prior density (dashed blue). This is due to the heavy-handedness of the privacy mechanism employed for this analysis, resulting in poor statistical utility of the privatised count t . Indeed, the mechanism injects Laplace noise with standard deviation of $\sqrt{2}\varepsilon^{-1}\Delta(q) = 8.48$ into a statistic clamped between $a_0 = 0$ and $a_1 = 6$. That t cannot be highly informative for the inferential problem at hand is correctly identified by the full Bayesian analysis which precisely accounts for the uncertainty induced by the privacy mechanism (grey lines). Furthermore, these precise posterior distributions are generally far from the bounds implied by Theorem 5.6.3; this re-enforces the shallowness of these bounds due to their validity for very general classes of the data model P_θ and priors π .

5.7 PUFFERFISH PRIVACY

As the classic formulation of differential privacy, pure ε -DP has inspired many variants, most of which closely resemble the original (for example by replacing D_{MULT} with some other distorting function d_{DIST} , see Remark 5.3.6). In contrast, Pufferfish privacy (Kifer and Machanavajjhala, 2014) is conceptually distinct from ε -DP in two ways. Firstly, while ε -DP conceptualises privacy as indistinguishability between pairs of comparable datasets $x, x' \in \mathcal{X}$ (Remark 5.3.7), Pufferfish reconceptualises privacy as indistinguishability between pairs of competing conjectures about the unobserved, confidential data x (as we will see in Remark 5.8.2). Secondly – and consequently – ε -DP is concerned solely with the design of the data-release mechanism, while the object of Pufferfish’s interest is the composition of the data-generating process and the data-release mechanism. We call this composite function the *data-provision procedure*:

Definition 5.7.1. Given a data-generating process $G(\theta, U_1)$ and a data-release mechanism $M(x, U_2)$, the

data-provision procedure $M_G : \Theta \times [0, 1]^2 \rightarrow \mathcal{T}$ is defined as

$$M_G(\theta, U_1, U_2) = M(G(\theta, U_1), U_2),$$

where U_1, U_2 are independent and (without loss of generality) identically distributed $\text{Unif}[0, 1]$.

Here U_1 and U_2 are the random components (i.e. *seeds*) of the data-generation G and of the data-release M respectively; θ is the data-generating model parameter; $X = G(\theta, U_1)$ is the (stochastic) dataset; and (as before) $T = M(X, U_2)$ is the released summary statistic.

As in Section 5.4, we require that the data universe \mathcal{X} is equipped with a σ -algebra \mathcal{G} ; that the data-release mechanism M is $(\mathcal{G} \otimes \mathcal{B}[0, 1], \mathcal{F})$ -measurable; and, further, that $G(\theta, \cdot)$ is $(\mathcal{B}[0, 1], \mathcal{G})$ -measurable for all $\theta \in \Theta$. (Recall that $\mathcal{B}[0, 1]$ is the Borel σ -algebra on $[0, 1]$ and \mathcal{F} is the σ -algebra of the output space \mathcal{T} .) We now also assume that the data-generating model parameter set Θ is in a one-to-one correspondence with the set of probability measures on $(\mathcal{X}, \mathcal{G})$. (So the data-generating model is not parametric in the usual sense of the term.)

Under these requirements, the seed U_1 induces a probability measure for the dataset X :

$$P_\theta(X \in E) = \lambda(\{u_1 \in [0, 1] : G(\theta, u_1) \in E\}), \quad (5.14)$$

(where λ is the Lebesgue measure) and – together with the seed U_2 – also for the output T of the data-provision procedure M_G :

$$\begin{aligned} P(T \in S \mid \theta) &= \lambda(\{u_1, u_2 \in [0, 1] : M_G(\theta, u_1, u_2) \in S\}) \\ &= \int_{\mathcal{X}} P_x(S) dP_\theta(x), \end{aligned} \quad (5.15)$$

(Equation (5.15) follows from the previous line by Fubini's theorem – see Proposition 18.) Recall from

Equation (5.6) that $P(T \in \cdot \mid \theta)$ is the privatised data probability. Because Pufferfish is concerned with the data-provision procedure M_G – and because $P(T \in \cdot \mid \theta)$ is the probability induced by M_G – the privatised data probability plays a central role in Pufferfish. In fact, we will see (in Definition 5.7.3) that Pufferfish can be conceived as a Lipschitz condition on the map $\theta \mapsto P(T \in \cdot \mid \theta)$, just as ε -DP is a Lipschitz condition on the map $x \mapsto P_x$.

Pufferfish provides a framework for developing tailored privacy definitions. The data custodian constructs their custom Pufferfish privacy definition according to their judgement of:

- A. The *attackers*: Against what kinds of background knowledge, or beliefs about the data, should the data-release mechanism M guard? (These knowledge and beliefs are modelled by probability distributions θ on the data X .)
- B. The *attackers' conjectures on confidential information*: Which parts of the dataset require protection (i.e. what are the confidential information?), and what conjectures may an attacker make about these information? (Conjectures are modelled as events $E \in \mathcal{G}$ on the data universe \mathcal{X} .)
- C. The *pairs of competing conjectures*: Which pairs of conjectures should remain indistinguishable to the attackers, even after observing (the realized value of) T ? (Or, for a Bayesian attacker, between which pairs of conjectures should it be impossible for the attacker to significantly improve their ability to distinguish?)

(Note that B. is only used as a stepping-stone for answering C.; it does not have an independent role to play in Pufferfish.) Putting the above discussion more formally, the data custodian specifies their privacy definition with a *Pufferfish instantiation*:

Definition 5.7.2. A *Pufferfish instantiation* $\varepsilon\text{-PufferFish}(\mathbb{D}, \mathbb{S})$ is a tuple with three components:

1. A set $\mathbb{D} \subset \Theta$ of *attackers*;
2. A set $\mathbb{S} \subset \mathcal{G} \times \mathcal{G}$ of *pairs of competing conjectures* about “confidential” information in the dataset x ;⁹ and

⁹Note that elsewhere in the literature (for example in Kifer and Machanavajjhala (2014)), what we term the “at-

3. A *privacy loss budget* $0 \leq \varepsilon \leq \infty$.

The sets \mathbb{D} and \mathbb{S} correspond to points A. and C. respectively in the above discussion. The privacy loss budget ε has the same role in Pufferfish as in ε -DP: it describes the degree of continuity of the map $\theta \mapsto P(T \in \cdot \mid \theta)$ – and hence, intuitively, the degree of privacy afforded to the data – with smaller ε corresponding to more continuity/privacy.

Pufferfish privacy is a Lipschitz condition on M_G :

Definition 5.7.3. Fix the data-generating process G . A data-release mechanism M *satisfies the Pufferfish instantiation* $\varepsilon\text{-PufferFish}(\mathbb{D}, \mathbb{S})$ if the associated data-provision procedure M_G satisfies the inequality:

$$D_{\text{MULT}}\left(P(T \in \cdot \mid \theta), P(T \in \cdot \mid \theta')\right) \leq \varepsilon d_{\mathbb{D}, \mathbb{S}}(\theta, \theta'),$$

for all $\theta, \theta' \in \Theta$.

Here $d_{\mathbb{D}, \mathbb{S}}$ is a metric on Θ which is given by the Pufferfish instantiation $\varepsilon\text{-PufferFish}(\mathbb{D}, \mathbb{S})$. It is defined as follows: Firstly, as Θ is in a one-to-one correspondence with the set of probability measures on $(\mathcal{X}, \mathcal{G})$, it is closed under conditioning. That is, for each $\theta \in \Theta$ and for each event $E \in \mathcal{G}$, there exists a unique $\theta' \in \Theta$ such that $P_{\theta'}(X \in \cdot) = P_{\theta}(X \in \cdot \mid X \in E)$ (provided that $P_{\theta}(X \in \cdot \mid X \in E)$ is well-defined¹⁰). Denote this θ' by $\theta|_E$.

tackers’ conjectures” are referred to as “secrets,” and the “pairs of competing conjectures” are referred to as “discriminative pairs.” Moreover, in [Kifer and Machanavajjhala \(2014\)](#), the set of discriminative pairs is denoted by $\mathbb{S}_{\text{pairs}}$ and \mathbb{S} instead denotes the set of secrets/conjectures. We choose to omit the set of secrets/conjectures from a Pufferfish instantiation as it is superfluous, and instead use \mathbb{S} to denote the set of pairs of competing conjectures.

¹⁰Pufferfish limits its consideration to $\theta \in \Theta$ and $E, E' \in \mathcal{G}$ for which $P_{\theta}(X \in \cdot \mid X \in E)$ and $P_{\theta}(X \in \cdot \mid X \in E')$ are well-defined, in order to ensure that the data-provision procedures $P(T \in \cdot \mid \theta, X \in E)$ and $P(T \in \cdot \mid \theta, X \in E')$

Then let the graph $G_{\mathbb{D}, \mathbb{S}}$ on Θ have edges (θ, θ') if there exists some $\theta^* \in \mathbb{D}$ and some $(E, E') \in \mathbb{S}$ such that $\theta = \theta^*|_E$ and $\theta' = \theta^*|_{E'}$ – i.e. such that P_θ is equal to the conditional distribution $P_{\theta^*}(X \in \cdot \mid X \in E)$ and $P_{\theta'}$ is equal to $P_{\theta^*}(X \in \cdot \mid X \in E')$. Finally, define $d_{\mathbb{D}, \mathbb{S}}(\theta, \theta')$ as the length of a shortest path between θ and θ' in $G_{\mathbb{D}, \mathbb{S}}$.

Therefore, for all $\theta^* \in \mathbb{D}$ and all $(E, E') \in \mathbb{S}$ (with $P_{\theta^*}(X \in \cdot \mid X \in E)$ and $P_{\theta^*}(X \in \cdot \mid X \in E')$ both well-defined¹⁰), the data-generating probabilities $\theta^*|_E$ and $\theta^*|_{E'}$ are adjacent in the graph $G_{\mathbb{D}, \mathbb{S}}$ and hence

$$D_{\text{MULT}}\left(P(T \in \cdot \mid \theta^*, X \in E), P(T \in \cdot \mid \theta^*, X \in E')\right) \leq \varepsilon. \quad (5.16)$$

The above discussion sheds light on the differences between Pufferfish and pure ε -DP: As observed earlier, ε -DP is concerned with indistinguishability of datasets $x, x' \in \mathcal{X}$. Hence, its starting point is the data universe \mathcal{X} and it is a Lipschitz condition on the data-release mechanism $M : \mathcal{X} \times [0, 1] \rightarrow \mathcal{T}$. On the other hand, Pufferfish is concerned with competing conjectures $\theta^*|_E$ and $\theta^*|_{E'}$ (for $\theta^* \in \mathbb{D}$ and $(E, E') \in \mathbb{S}$). Its starting point is thus the data-generating parameter set Θ and it is a Lipschitz condition on the data-provision procedure $M_G : \Theta \times [0, 1]^2 \rightarrow \mathcal{T}$. Yet the data custodian only has partial control over M_G . That is to say, while Pufferfish is a property of the data-provision procedure M_G , the data custodian can achieve this property only through the design of M . In contrast, the data custodian often has full

are themselves well-defined.

However, determining whether or not $P_\theta(X \in \cdot \mid X \in E)$ can be well-defined is beyond the scope of this paper. Answering this question to the necessary level of generality is difficult (see [Chang and Pollard \(1997\)](#) and references therein), but the majority of cases encountered in practice are covered by two approaches: When $P_\theta(X \in E)$ is non-zero, $P_\theta(X \in A \mid X \in E)$ is defined as $P_\theta(X \in A \cap E) / P_\theta(X \in E)$. And when $X = (Y, Z)$ has a canonical density $f(Y = y, Z = z)$ on a product measure $\mu = \mu_1 \times \mu_2$ with $\{X \in E\} = \{Z = z_E\}$ for some z_E , then $P_\theta(X \in \cdot \mid X \in E)$ is defined as the regular conditional probability $f_{Y|Z}(\cdot \mid z_E) d\mu_1(\cdot)$ where

$$f_{Y|Z}(y \mid z) = \begin{cases} \frac{f(y, z)}{f(z)} & \text{if } f(z) > 0, \\ \varphi(y) & \text{otherwise,} \end{cases}$$

with $f(z) = \int f(y, z) d\mu_1(y)$ and $\varphi(y)$ an arbitrary density on μ_1 ([Durrett, 2019](#), Example 4.1.6).

control of the object of ε -DP's interest, the data-release mechanism M .

While ε -DP allows for the use of an arbitrary distance d , Pufferfish makes a very particular choice for the distance on its input space: $d_{\mathbb{D}, \mathbb{S}}$. (However, beyond the interpretation of $d_{\mathbb{D}, \mathbb{S}}$ in terms of attackers and competing conjectures, there is no reason in principle that Pufferfish cannot be generalised to allow for arbitrary distances d on Θ . In fact, all of the results below generalise immediately from $d_{\mathbb{D}, \mathbb{S}}$ to any metric d on Θ which satisfies Assumption 5.3.3.)

Despite their differences, ε -DP and Pufferfish share one characteristic which is very important for our purposes: They both use the multiplicative distance D_{MULT} to measure the change in output variations. This means that Pufferfish is fundamentally linked to the concept of an interval of measures, just as ε -DP is. This connection to intervals of measures and the resulting implications on the indistinguishability of important inferential quantities are the subject of the next section.

5.8 AN IP VIEW OF PUFFERFISH PRIVACY

As an analog to Theorem 5.3.5 for pure ε -differential privacy, Theorem 5.8.1 below establishes the connection between Pufferfish and intervals of measures. Specifically, 5.8.1.II is the standard definition of Pufferfish as given in [Kifer and Machanavajjhala \(2014\)](#). The equivalence 5.8.1.I \Leftrightarrow 5.8.1.II justifies the formulation of Pufferfish as Lipschitz continuity. In addition, 5.8.1.III and 5.8.1.IV give novel formulations of Pufferfish in terms of intervals of measures.

Theorem 5.8.1. *Fix the data-generating process $G(\theta, U_1)$ and the data-release mechanism $M(x, U_2)$. For any Pufferfish instantiation $\varepsilon\text{-PufferFish}(\mathbb{D}, \mathbb{S})$ with privacy loss budget $\varepsilon < \infty$, the following statements are equivalent:*

I M satisfies $\varepsilon\text{-PufferFish}(\mathbb{D}, \mathbb{S})$.

II For all $S \in \mathcal{F}$, all competing conjectures $(E, E') \in \mathbb{S}$ and all attackers $\theta^* \in \mathbb{D}$ (such that $P_{\theta^*}(X \in \cdot \mid X \in E)$ and $P_{\theta^*}(X \in \cdot \mid X \in E')$ are both well-defined¹⁰), the following inequalities are satisfied:

$$\begin{aligned} P(T \in S \mid \theta^*, X \in E) &\leq e^\varepsilon P(T \in S \mid \theta^*, X \in E'), \\ P(T \in S \mid \theta^*, X \in E') &\leq e^\varepsilon P(T \in S \mid \theta^*, X \in E). \end{aligned}$$

III For all $\delta \in \mathbb{N}$ and all $\theta, \theta' \in \Theta$ with $d_{\mathbb{D}, \mathbb{S}}(\theta, \theta') = \delta$,

$$P(T \in \cdot \mid \theta') \in \mathcal{I}_1(L_{\theta, \delta\varepsilon}, U_{\theta, \delta\varepsilon}),$$

where $L_{\theta, \delta\varepsilon} = e^{-\delta\varepsilon} P(T \in \cdot \mid \theta)$ and $U_{\theta, \delta\varepsilon} = e^{\delta\varepsilon} P(T \in \cdot \mid \theta)$.

IV For all $\theta \in \Theta$ and all measures $\nu \in \Omega$, if $P(T \in \cdot \mid \theta)$ has a density $p(t \mid \theta)$ with respect to ν , then for every $d_{\mathbb{D}, \mathbb{S}}$ -connected¹¹ $\theta' \in [\theta]$, $P(T \in \cdot \mid \theta')$ also has a density $p(t \mid \theta')$ (with respect to ν) satisfying

$$p(t \mid \theta') \in p(t \mid \theta) \exp[\pm \varepsilon d_{\mathbb{D}, \mathbb{S}}(\theta, \theta')],$$

for all $t \in \mathcal{T}$.

Proof. First note that (5.16) is equivalent to both I and II. Specifically, that (5.16) implies I follows by applying the triangle inequality to a shortest path between θ and θ' in $G_{\mathbb{D}, \mathbb{S}}$, similar to the proof of 5.3.5.I \Leftrightarrow 5.3.5.II in Theorem 5.3.5. The remainder of the proof is analogous to that of Theorem 5.3.5. \square

Remark 5.8.2. Statement IV of Theorem 5.3.5 is the backbone for reasoning about indistinguishability between d -connected datasets $x, x' \in \mathcal{X}$ under ε -DP (see Remark 5.3.7). In contrast, statement IV of Theorem 5.8.1 provides the rationale for indistinguishability between $d_{\mathbb{D}, \mathbb{S}}$ -connected distributions $\theta, \theta' \in \Theta$ under Pufferfish privacy. Specifically, for $\theta = \theta^*|_E$ and $\theta' = \theta^*|_{E'}$ (with $\theta^* \in \mathbb{D}$ and $(E, E') \in \mathbb{S}$), an attacker cannot distinguish with certainty between θ and θ' because $p(t \mid \theta)/p(t \mid \theta')$ is bounded away from zero and infinity, regardless of the value of t . More generally, whenever θ is plausible (i.e. when

¹¹ Analogous to the concept of connected data $x, x' \in \mathcal{X}$ (Definition 5.3.2), we say that $\theta, \theta' \in \Theta$ are $d_{\mathbb{D}, \mathbb{S}}$ -connected if $d_{\mathbb{D}, \mathbb{S}}(\theta, \theta') < \infty$ and we define $[\theta]$ to be θ 's connected component: $[\theta] = \{\theta' \in \Theta \mid d_{\mathbb{D}, \mathbb{S}}(\theta, \theta') < \infty\}$.

$p(t \mid \theta) > 0$) then all $d_{\mathbb{D}, \mathbb{S}}$ -connected $\theta' \in [\theta]$ are also plausible (regardless of the choice of dominating measure ν).

We now turn to discussing the impact of Pufferfish privacy on statistical inference in both Bayesian and frequentist paradigms. From the Bayesian view, Pufferfish limits the ability of an attacker $\theta^* \in \mathbb{D}$ to discern between two competing conjectures $(E, E') \in \mathbb{S}$ relative to their prior (baseline) ability to do so:

$$e^{-\varepsilon} \leq \frac{P_{\theta^*}(X \in E \mid T = t)}{P_{\theta^*}(X \in E' \mid T = t)} \bigg/ \frac{P_{\theta^*}(X \in E)}{P_{\theta^*}(X \in E')} \leq e^{\varepsilon}, \quad (5.17)$$

where the attacker's “ability to discern between (E, E') ” is quantified as the odds of E against E' , so that (5.17) is a bound on the prior-to-posterior odds ratio. In fact, \mathcal{M} satisfies ε -PufferFish(\mathbb{D}, \mathbb{S}) if and only if \mathcal{M} satisfies (5.17) for all $\theta^* \in \mathbb{D}$, all¹⁰ $(E, E') \in \mathbb{S}$ and almost all $t \in \mathcal{T}$. (This result was first described in (Kifer and Machanavajjhala, 2014, p. 6) and follows from Statement 5.8.1.IV by setting $\theta = \theta^*|_E$ and $\theta' = \theta^*|_{E'}$, and then applying Bayes rule. We formally state and prove this result in Proposition 19 of Appendix C.5.)

As previous sections contend, there is an important type of competing conjectures that privacy mechanisms aim to make indistinguishable. These conjectures concern the values of the records in the dataset x . To this end, Pufferfish provides a Bayesian semantic guarantee that conforms to the structure of a *density ratio neighbourhood* Wasserman and Kadane (1992); Wasserman (1992), defined below.

Recall (from Definition 5.3.1) that Ω is the set of σ -finite measures on the measurable space $(\mathcal{T}, \mathcal{F})$.

Definition 5.8.3. The *density ratio neighbourhood* of $\mu \in \Omega$ with radius $r \geq 0$ is defined as

$$N_r(\mu) = \{\nu \in \Omega : d_{\text{DR}}(\mu, \nu) \leq r\},$$

where d_{DR} is the *density ratio metric*:

$$d_{\text{DR}}(\mu, \nu) = \begin{cases} 0 & \text{if } \mu = \nu = 0 \\ \text{ess sup}_{t, t' \in \mathcal{T}^\circ} \ln \left(\frac{f(t)}{f(t')} / \frac{g(t)}{g(t')} \right) & \text{else if } \mu, \nu \text{ are mutually absolutely continuous,} \\ \infty & \text{otherwise,} \end{cases} \quad (5.18)$$

with f and g densities of σ -finite measures μ and ν respectively, with respect to some common dominating measure $\tau \in \Omega$; $\mathcal{T}^\circ = \{t \in \mathcal{T} \mid 0 < f(t), g(t) < \infty\}$; and the essential supremum is with respect to τ .¹²

The definition of the density ratio metric d_{DR} is well-defined in the sense that $d_{\text{DR}}(\mu, \nu)$ does not depend on the choice of f, g and τ in (5.18). (See Appendix C.2 for details.)

The following theorem characterises Pufferfish privacy (under a particular choice of \mathbb{S}) as the requirement that an attacker θ 's posterior on X is in the ε -density ratio neighbourhood of their prior:

Theorem 5.8.4. *Fix some $\theta^* \in \mathbb{D}$. Let \mathcal{S}_X be a partition of \mathcal{X} such that $P_{\theta^*}(X \in E)$, for $E \in \mathcal{S}_X$, is given by a density $p_{\theta^*}(Z = z_E)$ of some marginalisation Z of X . Define $\mathbb{S} = \mathcal{S}_X \times \mathcal{S}_X$.*

If M satisfies ε -PufferFish(\mathbb{D}, \mathbb{S}), then

$$P_{\theta^*}(Z \in \cdot \mid T = t) \in N_\varepsilon(P_{\theta^*}(Z \in \cdot)), \quad (5.19)$$

for $P(T \in \cdot \mid \theta^)$ -almost all $t \in \mathcal{T}$.*

In the other direction, suppose that $P_\theta(X \in E)$, for $E \in \mathcal{S}_X$, is given by a density $p_\theta(Z = z_E)$ for all $\theta \in \mathbb{D}$. Then (5.19) holding for all $\theta \in \mathbb{D}$ and $P(T \in \cdot \mid \theta)$ -almost all $t \in \mathcal{T}$ implies that M satisfies ε -PufferFish(\mathbb{D}, \mathbb{S}).

¹²The property $f, g < \infty$ holds τ -almost everywhere (because μ and ν are σ -finite – see the proof of Lemma C.5.2), and $f, g > 0$ holds μ - and ν -almost everywhere. Hence, practically one may take the essential supremum in equation (5.18) over \mathcal{T} ; restricting to \mathcal{T}° simply removes the complications of dividing by zero or infinity.

The proof of Theorem 5.8.4 is immediate from (5.17).

Two special cases of Theorem 5.8.4 are worth noting. Firstly, when the partition $\mathcal{S}_{\mathcal{X}} = \{\{x\} : x \in \mathcal{X}\}$ consists of all the singleton subsets of \mathcal{X} , an ε -Pufferfish(\mathbb{D}, \mathbb{S}) mechanism is tasked with providing indistinguishability between competing conjectures $E = \{x\}$ and $E' = \{x'\}$. That is, a Pufferfish mechanism must protect against the conjecture $X = x$ versus $X = x'$, for any arbitrary choices of $x, x' \in \mathcal{X}$. This is a tall order, because the dataset may contain a large number of individual records, each with values x_i, x'_i that are nothing alike. The resulting privacy guarantee is thus a stringent one: For any $\theta \in \mathbb{D}$,

$$P_{\theta|t} \in N_{\varepsilon}(P_{\theta}),$$

where P_{θ} and $P_{\theta|t}$ are $P_{\theta}(X \in \cdot)$ and $P_{\theta}(X \in \cdot \mid T = t)$ respectively. In other words, the Bayesian attacker θ 's ability to discern between two arbitrary datasets relative to their prior discernability (i.e. the prior-to-posterior odds ratio) is limited by a multiplicative factor of e^{ε} . (This is closely related to the ‘no-free-lunch privacy’ of Kifer and Machanavajjhala (2011) and (Kifer and Machanavajjhala, 2014, Section 3.2).)

A second, and more pragmatic, special case arises when the partition $\mathcal{S}_{\mathcal{X}}$ consists of the level sets given by fixing a small number of records in the dataset – in particular, by fixing a single record. Assume that the datasets $x \in \mathcal{X}$ are vectors (x_1, \dots, x_n) of records and let \mathcal{R} be the set of all possible values that a record x_i can take, so that $\mathcal{X} \subset \bigcup_{n=1}^{\infty} \mathcal{R}^n$. Define

$$\mathcal{S}_{\mathcal{X}} = \{E(r, 1) : r \in \mathcal{R}\}, \tag{5.20}$$

where $E(r, i)$ is the level set which fixes the i th record to be the value r :

$$E(r, i) = \{x \in \mathcal{X} : x_i = r\}. \tag{5.21}$$

For this choice of $\mathcal{S}_{\mathcal{X}}$, the two densities $p_{\theta}(Z = z_{E(r,1)})$ and $p_{\theta}(Z = z_{E(r,1)} \mid T = t)$ are, respectively, the prior marginal density, $p_{\theta}(X_1 = r)$, and the posterior marginal density, $p_{\theta}(X_1 = r \mid T = t)$, of the first record taking the value r , where the marginalisation is over all the other records with respect to the data-generating process θ . Theorem 5.8.4 states that these prior and posterior marginal densities are restricted to the same density ratio neighbourhood of radius ε .

Remark 5.8.5. When the competing conjectures \mathbb{S} are given by the level sets, $\varepsilon\text{-PufferFish}(\mathbb{D}, \mathbb{S})$ has a close connection to pure ε -DP. Indeed, suppose that $\mathcal{X} = \mathcal{R}^n$ for a fixed n and let $\mathbb{S} = \bigcup_{i=1}^n \{E(r, i) : r \in \mathcal{R}\}^2$. If \mathbb{D} is the collection of distributions on \mathcal{X} which take the records X_1, \dots, X_n as mutually independent, then a mechanism M satisfying $\varepsilon\text{-PufferFish}(\mathbb{D}, \mathbb{S})$ is equivalent to M satisfying ε -DP with $d = d_{\text{Ham}}$ (Kifer and Machanavajjhala, 2014, Theorem 6.1). This result follows from observing

$$P(T \in S \mid \theta, X_i = x_i) = \int_{\mathcal{R}^{n-1}} P_x(S) dP_{\theta}(X_{-i} = x_{-i}),$$

for $\theta \in \mathbb{D}$, which implies that Statements 5.3.5.II and 5.8.1.II are equivalent.

Pufferfish also has a frequentist interpretation as a limit to the power of any attacker's level- α test between competing conjectures (c.f. Theorem 5.5.1):

Theorem 5.8.6. *A data-release mechanism M satisfies $\varepsilon\text{-PufferFish}(\mathbb{D}, \mathbb{S})$ if and only if, for all $\theta_0 \neq \theta_1 \in \Theta$, the power $1 - \beta$ of all size- α tests of*

$$H_0 : \theta = \theta_0 \text{ versus } H_1 : \theta = \theta_1,$$

is bounded by the inequalities:

$$\max(\alpha/\varphi, 1 - [1 - \alpha]\varphi) \leq 1 - \beta \leq \min(\alpha\varphi, 1 - [1 - \alpha]/\varphi), \quad (5.22)$$

where $\varphi = \exp[\varepsilon d_{\mathbb{D}, \mathbb{S}}(\theta_0, \theta_1)]$.

Proof. The result is immediate from IV of Theorem 5.8.1 and the Neyman-Pearson lemma. (The derivation is analogous to the second half of the proof of Theorem 5.5.1 and the proof of Corollary 5.5.2.) \square

As an application of Theorem 5.8.6, suppose an attacker $\theta^* \in \mathbb{D}$ is interested in testing the null hypothesis $X \in E$ against the alternative $X \in E'$, for some $(E, E') \in \mathbb{S}$. This is equivalent to setting $\theta_0 = \theta^*|_E$ and $\theta_1 = \theta^*|_{E'}$ in Theorem 5.8.6 (assuming that $\theta^*|_E$ and $\theta^*|_{E'}$ are well-defined¹⁰). Hence, any such test with level 0.05 will have power at most $0.05 \exp(\varepsilon)$ under ε -PufferFish(\mathbb{D}, \mathbb{S}).

Remark 5.8.7. Because the density ratio neighbourhood $N_r(\mu)$ is the closed ball $B_{d_{\text{DR}}}^r(\mu)$, it is a distortion model (Remark 5.3.6). Moreover, it is closely related to the *constant odds ratio model* Walley (1991), which is the distortion model associated with the distorting function d_{COR} Montes et al. (2020a). Here, d_{COR} is the constant odds ratio metric:

$$d_{\text{COR}}(\mu, \nu) = \begin{cases} 0 & \text{if } \mu = \nu = 0, \\ 1 - \inf_{S, S' \in \mathcal{F}^*} \frac{\mu(S)\nu(S')}{\mu(S')\nu(S)} & \text{else if } \mu, \nu \text{ are mutually absolutely continuous,} \\ 1 & \text{otherwise,} \end{cases}$$

for finite $\mu, \nu \in \Omega$, where $\mathcal{F}^* = \{S \in \mathcal{F} : \mu(S) > 0\}$.

Corollary C.5.5 in Appendix C.5 proves that, for finite $\mu, \nu \in \Omega$,

$$d_{\text{DR}}(\mu, \nu) = -\ln[1 - d_{\text{COR}}(\mu, \nu)].$$

Hence, when restricting to finite measures, the density ratio neighbourhood N_r is equal to the constant odds ratio model with distortion parameter $\delta = 1 - \exp(-r)$.

5.9 OPTIMALITY OF THIS PAPER'S RESULTS

The bounds presented in this paper cannot be improved without additional assumptions on the data-release mechanism \mathcal{M} , the data-generating model P_θ or the prior π . In this section, we provide examples which demonstrate the optimality of these bounds. However, it is important to reiterate that these bounds are only tight pointwise. Indeed it would be impossible for it to be otherwise, since the bounds are on probability measures, yet the bounds themselves are not probability measures.

Throughout this section, we rely on the Laplace mechanism \mathcal{M} for the count query $q(x) = \sum_i x_i$ (Example 5.2.2). The density of $T \sim \mathcal{M}(x, U)$ is $p_x(t) = \frac{\varepsilon}{2} \exp(-\varepsilon|t - q(x)|)$ when $\mathcal{X} = \{0, 1\}^n$ and the metric d on \mathcal{X} is the Hamming distance d_{Ham} . We assume that n is fixed, so that \mathcal{X} is d_{Ham} -connected and $\sup_{x, x' \in \mathcal{X}} d_{\text{Ham}}(x, x') = n < \infty$.

We begin with Theorem 5.4.1 which states that the privatised data probability $P(T \in \cdot \mid \theta)$ is bounded by $L_{\theta, \varepsilon}$ and $U_{\theta, \varepsilon}$ on \mathcal{T}_0 . Because $p(t \mid \theta) = \int p_x(t) dP_\theta(x)$ for a.e. $t \in \mathcal{T}_0$, the lower bound $p(t \mid \theta) \geq l_{\theta, \varepsilon}(t)$ is tight if $p_x(t) = l_{\theta, \varepsilon}(t)$ for P_θ -a.e. $x \in \text{supp}(x \mid t_0, \theta)$. Consider the Laplace mechanism under the setting given above. In this case, $\mathcal{T}_0 = \mathbb{R}$ for any t_0 and any θ . Further, the essential-supremum in $l_{\theta, \varepsilon}(t)$ is achieved by $x_* = (1, \dots, 1)$ when $t \leq 0$. Hence $l_{\theta, \varepsilon}(t) = p_{x_0}(t)$ where $x_0 = (0, \dots, 0)$. Therefore, $p(t \mid \theta)$ can be arbitrarily close to $l_{\theta, \varepsilon}(t)$ for $t \leq 0$ as P_θ concentrates on x_0 . This implies $P(T \in S \mid \theta)$ can be arbitrarily close to $L_{\theta, \varepsilon}(S)$ for a bounded, measurable set $S \subset \mathbb{R}^{\leq 0}$. The upper bound $P(T \in \cdot \mid \theta) \leq U_{\theta, \varepsilon}$ follows similarly, by considering $t \geq n$, $x_* = (0, \dots, 0)$ and $x_0 = (1, \dots, 1)$.

We now move to Theorem 5.5.1 which concerns the power of hypothesis tests in the private setting. To see that this result is tight, consider the model $\mathcal{P} = \{P_\theta \mid \theta \in \{0, 1\}^n\}$ where $P_\theta(X \in \cdot)$ is the point mass on $x = \theta$. Set $\theta_0 = (0, \dots, 0)$ and $\theta_1 = (1, \dots, 1)$. By examining the density $p_x(t)$ of the Laplace

mechanism for $x = (0, \dots, 0)$ and for $x = (1, \dots, 1)$, one can conclude that the Neyman-Pearson (NP) test must have a rejection region of the form $R = \{t > t_1\}$ for some t_1 . Moreover, for small enough ε , t_1 must be at least $n = d_{**}$ (assuming $\alpha < 0.5$). Then, $p(t \mid \theta_1) = \exp(\varepsilon n)p(t \mid \theta_0)$ for all $t \in R$, which means the NP test has power exactly $\alpha \exp(\varepsilon n)$.

Theorem 5.6.2 provides bounds on a Bayesian analyst's prior predictive probability. If one sets the prior π to be a point mass on a single θ_0 , then the prior predictive probability $P(T \in S) = \int_{\Theta} P(T \in S \mid \theta) d\pi(\theta)$ reduces to the privatised data probability $P(T \in S \mid \theta_0)$. In this case, proving optimality of Theorem 5.6.2 is analogous to proving that the bounds $L_{\theta_0, \varepsilon} \leq P(T \in \cdot \mid \theta_0) \leq U_{\theta_0, \varepsilon}$ are tight. Hence, the argument outline above for Theorem 5.4.1 can also be used to show optimality of Theorem 5.6.2.

For Theorem 5.6.3 – which demonstrates that a Bayesian's posterior is within a probability interval of the prior – take $\Theta = [0, 1]$ with the prior $\pi = \text{Unif}[0, 1]$. Let $P_{\theta}(x)$ be the point mass on $(1, \dots, 1)$ if $\theta = 1$ and the point mass on $(0, \dots, 0)$ otherwise. For $t > n$, we have $\pi(\theta = 1 \mid t) = \pi(\theta = 1) \exp(\varepsilon n)$. Thus, the bound in Theorem 5.6.3 is achieved since $d_{**} = n$.

Finally, we prove that our results on the inferential limits induced by Pufferfish privacy (Theorems 5.8.4 and 5.8.6) are tight. For this, consider the setting described in Remark 5.8.5. In this case, the Laplace mechanism \mathcal{M} with $\mathcal{X} = \{0, 1\}^n$ satisfies $\varepsilon\text{-PufferFish}(\mathbb{D}, \mathbb{S})$. Theorem 5.8.4 states that, for any $i \in \{1, \dots, n\}$ and any $\theta^* \in \mathbb{D}$, the posterior $P_{\theta^*}(X_i \in \cdot \mid T = t)$ is in the density ratio neighbourhood of radius ε that is centred at the prior $P_{\theta^*}(X_i \in \cdot)$. Let θ^* be such that

$$P_{\theta^*}(X = x_0) = P_{\theta^*}(X = x_1) = 0.5,$$

where $x_0 = (0, \dots, 0)$ and $x_1 = (1, 0, \dots, 0)$. Then

$$\frac{P_{\theta^*}(X_1 = 0 \mid T = t)}{P_{\theta^*}(X_1 = 1 \mid T = t)} \bigg/ \frac{P_{\theta^*}(X_1 = 0)}{P_{\theta^*}(X_1 = 1)} = \frac{p_{x_0}(t)}{p_{x_1}(t)},$$

by Bayes rule, where $p_x(t)$ is the density of the Laplace mechanism. Yet $p_{x_0}(t)/p_{x_1}(t) = \exp(\varepsilon)$ for $t \leq 0$.

Hence,

$$d_{\text{DR}}\left(P_{\theta^*}(X_1 \in \cdot \mid T = t), P_{\theta^*}(X_1 \in \cdot)\right) = \varepsilon,$$

and thus the bound (5.19) of Theorem 5.8.4 is tight.

Now we consider Theorem 5.8.6, which provides a bound on the power of any size- α test. As before, we use the Pufferfish instantiation given in Remark 5.8.5 and the Laplace mechanism \mathcal{M} . Let θ_0 and θ_1 be such that $P_{\theta_i}(X \in \cdot)$ is the point mass on x_i , where $x_0 = (0, \dots, 0)$ and $x_1 = (1, \dots, 1)$ with $|x_0| = |x_1| = n$. Then $d_{\mathbb{D}, \mathbb{S}}(\theta_0, \theta_1) = n$ and, furthermore, the test $H_0 : \theta = \theta_0$ versus $H_1 : \theta = \theta_1$ is exactly the test we examined when proving the optimality of Theorem 5.5.1. In that proof, we demonstrated that the Neyman-Pearson test has power $\alpha \exp(\varepsilon n)$ (assuming that ε is small and $\alpha < 0.5$). This implies the bounds (5.22) in Theorem 5.8.6 are tight.

5.10 DISCUSSION

The results we obtain in this paper make novel contributions to the differential privacy literature in the following ways. Firstly, the bounds we obtain in Theorems 5.4.1, 5.5.1, 5.6.2, 5.6.3, 5.8.4 and 5.8.6 are non-trivial, due to the validity of these results across a broad range of data models, privacy mechanisms and prior distributions. When the analyst has little knowledge or is only willing to make minimal assumptions about their model, these bounds are useful representations of the limits of statistical learning under privacy constraints. This draws a contrast with the existing DP literature, which has largely focused on asymptotic

lower bounds or on constructing (asymptotically-)optimal data-release mechanisms for specific data use cases [Smith \(2011\)](#); [Cai et al. \(2021\)](#); [Chhor and Sentenac \(2023\)](#); [Duchi et al. \(2018\)](#); [Bassily et al. \(2014\)](#); [Talwar et al. \(2015\)](#); [Dwork et al. \(2014\)](#); [Wasserman and Zhou \(2010\)](#); [Awan and Slavković \(2020\)](#). This literature aligns with *query-based access* [Hotz et al. \(2022\)](#) where the user can choose what statistics are released. Our results, on the other hand, are finite-sample and apply to the *dissemination mode* of data release where the mechanism is not tailored for the analyst’s use case. This setting is typical of official statistics (e.g. censuses and surveys) and, more generally, data products with multiple users, and is more common in the research community than query-based access [Hotz et al. \(2022\)](#).

Secondly, the generality of our bounds implies that they are inherent consequences of the privacy standards themselves, be it pure ε -DP or Pufferfish. Specifically, these bounds stem only from the requirement that the mechanism M is ε -DP – or ε -PufferFish(\mathbb{D}, \mathbb{S}) – and not on any particularities of P_θ, M or π . That these bounds are typically wide in practice – as can be seen from Examples 5.4.3 and 5.6.5 – is in part due to the near-total lack of assumption under which they are derived. While these bounds can approach vacuity as the data size n grows, in practical examples that need not be the case if, for example, the data analyst has probabilistic knowledge about the privacy mechanism (see e.g. Example 5.6.5) or the data space \mathcal{X} . For a given choice of P_θ, M and π , we may obtain tighter bounds than those in this paper. In addition to the asymptotic results in the aforementioned papers, sharp bounds for specific P_θ, M and π may be derivable from the existing literature on measurement errors (*errors-in-variables*) in statistics and econometrics, particularly in the case of point identification problems (see e.g. ([Carroll and Hall, 1988](#); [Horowitz and Manski, 1995](#))).

Through the lens of Theorems 5.5.1, 5.6.3, 5.8.4 and 5.8.6, we obtain valuable insights in both frequen-

tist and Bayesian paradigms on the *privacy-utility trade-off* which is fundamental to differential privacy as a quantitative privacy standard Hotz et al. (2022). (For details, see the discussion accompanying these theorems.) Qualitatively speaking, there exists an inherent tension between protecting private information and deriving scientific knowledge. To date, quantitative approaches to this trade-off predominantly rely on the privacy loss budget as the sole metric to balance this trade-off Abowd and Schmutte (2019); Hsu et al. (2014); Heffetz (2022). However, from the suite of IP analyses presented here, we see that other building blocks – notably the metric structure (\mathcal{X}, d) of the data universe, the associated database distances (such as d_* and d_{**}), and the clamping parameters on \mathcal{X} – are all relevant factors that, together with the privacy loss budget ϵ , collectively determine the limits to statistical learning for attackers and scientists alike. Therefore, ϵ is not the only parameter of concern – and perhaps not even the central concern – when assessing and trading-off privacy and utility Bailie et al. (2025b).

While this paper qualifies the tradeoff of ϵ -DP and Pufferfish privacy in concrete terms, it narrowly conceives privacy and utility in terms of the extent of statistical estimation attainable under either frequentist or Bayesian paradigms. There are, of course, many other aspects of utility that are worth examining – such as the ease of analysis, use of computational resources, facial validity and logical consistency boyd and Sarathy (2022); Hotz and Salvo (2022); Ruggles et al. (2019) – and other paradigms (in particular decision theory) with which the notions of privacy and utility can be quantified. In fact, both notions are multi-faceted and context-specific and, as one of the reviewers of Bailie and Gong (2023a) pointed out, a judicious conceptualisation of privacy and utility may improve their tradeoff’s efficiency frontier. Acknowledging the complex makeup of this tradeoff, we advocate for future design and analysis of data-release mechanisms to treat the conceptualisations of privacy and utility – and the roles that \mathcal{X}, d and

the distorting function d_{DIST} play in these conceptualisations – with scrutiny, given their scarcity in the current literature (Bailie et al., 2025c).

Tools from the IP literature harbour potential in aiding future endeavours to study statistical data privacy in a rigorous yet general manner. This work has examined some examples in which a DP definition can be formulated as the requirement that a mechanism’s probability P_x under one input $x \in \mathcal{X}$ is in a certain distortion model of its probability $P_{x'}$ under another input x' , whenever those two inputs x, x' are connected. In this case, the choice of distorting function d_{DIST} (partially) determines the flavor of the privacy guarantee. One direction for future research is thus to explore IP characterisations of other common flavors of DP, in particular (ε, δ) -DP (Dwork et al., 2006a; Machanavajjhala et al., 2008; Kifer et al., 2022), zero-concentrated DP (zCDP) (Dwork and Rothblum, 2016; Bun and Steinke, 2016), Rényi DP (Mironov, 2017) and Gaussian DP (Dong et al., 2022), which are popular in practice due to flexible privacy mechanism design, better privacy budget accounting and increased statistical efficiency. Since these variants, and others such as subspace DP (Gao et al., 2022), stem from changing \mathcal{X} , d or the distorting function d_{DIST} (Remark 5.3.6), three key questions are 1) how the distorting function d_{DIST} corresponding to a DP variant can be characterised as an IP object; 2) what IP properties does d_{DIST} have; and 3) what are the consequences on statistical inference from using d_{DIST} to constrain P_x and $P_{x'}$ (for connected $x, x' \in \mathcal{X}$). For example, our preliminary analysis shows that (ε, δ) -DP cannot be described by an interval of measures, at least not alone. Moreover, we have demonstrated the necessity of using D_{MULT} as the distorting function for ensuring the types of bounds on frequentist and Bayesian inference found in Sections 5.5 and 5.6 (see Remark 5.6.4). This implies other variants of DP which replace D_{MULT} with some other distorting function d_{DIST} cannot satisfy the Bayesian and frequentist semantics described in this article.

A second direction for future research is suggested by the characterisation of Pufferfish as a Lipschitz condition with the metric D_{MULT} on Ω and the metric $d_{\mathbb{D},\mathbb{S}}$ on Θ . Replacing D_{MULT} with some other distorting function d_{DIST} – for example the distorting functions corresponding to (ε, δ) -DP or to zCDP (see [Baillie et al. \(2025b\)](#) for the definitions of these distorting functions) – would generate new variants of Pufferfish. (Some such variants have already been proposed in [Zhang et al. \(2022\)](#); [Ding \(2024\)](#).) Similarly, replacing $d_{\mathbb{D},\mathbb{S}}$ with another metric on Θ would generalise Pufferfish beyond the interpretation of attackers and competing conjectures, and would provide a more-flexible framework for expressing privacy as ε -indistinguishability (Remarks 5.3.7 and 5.8.2) between distributions $\theta, \theta' \in \Theta$: By specifying their own metric on Θ , the data custodian can choose which θ and θ' should be ε -indistinguishable according to their knowledge and expertise, rather than being restricted to those (θ, θ') pairs which correspond to $(\theta^*|_E, \theta^*|_{E'})$ for some $\theta^* \in \mathbb{D}$ and some $(E, E') \in \mathbb{S}$.

Thirdly, a conceptually distinct IP approach to data privacy protection is the employment of SDL mechanisms that produce *set-valued* outputs. This approach has not been explicitly considered in this paper, although one can take the elements t of the output space \mathcal{T} to themselves be sets, and as such, the results from this paper may still be applied. SDL mechanisms which produce set-valued outputs – such as the “leaky” variant of Warner’s randomised response considered in [Li et al. \(2022\)](#) and the randomised censoring mechanisms considered in [Ding and Ding \(2022\)](#) – can be understood as an intentional *coarsening* of the data product [Heitjan and Rubin \(1990\)](#). A mechanism that produces set-valued outputs has a precise probability distribution that is given by the mass function associated with a *belief function* – a special type of coherent lower prevision [Shafer \(1976\)](#). As such, one can view set-valued mechanisms as inducing imprecise probabilities on \mathcal{T} – rather than precise probabilities P_x – where this imprecise probability is a

belief function. Compared to more general forms of IP, including the distortion models considered in this work, the mass function formulation of belief functions lends a computational advantage (particularly for Markov chain Monte Carlo – see e.g. [Jacob et al. \(2021\)](#)). On the other hand, it is less clear that set-valued outputs are practically acceptable for many of the real-world use-cases of SDL. Data users may anticipate point-valued data in most situations, and may not be prepared to conduct further statistical processing of set-valued outputs. In sum, the utility of the set-valued approach to SDL remains open to formulation and assessment in future research.

PART III

DIFFERENTIAL PRIVACY IN THE SURVEY CONTEXT

This page intentionally left blank.

6

Whose Data Is It Anyway? Towards a Formal Treatment of Differential Privacy for Surveys¹

6.1 INTRODUCTION

THE SURVEY IS THE WORKHORSE of statistical agencies. For example, the U.S. Census Bureau conducts more than 100 surveys annually (US Census Bureau, 2023o) including key data collections such as the American Community Survey (ACS), the Current Population Survey (CPS), the Survey of Income and Program Participation (SIPP) and the new Annual Business Survey (ABS). The data gathered through these surveys provide invaluable information on the US economy and on American society more generally. They are used by various stakeholders – for example, businesses, researchers, local and federal governments, media and not-for-profits – to make investment decisions, to inform policy and to allocate government funding, among many other uses.

¹Based on work coauthored with Jörg Drechsler.

On the other hand, national statistical organizations (NSOs) have acknowledged for decades their obligation to maintain the confidentiality of survey respondents due to legal and ethical considerations, but also to safeguard institutional trust and thus sustain the quality of their data products. To address these conflicting goals, various methods have been proposed over the years to protect the confidentiality of survey respondents while still maintaining the value of the data for the different stakeholders involved. In the last two decades, a new framework for assessing the privacy of statistical data products has emerged: *differential privacy* (DP) (Dwork et al., 2006b). This framework is mathematically appealing as it offers a formal guarantee: any single unit's influence on the probability of observing a specific output is bounded. This guarantee translates into quantifiable measures of protection against an adversary seeking to learn the confidential responses, although not without some complications (Cuff and Yu, 2016; Tschantz et al., 2020; Kifer and Machanavajjhala, 2011; Bailie et al., 2025e). Beyond the formal guarantees, DP is attractive as it allows for full transparency of the methods that were used to protect the output, without further loss to respondents' privacy beyond that associated with publishing the outputted statistics and the DP specification. This is in contrast to many methods that are currently employed at statistical agencies, which rely on hiding some of the parameters of the privacy-protection mechanism (such as the variance of the noise term when noise addition is used to protect a continuous attribute) to ensure privacy. With DP, agencies typically release all details about the mechanism including the levels of the privacy parameters. This implies that – at least in principle – users of the protected data will be able to account for the additional uncertainty introduced through the protection step (although this turns out to be difficult for many of the algorithms used in practice so far). Finally, DP offers several additional properties such as immunity to postprocessing and composition of privacy budget (see for example Dwork and Roth (2014)). This

second property makes the DP framework specifically interesting for statistical agencies as it allows for the quantification of the privacy loss over multiple data releases.

These attractive features have motivated the adoption of DP in the private sector (Erlingsson et al., 2014; Apple’s Differential Privacy Team, 2017; Ding et al., 2017; Messing et al., 2020a; Uber Security, 2017), as well as at some NSOs such as the Census Bureau (Machanavajjhala et al., 2008; Foote et al., 2019; Abowd and Hawes, 2023). Still, all deployments of DP so far have focused on situations in which the data to be protected coincided with the population of interest. As pointed out above, this is rarely the case for data collected by NSOs. Except for censuses – which are typically only conducted every five to ten years – and some administrative databases, most data at statistical agencies are collected via probability surveys. In the survey context, information is only gathered from a small fraction of the population, but the careful design of the selection process and several adjustment steps after the survey has been conducted (such as weighting, editing and imputation) ensure that the resulting data can be used to obtain approximately unbiased estimates for the population of interest. (We will offer a more detailed review of the survey process in Section 6.2.3.) However, how to properly account for these particularities within the framework of DP is currently poorly understood (see Reiter (2019) and Drechsler (2023) for an in-depth discussion of the challenges that will arise in this context). Gaining a better understanding is especially critical as the Census Bureau has publicly committed to adopting DP for all its data products (US Census Bureau, 2018) – a resolution that has been recently reaffirmed in US Census Bureau (2022c). (In fact, the Census Bureau only recommitted to adopting “formal privacy”; however we are not aware of any other formal privacy frameworks for statistical data apart for DP.) In the same 2022 press release, the Census Bureau concluded that “the science does not yet exist” to implement DP for their flagship survey – the ACS – highlighting

the need for additional research in this area.

We are aware of only a few papers that address DP in the survey context and, moreover, all these papers only focus on specific aspects of this process. [Lin et al. \(2023\)](#) study how to estimate the mean of a binary variable under DP assuming stratified sampling using proportional allocation and simple random sampling within the strata. [Bun et al. \(2022\)](#) investigate if the complex sampling designs commonly used in the survey context can offer increased privacy protection building on previous results showing that simple sampling procedures such as simple random sampling or Poisson sampling will amplify the privacy protection ([Balle et al., 2020](#)). We will summarize their findings in Section 6.4.1. Finally, in some preliminary work, [Das et al. \(2022\)](#) study the effects of imputation. They find that if DP is only considered when analyzing the imputed data, the required privacy loss budget can increase linearly with the number of missing cases. They also show that this problem can be avoided – at least for certain imputation schemes – if DP is already considered during imputation.

This paper aims to establish a framework for DP in the survey context by discussing the implications of (for example) whether the privacy guarantees should hold only for the sampled units or the entire population. We identify ten settings that vary in their assumptions about the data at different levels (the responding sample, the selected sample, the sampling frame, and the target population). Building on the framework introduced in [Bailie et al. \(2025b\)](#), we formalize the DP flavors for these settings and discuss their implications on both data utility and privacy.

6.2 BACKGROUND

6.2.1 NOTATION

We typically denote sets by upper-case calligraphic letters (for example, \mathcal{S} , \mathcal{T} or \mathcal{D}) and sets of sets by upper-case script letters (for example, \mathscr{D} or \mathscr{F}). Datasets are denoted by fraktur lower-case letters (for example, \mathfrak{d} , \mathfrak{d}' , \mathfrak{p} , \mathfrak{f} or \mathfrak{s}) when they are not stochastic, and by upper-case letters (for example, \mathfrak{D} , \mathfrak{D}' , \mathfrak{P} , \mathfrak{F} or \mathfrak{S}) when they are random variables. In general, we follow the convention that lower-case letters denote realizations of the corresponding upper-case random variable. However, we use the sans-serif superscript R to denote a random set (for example, \mathcal{S}^{R}); an upper-case calligraphic letter without this superscript often denotes a realization of the corresponding random set (for example, \mathcal{S} denotes a realization of the random set \mathcal{S}^{R}).

A record r is a set of attributes and a dataset \mathfrak{d} is a set of records. Every record r is associated with a unit, which we denote by $u(r)$. The units of a dataset \mathfrak{d} are given by the set $\mathcal{U}(\mathfrak{d}) = \{u(r) \mid r \in \mathfrak{d}\}$. We assume throughout that every unit is associated with at most one record in any given dataset, although a unit will often have multiple records spread across different datasets. The unique record in the dataset \mathfrak{d} associated with unit $i \in \mathcal{U}(\mathfrak{d})$ is denoted by \mathfrak{d}_i .

As an example, a unit could be a person, and the attributes of a record could describe some of the characteristics of that person, such as their age, income and occupation, as well as some identifiers, such as their name and address. Alternatively, a unit could be a company, and a record associated with a company could detail some business characteristics of that company. Less frequently, a unit may represent a group of people, or a population – in this way, we can encode population-level information in a dataset. Occasionally, it will be important to distinguish between the unit – which is an abstraction – and the real-world entity

that is represented by the unit. Beyond their philosophical differences, discrepancies between a unit's data and the corresponding entity's characteristics can arise due to measurement error, non-response or imputation. Moreover, there can be multiple units which represent the same entity. Such over-counting can occur when, for example, units are constructed from a register of addresses (or phone numbers, identification numbers, etc.) because a single entity can have multiple addresses. Duplication is a common problem in surveying, particularly in the context of business statistics, and – as we will see – poses a complication for DP.

An attribute is a value of a variable. More exactly, an attribute of a unit i is the value of a variable that is taken by i . (For example, an attribute could be the value 40 and the associated variable could be Age (in years). This would signify that unit i has an age of 40 years.) Therefore, a record r is uniquely specified by its unit $u(r)$ and the variables associated to its attributes. Denote the set of the variables in a record r by $\mathcal{V}(r)$ and the variables in a dataset \mathfrak{d} by $\mathcal{V}(\mathfrak{d}) = \bigcup_{i \in \mathcal{U}(\mathfrak{d})} \mathcal{V}(\mathfrak{d}_i)$. Although we do not require it, usually every record in a dataset has the same variables: $\mathcal{V}(\mathfrak{d}) = \mathcal{V}(\mathfrak{d}_i)$ for all $i \in \mathcal{U}(\mathfrak{d})$.

Given a set of units \mathcal{U} and a set of variables \mathcal{V} , let $\mathfrak{d}(\mathcal{U}, \mathcal{V})$ denote the dataset $\{r \mid u(r) \in \mathcal{U}, \mathcal{V}(r) = \mathcal{V}\}$. This dataset $\mathfrak{d}(\mathcal{U}, \mathcal{V})$ is well-defined because every record is determined by its variables and its unit. Given a variable x and a unit i , let x_i denote i 's value of the variable x . We can re-express $\mathfrak{d}(\mathcal{U}, \mathcal{V})$ as

$$\mathfrak{d}(\mathcal{U}, \mathcal{V}) = \left\{ \{x_i \mid x \in \mathcal{V}\} \mid i \in \mathcal{U} \right\}.$$

6.2.2 DIFFERENTIAL PRIVACY

DP studies data-release mechanisms – functions T which take as input a dataset \mathfrak{d} and a random seed ω , and output a stochastic summary $T(\mathfrak{d}, \omega)$ of \mathfrak{d} .

Definition 6.2.1. A *data-release mechanism* is a function $T : \mathcal{D}_0 \times \Omega \rightarrow \mathcal{T}$ where

- \mathcal{D}_0 is the data space, the set of all theoretically-possible datasets \mathfrak{d} ;
- Ω is the probability space of the seed ω with σ -algebra \mathcal{F}_Ω and probability P ;
- \mathcal{T} is equipped with a σ -algebra $\mathcal{F}_\mathcal{T}$; and
- $T(\mathfrak{d}, \cdot)$ is measurable for all $\mathfrak{d} \in \mathcal{D}_0$.

(See [Bailie et al. \(2025b\)](#) for a slightly more general definition and for additional context.)

Intuitively speaking, \mathfrak{d} is the data that is considered confidential and hence must not be disclosed by the summary $T(\mathfrak{d}, \omega)$. DP measures how the probabilistic noise induced by the seed ω masks this input dataset \mathfrak{d} .

We emphasize that, in order for T to be well-defined (as a function $\mathcal{D}_0 \times \Omega \rightarrow \mathcal{T}$), its input \mathfrak{d} must contain all the data which has a non-zero probability (with respect to P) of being used by T . That is to say, the output $T(\mathfrak{d}, \omega)$ can only depend on data which is in \mathfrak{d} , or data that is generated from \mathfrak{d} and ω , but not on other data. While it may seem we are belaboring an obvious point – of course, by definition $T(\mathfrak{d}, \omega)$ cannot be a function of anything but \mathfrak{d} and ω – the input dataset \mathfrak{d} is surprisingly slippery to specify in the context of surveying, as we illustrate with the following simplistic example.

Example 6.2.2. Suppose that a government agency is conducting a survey on the health of people in Massachusetts. The agency has a list of Massachusettsans (a *frame* \mathfrak{f} , see Subsection 6.2.3 below) from which they will randomly select a sample of individuals. They will then collect data \mathfrak{S} on some of the health characteristics of the sampled individuals (e.g. blood pressure, heart rate, etc.) and publish some aggregate statistics based on these collected data.

As we will expand upon later in this article, the agency may decide to include the sampling procedure in their data-release mechanism T , since this can potentially increase the efficiency of the privacy-utility tradeoff (see Subsection 6.4.1). In this case, T takes as input the frame \mathbf{f} ; it “performs” the sampling and data collection steps outlined above; and then it calculates and outputs the aggregate statistics. There are two options for how T can “collect” the data \mathfrak{S} . The first option is that the data \mathfrak{S} is generated (or modelled, depending on one’s perspective) within the data-release mechanism T – i.e. \mathfrak{S} is a function of T ’s input data and seed. The second option is that the data \mathfrak{S} is itself included as part of T ’s input data.

We will see in Section 6.5 that the DP guarantee does not necessarily apply to data generated within a DP mechanism – it only applies to the mechanism’s input data.² Hence, the first option is not appropriate if we want to guarantee the privacy protection of the sampled individuals’ health characteristics. We must therefore resort to the second option and include the data \mathfrak{S} as input to T . However, we do not know a-priori which individuals will be sampled. Since any individual in the frame \mathbf{f} has a non-zero probability of being sampled, any of the records in $\mathfrak{d}(\mathcal{U}(\mathbf{f}), \mathcal{V}(\mathfrak{S}))$ may appear in the sample data \mathfrak{S} . As such, all of these records must be included as input – that is, T requires as input $\mathbf{f}^* = \mathfrak{d}(\mathcal{U}(\mathbf{f}), \mathcal{V}(\mathbf{f}) \cup \mathcal{V}(\mathfrak{S}))$.

We refer to \mathbf{f}^* as the *augmented frame*, since it includes all the variables that are collected in the survey as well as all the frame variables. In the context of survey sampling, \mathbf{f}^* is never observed. Yet, it must nevertheless serve as input to any data-release mechanism T , whenever T includes a sampling step and we wish to provide the sample data with a DP guarantee. The data in $\mathfrak{d}(\mathcal{U}(\mathbf{f}), \mathcal{V}(\mathfrak{S}))$ are not available to the government agency at the time it starts its data collection. Rather, $\mathfrak{d}(\mathcal{U}(\mathbf{f}), \mathcal{V}(\mathfrak{S}))$ is the ‘theoretical’ dataset from which the agency collects the survey data.

While the input \mathbf{f}^* described above can be observed if the agency surveyed all units in the frame, in some

²This discussion is still missing at this point, but will be included in the final version of the paper.

situations it is not even theoretically possible to observe the input to a DP data-release mechanism. It is not uncommon that a survey includes a minor intervention as part of its data collection. For example, the Massachusetts health survey could require administering an oral glucose load as part of a glucose tolerance test in the diagnosis of diabetes (Phillips, 2012), or it could direct the survey respondent to exercise on a stationary bike as part of a cardiac stress test (Bruce and McDonough, 1969). Alternatively, in the context of a medical trial, the sampled individuals could be randomly assigned to receive a treatment or a placebo. In these cases, the data we wish to protect – the outcomes of these health tests – are only realized during the data collection process. When this data collection process is included within the data-release mechanism – as must necessarily be the case when the data-release mechanism T includes the sampling step of the survey – these data cannot possibly be included as input into T , because they do not even exist at the time the data-release mechanism begins! (One may argue that the process of any data collection or measurement – such as checking blood pressure – is itself an intervention and the collected data only come into existence at the point of collection. Under this perspective, the following remarks apply to all data.)

In such cases, the input data must necessarily include the potential outcome of each of the possible interventions (or treatments). To those familiar with causal inference, the dataset of these potential outcomes is known as the *science table* (Rubin, 2005). The science table is never fully observable because the potential outcome under a counterfactual treatment is always unknown. Yet, if we want to protect the outcome under the non-counterfactual treatment – which is unknown at the start of T – we must include it as input to T , and we can only ensure it is included as input if we include all the potential outcomes as input.

We end this example by noting that, if T does not include a sampling step, then T need not include the

data collection step either. As such, T 's input data is simply the collected data, without concern to the counterfactual potential outcomes.

It is convenient to think of a data-release mechanism as a function $\mathfrak{d} \mapsto P_{\mathfrak{d}}(T \in \cdot)$. Here the probability distribution $P_{\mathfrak{d}}(T \in \cdot)$ of the summary $T(\mathfrak{d}, \omega)$ is the push-forward measure induced by the distribution P of the random seed $\omega \in \Omega$, taking \mathfrak{d} as fixed:

$$P_{\mathfrak{d}}(T \in E) := P(\{\omega \in \Omega : T(\mathfrak{d}, \omega) \in E\}),$$

where $E \in \mathcal{F}_{\mathcal{T}}$ is any measurable subset of the output space \mathcal{T} . DP is the condition that the data-release mechanism is Lipschitz continuous – i.e. that the distance $D_{\text{Pr}}(P_{\mathfrak{d}}, P_{\mathfrak{d}'})$ between outputs $P_{\mathfrak{d}}$ and $P_{\mathfrak{d}'}$ is at most a multiplicative factor of the distance $d_{\mathcal{D}_0}(\mathfrak{d}, \mathfrak{d}')$ between the corresponding inputs \mathfrak{d} and \mathfrak{d}' .

Example 6.2.3. For pure ε -DP, as defined in [Dwork et al. \(2006b\)](#), the multiplicative factor is ε ; the distance between inputs \mathfrak{d} and \mathfrak{d}' is the Hamming distance; and the distance between outputs $P_{\mathfrak{d}}$ and $P_{\mathfrak{d}'}$ is the multiplicative distance:

$$D_{\text{MULT}}(P_{\mathfrak{d}}, P_{\mathfrak{d}'}) = \sup_{E \in \mathcal{F}_{\mathcal{T}}} \left| \ln \frac{P_{\mathfrak{d}}(T \in E)}{P_{\mathfrak{d}'}(T \in E)} \right|,$$

(For readers that are familiar with the definition of pure ε -DP in terms of neighboring datasets \mathfrak{d} and \mathfrak{d}' , the Lipschitz condition for non-neighbors is implied by group privacy. Hence, the neighbor definition of pure ε -DP is the equivalent to the above definition.)

For approximate (ε, δ) -DP ([Dwork et al., 2006a](#)), the multiplicative factor is again ε ; the distance be-

tween inputs is given by

$$d_{\mathcal{D}_0}^{\text{neighbors}}(\mathfrak{d}, \mathfrak{d}') = \begin{cases} 0 & \text{if } \mathfrak{d} = \mathfrak{d}', \\ 1 & \text{if } \mathfrak{d} \text{ and } \mathfrak{d}' \text{ are neighbors,} \\ \infty & \text{otherwise;} \end{cases}$$

and the distance between outputs is given by

$$D_{\text{MULT}}^{\delta}(\mathbb{P}_{\mathfrak{d}}, \mathbb{P}_{\mathfrak{d}'}) = \sup_{E \in \mathcal{F}_T} \left\{ \ln \frac{[\mathbb{P}_{\mathfrak{d}}(T \in E) - \delta]^+}{\mathbb{P}_{\mathfrak{d}'}(T \in E)}, \ln \frac{[\mathbb{P}_{\mathfrak{d}'}(T \in E) - \delta]^+}{\mathbb{P}_{\mathfrak{d}}(T \in E)}, 0 \right\},$$

(where $[x]^+ = \max\{x, 0\}$). Note that $d_{\mathcal{D}_0}^{\text{neighbors}}$ and D_{MULT}^{δ} are not distances in the mathematical sense of a metric; we will instead refer to them as *premetrics* from herein. Since D_{MULT}^{δ} does not satisfy the triangle inequality, approximate (ε, δ) -DP's group privacy budget does not increase linearly with the group size; hence we cannot replace $d_{\mathcal{D}_0}^{\text{neighbors}}$ with the Hamming distance, as we did for pure ε -DP.

By definition, a data-release mechanism T satisfies DP if it is Lipschitz continuous. There are different *flavors* (i.e. types or versions) of DP; each of these flavors correspond to different ways to specify continuity. For our purposes, there are four components to the specification of Lipschitz continuity. Most obviously, there are the premetrics $d_{\mathcal{D}_0}(\mathfrak{d}, \mathfrak{d}')$ and $D_{\text{Pr}}(\mathbb{P}_{\mathfrak{d}}, \mathbb{P}_{\mathfrak{d}'})$. These premetrics measure the ‘distance’ between any two inputs \mathfrak{d} and \mathfrak{d}' , or between any two output probabilities $\mathbb{P}_{\mathfrak{d}}$ and $\mathbb{P}_{\mathfrak{d}'}$. Secondly, there is the domain \mathcal{D}_0 of the data-release mechanism, which – as we shall see – serves as the parameter space of the attacker’s inferential model.³ Finally, there is the data multiverse \mathcal{D} , which allows the data custodian to restrict the Lipschitz continuity condition to certain pairs of inputs – as is often desirable in practice. For example, we may only want to compare samples drawn from the same population. This restriction is achieved by

³This discussion is still missing at this point, but will be included in the final version of the paper.

specifying the data multiverse \mathcal{D} .

Definition 6.2.4 (Bailie et al. (2025b)). A *differential privacy flavor* is a quadruple $(\mathcal{D}_0, \mathcal{D}, d_{\mathcal{D}_0}, D_{\text{Pr}})$

where:

1. The *domain* \mathcal{D}_0 is the *data space* – the set of all (theoretically-possible) input datasets.
2. The *multiverse* $\mathcal{D} \subset 2^{\mathcal{D}_0}$ is a set of *universes*, which are denoted by \mathcal{D} or \mathcal{D}' .
3. The *input premetric* $d_{\mathcal{D}_0}$ is a premetric on \mathcal{D}_0 – i.e. a function $\mathcal{D}_0 \times \mathcal{D}_0 \rightarrow \mathbb{R}^{\geq 0}$ such that $d_{\mathcal{D}_0}(\mathfrak{d}, \mathfrak{d}) = 0$ for all $\mathfrak{d} \in \mathcal{D}_0$.
4. The *output premetric* D_{Pr} is a premetric on the space of all probability distributions \mathcal{P} – i.e. a function $\mathcal{P} \times \mathcal{P} \rightarrow \mathbb{R}^{\geq 0}$ of probabilities $P, Q \in \mathcal{P}$ such that
 - $D_{\text{Pr}}(P, P) = 0$ for all $P \in \mathcal{P}$; and
 - $D_{\text{Pr}}(P, Q) = \infty$ for probabilities P, Q which live on different measurable spaces.

Once we have specified the four components for Lipschitz continuity via a DP flavor, we also need to specify the multiplicative constant (known as the Lipschitz constant) which controls the rate between input and output variations. Together, choices for these five components are called a DP specification:

Definition 6.2.5. A *differential privacy specification* is a quintuple $(\mathcal{D}_0, \mathcal{D}, d_{\mathcal{D}_0}, D_{\text{Pr}}, \varepsilon_{\mathcal{D}})$ consisting of a DP flavor $(\mathcal{D}_0, \mathcal{D}, d_{\mathcal{D}_0}, D_{\text{Pr}})$ and a privacy-loss budget $\varepsilon_{\mathcal{D}} : \mathcal{D} \rightarrow \mathbb{R}^{\geq 0}$. We denote a DP specification by $\varepsilon_{\mathcal{D}}\text{-DP}(\mathcal{X}, \mathcal{D}, d_{\mathcal{X}}, D_{\text{Pr}})$.

A data-release mechanism $T : \mathcal{D}_0 \times \Omega \rightarrow \mathcal{T}$ satisfies the DP specification $\varepsilon_{\mathcal{D}}\text{-DP}(\mathcal{X}, \mathcal{D}, d_{\mathcal{X}}, D_{\text{Pr}})$ if, for all data universes $\mathcal{D} \in \mathcal{D}$, and all $\mathfrak{d}, \mathfrak{d}' \in \mathcal{D}$,

$$D_{\text{Pr}}[P_{\mathfrak{d}}(T \in \cdot), P_{\mathfrak{d}'}(T \in \cdot)] \leq \varepsilon_{\mathcal{D}} d_{\mathcal{D}_0}(\mathfrak{d}, \mathfrak{d}'). \quad (6.1)$$

Let $\mathcal{M}(\mathcal{D}_0, \mathcal{D}, d_{\mathcal{D}_0}, D_{\text{Pr}}, \varepsilon_{\mathcal{D}})$ denote the set of data-release mechanisms which satisfy the DP specification $\varepsilon_{\mathcal{D}}\text{-DP}(\mathcal{X}, \mathcal{D}, d_{\mathcal{X}}, D_{\text{Pr}})$.

For the purposes of understanding DP in the context of survey sampling, the relevant components of a DP flavor $(\mathcal{D}_0, \mathcal{D}, d_{\mathcal{D}_0}, D_{\text{Pr}})$ are its domain \mathcal{D}_0 and its multiverse \mathcal{D} .

We need the following technical definition:

Definition 6.2.6. Let \mathcal{D}_0 be a domain and $\mathcal{D}, \mathcal{D}' \subset 2^{\mathcal{D}_0}$ be two multiverses of \mathcal{D}_0 . We say \mathcal{D}' is a *coarsening* of \mathcal{D} if, for all $\mathcal{D} \in \mathcal{D}$, there exists $\mathcal{D}' \in \mathcal{D}'$ with $\mathcal{D} \subset \mathcal{D}'$.

When \mathcal{D}' is a coarsening of \mathcal{D} we write $\mathcal{D} \leq \mathcal{D}'$. The following lemma justifies this notation by establishing that \mathcal{D} is a weaker condition than \mathcal{D}' if $\mathcal{D} \leq \mathcal{D}'$.

Lemma 6.2.7. Let \mathcal{D}_0 be a domain and $\mathcal{D}, \mathcal{D}' \subset 2^{\mathcal{D}_0}$ be multiverses such that $\mathcal{D} \leq \mathcal{D}'$. Then, for all budgets $\varepsilon_{\mathcal{D}'} : \mathcal{D}' \rightarrow \mathbb{R}^{\geq 0}$,

$$\mathcal{M}(\mathcal{D}_0, \mathcal{D}', d_{\mathcal{D}_0}, D_{\text{Pr}}, \varepsilon_{\mathcal{D}'}) \subset \mathcal{M}(\mathcal{D}_0, \mathcal{D}, d_{\mathcal{D}_0}, D_{\text{Pr}}, \varepsilon_{\mathcal{D}}),$$

where $\varepsilon_{\mathcal{D}} = \inf\{\varepsilon_{\mathcal{D}'} : \mathcal{D}' \in \mathcal{D}' \text{ s.t. } \mathcal{D} \subset \mathcal{D}'\}$.

Definition 6.2.8. Given a DP flavor $(\mathcal{D}_0, \mathcal{D}, d_{\mathcal{D}_0}, D_{\text{Pr}})$, the multiverse \mathcal{D} is *complete* if $d_{\mathcal{D}_0}(\mathfrak{d}, \mathfrak{d}') < \infty$ for all $\mathfrak{d}, \mathfrak{d}' \in \mathcal{D}$ and all $\mathcal{D} \in \mathcal{D}$.

Definition 6.2.9. Given a DP flavor $(\mathcal{D}_0, \mathcal{D}, d_{\mathcal{D}_0}, D_{\text{Pr}})$, two datasets $\mathfrak{d}, \mathfrak{d}' \in \mathcal{D}_0$ are *comparable* when 1) $\mathfrak{d} \neq \mathfrak{d}'$; 2) $d_{\mathcal{D}_0}(\mathfrak{d}, \mathfrak{d}') < \infty$ or $d_{\mathcal{D}_0}(\mathfrak{d}', \mathfrak{d}) < \infty$; and 3) there exists a data universe $\mathcal{D} \in \mathcal{D}$ such that $\mathfrak{d}, \mathfrak{d}' \in \mathcal{D}$.

Definition 6.2.10. Given a DP flavor $(\mathcal{D}_0, \mathcal{D}, d_{\mathcal{D}_0}, D_{\text{Pr}})$, denote the protection objects *connected* to $\mathfrak{d} \in \mathcal{D}_0$ by

$$[\mathfrak{d}] = \{\mathfrak{d}' \in \mathcal{D}_0 : d_{\mathcal{D}_0}(\mathfrak{d}, \mathfrak{d}') < \infty\}.$$

Then the *completion* $\overline{\mathcal{D}}$ of the data multiverse \mathcal{D} is defined as

$$\overline{\mathcal{D}} = \{\mathcal{D} \cap [\mathfrak{d}] : \mathcal{D} \in \mathcal{D}, \mathfrak{d} \in \mathcal{D}\}.$$

Lemma 6.2.11. *Let $(\mathcal{D}_0, \mathcal{D}, d_{\mathcal{D}_0}, D_{\text{Pr}})$ be a DP flavor where $d_{\mathcal{D}_0}$ is a metric. Then, the completion $\overline{\mathcal{D}}$ of \mathcal{D} is complete and, for all budgets $\varepsilon_{\mathcal{D}} : \mathcal{D} \rightarrow \mathbb{R}^{\geq 0}$,*

$$\mathcal{M}(\mathcal{D}_0, \overline{\mathcal{D}}, d_{\mathcal{D}_0}, D_{\text{Pr}}, \varepsilon_{\mathcal{D}'}) = \mathcal{M}(\mathcal{D}_0, \mathcal{D}, d_{\mathcal{D}_0}, D_{\text{Pr}}, \varepsilon_{\mathcal{D}}),$$

where

$$\varepsilon_{\mathcal{D}'} = \inf\{\varepsilon_{\mathcal{D}} : \mathcal{D} \in \mathcal{D} \text{ s.t. } \mathcal{D}' \subset \mathcal{D}\}.$$

6.2.3 SURVEY SAMPLING

Surveys are conducted to learn some characteristics of a well-defined population by collecting information from a random subset of this population. Most survey sampling processes rely on three key ingredients: the *target population* of interest; the *sampling frame* from which the random sample to be surveyed is drawn; and the *sampling design* for drawing this sample.

The target population (also known as the universe in some survey sampling texts, although we will not use this term to avoid confusion with the notion of a universe \mathcal{D} in a DP flavor) is the scope of the survey; it is the population the survey is aiming to learn about. It is typically defined conceptually, while the sampling frame \mathfrak{f} , on the other hand, is an existent register containing names (or other identifiers), contact information (postal or physical address, email, and/or telephone number) and possibly some basic demographic information of the survey units. The sampling frame \mathfrak{f} serves as the source from which the sample is drawn. For the discussions in the remainder of the paper it is important to clearly distinguish between the target population and the sampling frame \mathfrak{f} . While the sampling frame aims to cover all the units from

the target population, it might include units that are not part of the target population (*overcoverage*), and it might also miss units that should be included (*undercoverage*). To formalize the difference between the target population and the sampling frame, we suppose that the frame is not constructed from the target population data, but from a fixed dataset we term the *pseudo-population* dataset \mathfrak{p} . Typically, the frame is constructed from previous censuses' data, administrative records and canvassing. The pseudo-population dataset \mathfrak{p} is the collection of all such data, so that $\mathcal{U}(\mathfrak{f}) \subset \mathcal{U}(\mathfrak{p})$. By introducing the concept of the pseudo-population, we allow for undercoverage and overcoverage, as well as duplications in the frame (where a single unit in the target population corresponds to multiple units in the frame).

The sample is the set $\mathcal{U}^R(\mathfrak{S})$ of units of the sample dataset \mathfrak{S} . The sample is a random set whose distribution is given by the sampling design. The sampling design is defined as a probability measure $\tau_{\mathfrak{f}}$ on $2^{\mathcal{U}(\mathfrak{f})}$. The units $\mathcal{U}^R(\mathfrak{S})$ of the sample dataset \mathfrak{S} are a draw from $\tau_{\mathfrak{f}}$. That is, $\mathcal{U}^R(\mathfrak{S}) \sim \tau_{\mathfrak{f}}$. For each subset $\mathcal{S} \subset \mathcal{U}(\mathfrak{f})$, $\tau(\mathcal{S})$ is the probability that the realized sample $\mathcal{U}(\mathfrak{s})$ is \mathcal{S} . Sometimes the frame \mathfrak{f} contains basic demographic information on the survey units, which can be used to construct the sample selection probabilities $\tau(\mathcal{S})$. The sampling designs used in practice are often complex multi-stage designs, with different sampling strategies (e.g. cluster sampling, stratified sampling, probability-proportional-to-size (PPS) sampling) for each of the different stages. In determining the sample design $\tau_{\mathfrak{f}}$, the frame \mathfrak{f} is usually taken into consideration, which can complicate the deployment of DP.

To illustrate the relevance of this discussion, we look at the Current Population Survey (CPS) conducted by the Census Bureau for the Bureau of Labor Statistics (BLS). The *target population* of the CPS is the civilian noninstitutionalized population in the US, or, more exactly,

“all people residing in the 50 states [of the US] and the District of Columbia who are not confined to institutions such as nursing homes and prisons, and who are not on active duty in the US Armed Forces. Included are citizens of foreign countries who reside in the United

States but do not live on the premises of an embassy. The civilian noninstitutional population ages 16 and older is the base population group used for CPS statistics” (US Bureau of Labor Statistics, 2018a).

The survey uses two different *sampling frames*: one for households and one for group quarters. Both are derived from the master address file (MAF) of the Census Bureau: “The MAF is a national inventory of addresses that is continually updated by the U.S. Census Bureau to support its decennial programs and demographic surveys” (US Census Bureau, 2019a). The CPS uses a stratified two-stage *sampling design*. In the first stage, the population is divided into geographical clusters and one cluster is sampled within each stratum using PPS sampling. A small group of households is selected in the second stage using systematic sampling based on a list sorted by demographic composition and geographic proximity. (See Section 2.2 in US Census Bureau (2019a) for a full description of the sampling methodology.)

6.2.4 SURVEY WEIGHTS

A distinctive feature of survey data is that they typically contain survey weights. Survey weights are provided by statistical agencies as a convenient tool to account for the sampling design and additional data preparation steps such as nonresponse adjustments when analyzing the data. Because complex sampling designs are often used (as we described in the previous section) and because not all sampled units actually respond to the survey, the resulting dataset cannot be treated as a simple random sample from the target population. Most estimators need to be adjusted to take these complications into account. For example, the (unweighted) sample mean can no longer be treated as an unbiased estimator for the mean in the population if the probability of being included in the responding-sample varies between the units. Instead, it is typical to use weighted estimators, where individual data points are weighted according to their survey weights.

Survey weights are typically generated in three stages. In the first stage, *design weights* are generated that reflect the sampling design. In the second stage, *nonresponse adjustment weights* are used to account for different response propensities in different subgroups of the population. Finally, *calibration weights* try to correct for any deficiencies in the sampling frame and also help to reduce the variance of the final estimates.

The design weights w^D are defined as the inverse of the probability of selection: $w_i^D = 1/\pi_i$, where π_i is the probability that unit $i \in \mathcal{U}(\mathbf{f})$ is selected into the sample $\mathcal{U}^R(\mathfrak{S})$. Nonresponse adjustment weights try to adjust for potential biases that might arise due to unequal response propensities. The idea is to estimate the probability to respond for each unit. The nonresponse adjustment weights w^{NR} are calculated as the inverse of the estimated response probabilities p_i^R ; that is, $w_i^{NR} = 1/p_i^R$ for $i \in \mathcal{U}^R(\mathfrak{S})$. Note that the response probabilities can be used to compute the final probability to be included in the sample: $p_i^{(inc)} = \pi_i p_i^R$. Hence, the inverse of $p_i^{(inc)}$ can be used as a weight that accounts for both the complex sampling design and the nonresponse.

The final weighting step is commonly to calibrate the survey data to information that is known about the population of interest from other sources. For example, the total number of people living in the U.S. by age and gender might be known from the previous Census. Common calibration techniques are post-stratification, raking or the GREG estimator. Describing the details of these adjustment methods is beyond the scope of this paper (see Valliant et al. (2018) for further details). It suffices to note that all these methods can be reflected by adjusting the survey weights obtained from the previous two steps.

6.3 DP FLAVORS FOR SURVEY STATISTICS

As we have seen in Subsection 6.2.3, there are multiple phases in the creation of survey statistics: defining the target population, compiling the sampling frame, selecting the sample according to the sampling de-

sign, and collecting data from the responding units. (From herein, we use the term ‘target sample’ to refer to the sample of units selected by the sampling design from the sampling frame, in order to differentiate this sample with the responding sample – the set of units which were selected and responded.) The data output by each phase of this pipeline is fed into the subsequent phase as input. For example, data about the target population is used to compile the frame and data on the frame is used to select the sample.

The data custodian (e.g. the NSO) could plausibly start the data-release mechanism T at any point along this data pipeline. That is, the data-release mechanism could take as input the dataset corresponding to any of the various phases. Moreover, the custodian could also plausibly condition on previous phases in the data pipeline (taking their data as invariant). Thus, the data custodian is faced with two decisions: what should the protection domain \mathcal{D}_0 be? And what should the data multiverse \mathcal{D} be?

In this section, we formalize the various options for these two decisions in terms of their corresponding DP flavors. In Sections 6.4 and 6.5, we show why these two decisions are important by describing the consequences of each option on both data utility and privacy.

Definition 6.3.1. Let $\mathcal{D}_0^{\text{pp}}$ be the set of all possible pseudo-population datasets; $\mathcal{D}_0^{\text{fr}}$ the set of all possible frames (from all possible pseudo-populations); $\mathcal{D}_0^{\text{samp}}$ the set of all possible target sample datasets (from all possible frames); and $\mathcal{D}_0^{\text{resp}}$ the set of all possible responding sample datasets (from all possible target samples). We say that a DP flavor $(\mathcal{D}_0, \mathcal{D}, d_{\mathcal{D}_0}, D_{\text{Pr}})$ is *population-level* if $\mathcal{D}_0 = \mathcal{D}_0^{\text{pp}}$. The definitions of *frame-level*, *(target-)sample-level* and *responding-sample-level* DP flavors are analogous.

In the above definition, we have been deliberately vague in specifying $\mathcal{D}_0^{\text{pp}}$. The precise definition of the set $\mathcal{D}_0^{\text{pp}}$ depends on the data custodian’s assessment of what pseudo-populations are considered ‘possible’. In general, ‘possible’ should be interpreted liberally, so that this set $\mathcal{D}_0^{\text{pp}}$ is generously large. (See Section 6.5

for an explanation of why this matters and [Bailie et al. \(2025b\)](#) for a more extensive discussion.)

We can be more specific in the definition of $\mathcal{D}_0^{\text{fr}}$, since the construction of a frame is a real-world process undertaken by an NSO (although in practice this process is often messy, complex and hard to precisely describe). This process takes as input a pseudo-population $\mathbf{p} \in \mathcal{D}_0^{\text{pp}}$ and outputs a frame for that population. Then $\mathcal{D}_0^{\text{fr}}$ is the set of all outputs from this process, across all possible pseudo-populations $\mathbf{p} \in \mathcal{D}_0^{\text{pp}}$.

When defining the set $\mathcal{D}_0^{\text{samp}}$ of all possible samples, we assume that there is a given sampling design $\tau_{\mathbf{f}}$ and we only consider those sample datasets \mathbf{s} with non-zero probability $\tau_{\mathbf{f}}(\mathcal{U}(\mathbf{s})) > 0$. However, as is frequently the case, the sampling design $\tau_{\mathbf{f}}$ can depend on the realized frame \mathbf{f} . (For example, the stratum sample sizes are part of a stratified sampling design, and these sizes are partially based on the sizes of the strata in the frame \mathbf{f} .) Thus,

$$\mathcal{D}_0^{\text{samp}} = \{\mathbf{s} : \tau_{\mathbf{f}}(\mathcal{U}(\mathbf{s})) > 0, \mathbf{f} \in \mathcal{D}_0^{\text{fr}}\}.$$

Definition 6.3.2 (Primitive data multiverses). Define the primitive data multiverses:

1. $\mathcal{D}_{\text{fr}|\text{pp}} = \{\mathcal{D}_{\mathbf{p}} : \mathbf{p} \in \mathcal{D}_0^{\text{pp}}\}$, where $\mathcal{D}_{\mathbf{p}}$ is the set of all possible frames constructed from the pseudo-population $\mathbf{p} \in \mathcal{D}_0^{\text{pp}}$;
2. $\mathcal{D}_{\text{samp}|\text{pp}} = \{\mathcal{D}_{\mathbf{p}} : \mathbf{p} \in \mathcal{D}_0^{\text{pp}}\}$, where $\mathcal{D}_{\mathbf{p}}$ is the set of all possible target sample datasets drawn from the pseudo-population $\mathbf{p} \in \mathcal{D}_0^{\text{pp}}$;

$$\mathcal{D}_{\mathbf{p}} = \{\mathbf{s} : \tau_{\mathbf{f}}(\mathcal{U}(\mathbf{s})) > 0, \mathbf{f} \text{ is a possible frame constructed from the pseudo-population } \mathbf{p}\}.$$

3. $\mathcal{D}_{\text{samp}|\text{fr}} = \{\mathcal{D}_{\mathbf{f}} : \mathbf{f} \in \mathcal{D}_0^{\text{fr}}\}$, where $\mathcal{D}_{\mathbf{f}}$ is the set of all possible target samples drawn from the frame $\mathbf{f} \in \mathcal{D}_0^{\text{fr}}$:

$$\mathcal{D}_{\mathbf{f}} = \{\mathbf{s} : \tau_{\mathbf{f}}(\mathcal{U}(\mathbf{s})) > 0\}.$$

4. The data multiverses $\mathcal{D}_{\text{resp}|\text{pp}}$, $\mathcal{D}_{\text{resp}|\text{fr}}$ and $\mathcal{D}_{\text{resp}|\text{samp}}$ can be defined analogously, as the set of data universes $\mathcal{D}_{\mathbf{d}}$, with $\mathcal{D}_{\mathbf{d}}$ the set of all possible responding samples drawn from, respectively, the population, frame, or target sample \mathbf{d} .

Definition 6.3.3 (Population-, frame- and sample-invariance). A DP flavor $(\mathcal{D}_0, \mathcal{D}, d_{\mathcal{D}_0}, D_{Pr})$ is:

1. *population-invariant* if $\overline{\mathcal{D}} \leq \mathcal{D}_{|pp}$, where
 - for frame-level flavors: $\mathcal{D}_{|pp} = \mathcal{D}_{fr|pp}$,
 - for sample-level flavors: $\mathcal{D}_{|pp} = \mathcal{D}_{smp|pp}$, and
 - for responding-sample-level flavors: $\mathcal{D}_{|pp} = \mathcal{D}_{resp|pp}$;
2. *frame-invariant* if $\overline{\mathcal{D}} \leq \mathcal{D}_{|fr}$, where
 - for sample-level flavors: $\mathcal{D}_{|fr} = \mathcal{D}_{smp|fr}$, and
 - for responding-sample-level flavors: $\mathcal{D}_{|fr} = \mathcal{D}_{resp|fr}$;
3. *sample-invariant* if $\overline{\mathcal{D}} \leq \mathcal{D}_{resp|smp}$ (for responding-sample-level flavors).

The intuition behind these definitions is very simple. The idea is to restrict the comparable datasets (Definition 6.2.9). Population-invariance means that comparable frames (or samples or responding samples) must be from the same pseudo-population. (That is, a pair of frames are comparable only if they are constructed from the same pseudo-population.) Analogously, frame-invariance means that comparable sample datasets must be drawn from the same frame.

If a DP flavor is not population-invariant (resp. frame-invariant or sample-invariant), then we say it is *population-agnostic* (resp. *frame-agnostic* or *sample-agnostic*). Frame-agnosticism implies that there are two comparable samples which are drawn from different frames.

Because invariance at one level implies invariance at previous data pipeline phases, we identify ten settings (which together exhaust the potential options for where the DP mechanism starts and which phases are taken as invariant): one setting for population-level flavors; two for frame-level flavors; three for sample-level flavors; and four for responding-sample-level flavors (see Table 6.1 for illustration.)

$\mathcal{D}_0^{\text{pp}}$	population agnostic			
$\mathcal{D}_0^{\text{fr}}$	population agnostic	population invariant		
$\mathcal{D}_0^{\text{samp}}$	+frame agnostic	+frame agnostic	+frame invariant	
$\mathcal{D}_0^{\text{resp}}$	+sample agnostic	+sample agnostic	+sample agnostic	+sample invariant

Table 6.1: Overview of the possible settings for the different levels.

6.4 UTILITY CONSIDERATIONS

In this section we consider the possible implications of the different DP flavors on the achievable level of accuracy of the noisy outcome given a desired level of privacy (expressed by fixing the privacy parameters). Two components are relevant when evaluating the accuracy for DP estimates from survey data: the privacy amplification effects from sampling, which imply that less noise needs to be infused to achieve a given privacy level and the increased sensitivity of the weighted estimator (where weights are included to account for the sampling design, nonresponse, and potentially for other data deficiencies such as over- or undercoverage of the sampling frame), which typically implies that more noise is required. We discuss the effects of the different flavors on both components in the following chapters.

6.4.1 PRIVACY AMPLIFICATION VIA SAMPLING

Previous research has shown that simple sampling designs offer privacy amplification, that is, the privacy offered when running a DP algorithm on a random subset of the population is higher than if the same algorithm with the same privacy parameters is run on the full population. [Balle et al. \(2018\)](#) proof the following theorem for simple random sampling with replacement (they also obtain similar results for Poisson sampling and simple random sampling without replacement):

Theorem 6.4.1 ([Balle et al. \(2018\)](#)). *Let \mathcal{C} be a sampling scheme that uniformly randomly samples n values out of N possible values without replacement. Given an (ϵ, δ) -bounded differentially private mechanism \mathcal{M} ,*

we have that $\mathcal{M} \circ \mathcal{C}$ is (ε', δ') -bounded differentially private for $\varepsilon' = \log(1 + \frac{n}{N}[\varepsilon^\varepsilon - 1])$ and $\delta' = \frac{n}{N}\delta$.

In this theorem bounded differential privacy refers to the scenario in which neighboring datasets are obtained by changing the values of one record in the data while keeping the size of the data fixed. Note that for small ε and small sampling rates this implies that $\varepsilon' \approx n/N\varepsilon$, i.e., the amplification is proportional to the sampling rate. Based on these results [Bun et al. \(2022\)](#) studied to what extent privacy amplification can also be achieved for the more complex sampling designs commonly used at statistical agencies. Their findings can be summarized as follows:

- Cluster sampling using simple random sampling without replacement to draw the clusters offers negligible amplification in practice except for small ε (less than 0.5) and very small cluster sizes (less than 15 units).
- With minor adjustments, stratified sampling using proportional allocation can provide privacy amplification.
- Data dependent allocation functions such as Neyman allocation for stratified sampling will likely result in privacy degradation (the effects will depend on the sensitivity of the allocation function).
- With PPS sampling at the individual level, the privacy amplification will linearly depend on the maximum probability of inclusion (for small ε).
- Systematic sampling will only offer amplification if the ordering of the population is truly random. In all other cases, systematic sampling will suffer from the same effects as cluster sampling leading to no amplification (assuming the ordering is known to the attacker).

In practice this implies that for the multi-stage sampling designs that typically start with (multiple stages of) stratified cluster sampling, amplification effects can generally only be expected from those stages at which individual units are selected (typically the last stage of selection).

6.4.2 PRIVACY AMPLIFICATION FOR DIFFERENT DP FLAVORS

Before discussing the implications of the DP flavors introduced in Section 6.3, it is important to consider at which stages of the data production pipeline amplification effects could occur. Conceptually, three different sampling steps can be defined when moving from the population to the responding sample. The most obvious step (and the only one that is fully controlled) is the selection of the target sample from the sampling frame. However, if nonresponse is treated as a stochastic process (as is commonly done in the survey literature), moving from the target to the responding sample can be interpreted as another sampling step. The same is true when moving from the pseudo-population to the sampling frame if we assume that each unit in the pseudo-population has a certain probability to be included in the frame. Still, the amplification effects of these two steps are difficult to take into account in practice as the inclusion probabilities are unknown and would need to be estimated. Errors when modeling these probabilities would lead to invalid statements regarding the amplification effects. Besides, the amplification effects when moving from the pseudo-population to the sampling frame will typically be negligible given that the probability to be included in the frame should be well above 90% for high quality frames.

Considering the DP flavors, we can distinguish four scenarios: If the responding sample dataset is given as input, the DP mechanism can only be applied at the responding sample level. This scenario boils down to the standard setting considered in most DP papers. There is no (sub)sampling step within the data release mechanism T and thus there is no amplification effect. Interestingly, this scenario offers the same

privacy guarantees as the more restrictive assumption that the attacker knows who participated in the survey. In all other scenarios, privacy is amplified through the (sub)sampling process.

In the second scenario, the DP flavor is at the target sample-level. In this scenario, amplification can only arise from the subsampling step when moving from the target sample to the responding sample. As response rates are often less than 20% in practice, this subsampling might offer some privacy amplification. However, as mentioned earlier, quantifying this effect will be difficult in practice as response probabilities are unknown and will likely differ between the units. In the third scenario, the (augmented) frame \mathbf{f}^* is taken as input to the data-release mechanism. This scenario will offer privacy amplification as discussed in [Bun et al. \(2022\)](#) in addition to the theoretical amplification offered from nonresponse. Finally, if the DP flavor has domain $\mathcal{D}_0^{\text{PP}}$, a third layer of amplification is possible by moving from the pseudo-population to the sampling frame. As discussed above, this layer will typically be negligible for sampling frames commonly used in practice.

6.4.3 WEIGHTING

Using weighted estimators generally increases the amount of noise that needs to be added to achieve a desired level of privacy protection. This is because the sensitivity of the result, i.e., the maximum possible change in the result when changing a single record, increases when incorporating the survey weights. To illustrate, we can consider the simple example of a counting query. A counting query simply counts the number of units in a database that satisfy a given set of conditions, for example, the total number of unemployed men between 30 and 40. Counting queries are attractive under DP as they have low sensitivity and thus require limited amounts of noise to achieve DP (as the noise scales with the sensitivity of the query). Under unbounded DP (i.e., defining neighboring datasets by adding or removing one record) the

sensitivity of a counting query is 1.

In the survey literature a counting query is called a total and the most convenient way to estimate a total for complex sampling designs is to use the Horvitz-Thompson estimator (Horvitz and Thompson, 1952), which provides approximately unbiased estimates for most sampling designs. The Horvitz-Thompson estimator for a total is given as $\hat{t}_x = \sum_{i \in \mathcal{U}^R(\mathfrak{S})} w_i x_i$, where \hat{t}_x is the estimated total in the population for the target variable x and w_i is the survey weight for unit i . In our example, x_i is a binary indicator which equals 1 if unit i satisfies the conditions of interest (i.e. unit i is unemployed, male, and between 30 and 40 years old) and is zero otherwise. Using the Horvitz-Thompson estimator, the L_1 -sensitivity increases to $\max(w_i)$ (where the maximum is taken either over the records in the sample (under target sample invariance), the frame (under frame invariance) or over the entire population (under population invariance)).

Since the amount of noise that is required typically scales with the sensitivity of the output, this implies that much more noise needs to be added when trying to protect a weighted survey estimate. However, the considerations so far assume that the weights can be considered fixed. This assumption is never justified for the final survey weights. This is because the nonresponse adjustments and calibration steps rely on models that are estimated from the data. Changing one record in the data will change these models and thus the weights. How to account for this variability in the final weights when computing the sensitivity of a survey weighted estimate has not been addressed in the DP literature so far.

But even if we only consider the design weights, the assumption of constant weights is only justified, if changing one record in the database does not change the probability of inclusion for any of the records in the pseudo-population. Whether this is a realistic assumption will depend on the DP flavor to be considered but also on the properties of the sampling design.

In general, the design weights can only be treated as fixed under the frame-invariant or target sample

DP Setting	Effects on Design Weights
Target sample invariance	Can be treated as fixed
Frame invariance	Can be treated as fixed
Population invariance	Sensitivity needs to be considered
Population agnostic	Sensitivity needs to be considered

Table 6.2: Overview of the implications on the design weights for different types of invariance. We note that the final weights that also account for nonresponse can never be treated as fixed.

invariant scenario. In all other scenarios the weights will typically change. How much the weights will change will depend on the sensitivity of the sampling design, which in turn depends on how data dependent the sampling design is. To illustrate, data dependence will be small for single stage cluster sampling designs especially if the clusters are selected using simple random sampling (such a design is used for example for the German Microcensus). For such a design, the probability of selection does not change over neighboring frames (as long as the definition of the clusters does not change). On the other hand, PPS sampling will generally be highly data dependent as the probability of selection directly depends on some features of the data. This will be less problematic if PPS sampling is used to select the clusters as the probability of selection will only depend on the size of the clusters and these sizes will only change by one record over neighboring databases. However, if PPS sampling is used to select individual units, the probabilities of selection can change arbitrarily over neighboring datasets. Thus, for these designs the sensitivity of the final estimate might increase considerably and it seems difficult to correctly quantify this sensitivity in practice.

Tables 6.2 and 6.3 summarize the implications of the different DP flavors considered in this paper. Together they highlight the inherent trade-off between the various flavors of DP for survey estimators. For example, considering the frame as invariant implies that the DP flavor is at the target- or responding-sample

DP Setting	Amplification from
Responding-sample level $\mathcal{D}_0^{\text{resp}}$	–
Target-sample level $\mathcal{D}_0^{\text{samp}}$	NR
Frame level $\mathcal{D}_0^{\text{fr}}$	NR&S
Population level $\mathcal{D}_0^{\text{pp}}$	FR&S&NR

Table 6.3: Overview of the implications on privacy amplification for different levels of DP. (The abbreviations are NR=nonresponse,S=sampling,FR=frame).

level and hence no utility improvements through amplification by sampling can be achieved. On the other hand, frame invariance allows treating the weights as fixed, which will generally reduce the sensitivity of the final estimates and thus the noise that needs to be added to ensure privacy. For the other flavors, utility improvements could be achieved through privacy amplification, but this benefit comes at the cost that the sensitivity of the weights needs to be considered. This increase might outweigh the benefits of amplification from sampling, especially since as [Bun et al. \(2022\)](#) have shown, the amplification effects tend to be small for sampling designs commonly used in practice. Which of the flavors will be most attractive from a utility perspective will crucially depend on the sampling design in practice as the design will have an effect on both the amplification and the sensitivity of the weights. It will also depend on the question whether response probabilities can be determined reliably.

6.4.4 SENSITIVITY REDUCTION FROM THE SAMPLING DESIGN

When the DP flavor is frame-invariant, the sampling design τ_f can reduce the sensitivity of a query such as the Horvitz-Thompson estimator. This is because only samples with non-zero probability are considered. Comparable sample datasets $\mathfrak{s}, \mathfrak{s}'$ must both have non-zero probability of being realized under the same sampling design τ_f . This restricts the number of comparable sample datasets, and hence potentially reduces the sensitivity of a query.

For example, if the sampling design $\tau_{\mathfrak{f}}$ includes stratification, then the stratum sample sizes are constant between comparable sample datasets $\mathfrak{s}, \mathfrak{s}'$. Thus, if \mathfrak{s} and \mathfrak{s}' differ only on a single record, that record must belong to the same stratum in both \mathfrak{s} and \mathfrak{s}' . When the difference between the possible values of x_i within strata is smaller than their difference across strata (which typically is the case whenever stratification is used to reduce the uncertainty in survey estimates), the sensitivity of the Horvitz-Thompson estimator is reduced when the DP flavor is frame-invariant.

6.4.5 UTILITY IMPLICATIONS FOR THE HORVITZ-THOMPSON ESTIMATOR

In this section, we use the Horvitz-Thompson estimator $\hat{t}_x = \sum_{i \in \mathcal{U}^R(\mathfrak{S})} w_i x_i$ discussed in Section 6.4.3 to illustrate the utility implications of the different settings. For simplicity, we assume the output of \hat{t}_x is protected using the Laplace mechanism (we do not claim this mechanism is optimal for this estimator).

THE LAPLACE MECHANISM FOR THE HORVITZ-THOMPSON ESTIMATOR

If the Horvitz-Thompson estimator must be differentially private, the corresponding Laplace mechanism can be used in place of \hat{t}_x .

Definition 6.4.2 (Dwork et al. (2006b)). Let $\varepsilon_{\mathcal{D}}\text{-DP}(\mathcal{X}, \mathcal{D}, d_{\mathcal{X}}, D_{\text{Pr}})$ be a DP specification with $D_{\text{Pr}} = D_{\text{MULT}}$. Suppose $q : \mathcal{D}_0 \rightarrow \mathbb{R}^k$ is a non-stochastic function. The *Laplace mechanism corresponding to q* is the data-release mechanism

$$T_{q, \text{LAP}}(\mathfrak{d}, \omega) = q(\mathfrak{d}) + \Delta_q([\mathfrak{d}]_{\mathcal{D}})\omega,$$

where

- the seed $\omega \in \mathbb{R}^k$ is a vector of k iid Laplace random variables, each with PDF $f(\omega_i) = \frac{1}{2} \exp(-|\omega_i|)$,

- $[\mathfrak{d}]_{\mathcal{D}}$ is the connected component

$$[\mathfrak{d}]_{\mathcal{D}} = \{\mathfrak{d}' \in \mathcal{D}_0 \mid \text{there exists } \mathcal{D} \in \mathcal{D} \text{ s.t. } \mathfrak{d}, \mathfrak{d}' \in \mathcal{D} \text{ and } d_{\mathcal{D}_0}(\mathfrak{d}, \mathfrak{d}') < \infty\},$$

- for $\mathcal{D}^* \subset \mathcal{D}_0$, $\Delta_q(\mathcal{D}^*)$ is the ε -adjusted L_1 -sensitivity

$$\Delta_q(\mathcal{D}^*) = \sup_{\substack{\mathcal{D} \in \mathcal{D} \\ \mathcal{D} \subset \mathcal{D}^*}} \sup_{\mathfrak{d}, \mathfrak{d}' \in \mathcal{D}} \frac{\|q(\mathfrak{d}) - q(\mathfrak{d}')\|_1}{\varepsilon_{\mathcal{D}} d_{\mathcal{D}_0}(\mathfrak{d}, \mathfrak{d}')},$$

(with $\|\cdot\|_1$ the L_1 -norm, $0/0 := 0$ and $\sup \emptyset := 0$).

Theorem 6.4.3. *The Laplace mechanism $T_{q, \text{LAP}}$ satisfies $\varepsilon_{\mathcal{D}}\text{-DP}(\mathcal{X}, \mathcal{D}, d_{\mathcal{X}}, D_{\text{Pr}})$.*

SENSITIVITY OF THE HORVITZ-THOMPSON ESTIMATOR

Suppose that $d_{\mathcal{D}_0}$ is the Hamming distance; the budget $\varepsilon_{\mathcal{D}} = \varepsilon$ is constant in \mathcal{D} ; and q is the Horvitz-Thompson estimator $\hat{t}_x = \sum_{i \in \mathcal{U}^R(\mathfrak{s})} w_i x_i$, with w_i the design weights. Consider sample-level DP: the domain \mathcal{D}_0 is the set of all possible samples \mathfrak{s} . For

$$\mathcal{D}_{\mathfrak{f}} = \{\mathfrak{s} : \tau_{\mathfrak{f}}(\mathcal{U}(\mathfrak{s})) > 0\},$$

define the (unadjusted) L_1 -sensitivity as

$$\Delta_q(\mathcal{D}_{\mathfrak{f}}) = \sup_{\mathfrak{s}, \mathfrak{s}' \in \mathcal{D}_{\mathfrak{f}}} |q(\mathfrak{s}) - q(\mathfrak{s}')|.$$

In this section, we will prove that the L_1 -sensitivity $\Delta_q(\mathcal{D}_{\mathfrak{f}})$ is bounded by $|\max_{i \in \mathfrak{f}}(w_i x_i) - \min_{i \in \mathfrak{f}}(w_i x_i)|$.

This is the relevant L_1 -sensitivity for frame-invariant DP flavors. For frame-agnostic DP flavors, the relevant L_1 -sensitivity is the global L_1 -sensitivity $\Delta_q(\mathcal{D}_0)$, which can only be bounded by the worst-case

$$|\max w_i x_i - \min w_i x_i| + (n - 1)(\max w_i - \min w_i)(|\max x_i| \vee |\min x_i|),$$

where n is the (fixed) size of the target sample and the maximums and minimums are all over $i \in \mathcal{U}(\mathbf{p})$ and all possible \mathbf{p} because, in general, changing a single record may change the design weights of all other records.

6.5 PRIVACY CONSIDERATIONS

6.5.1 PRIVACY SEMANTICS

POSTERIOR-TO-POSTERIOR COMPARISONS

The aim of the posterior-to-posterior framework is to compare what an attacker would learn about a single unit, if this unit is included in the input dataset relative to a counterfactual world in which the unit is not included or his or her record is not used.

Adopting notation similar to that of [Kifer et al. \(2022\)](#), let P_A be the attacker's prior on the domain \mathcal{D}_0 , i.e., the prior implies that the input dataset is treated as a random variable \mathcal{D} on the space \mathcal{D}_0 . The goal of the attacker is to infer information about the record \mathcal{D}_i of a single unit $i \in \mathcal{U}^R(\mathcal{D})$ in the input dataset \mathcal{D} . For this to be well-defined, we must assume that the units of \mathcal{D} are fixed (that is, $\mathcal{U}(\mathcal{D})$ is a fixed set). A common practice in the literature is to assume that the units of \mathcal{D} are identified by the indices $1, \dots, n$, where $n = |\mathcal{U}(\mathcal{D})|$. Throughout this section we assume $d_{\mathcal{D}_0}$ is the Hamming distance. For simplicity, we also assume that \mathcal{D}_0 and \mathcal{T} are countable spaces.

Let $t \in \mathcal{T}$ denote a realized output of the data release mechanism T . The posterior-to-posterior framework as adopted in [Kifer et al. \(2022\)](#) compares the posterior distribution $P_A(\mathcal{D}_i \in \cdot \mid T(\mathcal{D}) = t)$ with the counterfactual world in which the information of the selected unit is replaced by a random draw from the posterior distribution of the attacker assuming knowledge of everybody else. Let $psample[\mathfrak{d}] \sim P_A(\mathcal{D} \mid \mathcal{D}^- = \mathfrak{d}^-)$ denote this random draw, where \mathcal{D}^- and \mathfrak{d}^- denote the random variable \mathcal{D} and

dataset \mathfrak{D} with the selected record (\mathfrak{D}_i or \mathfrak{d}_i respectively) being removed. As shown in Kifer et al. (2022)

the ratio of these two posteriors is given by

$$\frac{P_A(\mathfrak{D}_i = r \mid T(\mathfrak{D}) = t)}{P_A(\mathfrak{D}_i = r \mid T(\text{psample}[\mathfrak{D}]) = t)} = \frac{\sum_{\mathfrak{d}^-} P_A(\mathfrak{d}^-) P_A(r \mid \mathfrak{d}^-) P_{\mathfrak{d}^- \cup \{r\}}(T = t)}{\sum_{\mathfrak{d}^-} P_A(\mathfrak{d}^-) P_A(r \mid \mathfrak{d}^-) \sum_{r'} P_A(r' \mid \mathfrak{d}^-) P_{\mathfrak{d}^- \cup \{r'\}}(T = t)}. \quad (6.2)$$

For ε -DP,

$$e^{-\varepsilon} \leq \frac{P_{\mathfrak{d}^- \cup \{r\}}(T = t)}{P_{\mathfrak{d}^- \cup \{r'\}}(T = t)} \leq e^{\varepsilon},$$

and hence the ratio of posteriors (6.2) is bounded between $e^{-\varepsilon}$ and e^{ε} , for all possible values r of \mathfrak{D}_i (see Theorem 7.1 in Kifer et al. (2022)).

IMPLICATIONS FOR THE DIFFERENT SETTINGS

The posterior-to-posterior semantics apply to the possible values r of a record \mathfrak{d}_i from the input dataset $\mathfrak{D} \in \mathcal{D}_0$, which varies depending on the DP setting. Of particular importance is the domain \mathcal{D}_0 of the DP flavor, since this determines what dataset – the (augmented) pseudo-population dataset \mathfrak{p}^* , the (augmented) frame \mathfrak{f}^* , the sample dataset \mathfrak{s} , or the responding-sample dataset \mathfrak{r} – is protected. Although not explicitly stated, the classical framework considered in most of the DP literature assumes the responding-sample-level setting, in which the domain \mathcal{D}_0 is the set of possible responding-sample datasets, $\mathcal{D}_0^{\text{resp}}$. In this case, the data-release mechanism takes as input the fixed responding sample \mathfrak{r} . As such, the protections supplied by the data-release mechanism – as measured by the posterior-to-posterior framework – apply to a record \mathfrak{r}_i from the responding sample. That is, an ε -DP mechanism with domain $\mathcal{D}_0^{\text{resp}}$ ensures that the posterior-to-posterior ratio for a responding sample record \mathfrak{r}_i is bounded in the interval $[e^{-\varepsilon}, e^{\varepsilon}]$.

If we change the DP flavor to be at the frame-level – so that we may benefit from privacy amplification by sampling – then the input to the data-release mechanism is the augmented frame \mathfrak{f}^* . As such, an ε -

DP mechanism under this setting will protect an augmented frame record \mathbf{f}_i^* – rather than a responding sample record \mathbf{r}_i – within the nominal interval $[e^{-\varepsilon}, e^{\varepsilon}]$. This distinction is important, because protection at one level does not imply protection at another level. In fact, we will see in Subsection 6.5.1 that whenever there is privacy amplification due to sampling, a sample record’s posterior-to-posterior ratio is not bounded within $[e^{-\varepsilon}, e^{\varepsilon}]$ for an ε -DP mechanism at the frame-level.

Beyond looking at the different starting points of the data release mechanism, it is also important to consider the impacts of different types of invariances. For example, treating the frame as invariant implies that neighboring datasets must come from the same fixed frame. This enforces restrictions on the possible values r of \mathcal{D}_i . As a consequence two data release mechanisms that start at the same level, for example, $\mathcal{D}_0^{\text{fr}}$ and use the same privacy loss budget ε , will offer different privacy guarantees, if one of them is frame-invariant while the other is frame-agnostic. This illustrates the ever existing trade-off between utility and privacy. From a utility perspective, it seems desirable to identify scenarios, in which enforcing invariance substantially restricts the possible values of \mathcal{D}_i as this might considerably reduce the sensitivity of the query of interest. On the other hand, shrinking the data universe $\mathcal{D} \in \mathcal{D}$ will implicitly reduce the privacy guarantees even if the privacy loss parameter is held constant.

NO PRIVACY AMPLIFICATION IF THE ATTACKER KNOWS THAT UNIT i IS IN THE SAMPLE

In this section, we show that we cannot hope for privacy amplification by sampling if we assume that the attacker knows that unit i is included in the sample $\mathcal{U}^R(\mathcal{S})$. This is a risk scenario that statistical agencies commonly need to consider in practice. In the statistical disclosure control literature, this is often referred to as the “nosy neighbor” scenario, since a possible scenario in which this kind of knowledge is realistic is the situation in which a neighbor witnesses an interviewer entering the house next door and then hopes

to learn sensitive information about the neighbor by trying to reidentify him or her in the data.

To illustrate, we consider a data-release mechanism that starts at the frame level and thus should offer privacy amplification from sampling. Specifically, suppose that T is a data-release mechanism at the sample-level and let $\mathcal{S}(\cdot)$ be the sampling function, which takes an frame \mathbf{f} and outputs the sample according to the given sample design τ . (That is, $P(\mathcal{S}(\mathbf{f}) = \mathcal{S}) = \tau(\mathcal{S})$ for all $\mathcal{S} \subset \mathcal{U}(\mathbf{f})$.) The data-release mechanism that starts at the frame level is therefore the composition $T' = T \circ \mathcal{S}$. Conditioning on the fact that unit i is included in the sample, the lower bound of the posterior-to-posterior ratio under the assumption that T is ε -DP is (below we write \mathcal{S}^R for the sample $\mathcal{U}^R(\mathfrak{S})$):

$$\begin{aligned}
& \frac{P_A(\mathfrak{F}_i^* = r \mid T'(\mathfrak{F}^*) = t, i \in \mathcal{S}^R)}{P_A(\mathfrak{F}_i^* = r \mid T(\text{psample}[\mathfrak{F}^*]) = t, i \in \mathcal{S}^R)} \\
&= \frac{\sum_{\mathbf{f}^{*-}} P_A(\mathbf{f}^{*-} \mid i \in \mathcal{S}^R) P_A(r \mid \mathbf{f}^{*-}, i \in \mathcal{S}^R) P_{\mathbf{f}^{*-} \cup \{r\}}(T' = t \mid i \in \mathcal{S}^R)}{\sum_{\mathbf{f}^{*-}} P_A(\mathbf{f}^{*-} \mid i \in \mathcal{S}^R) P_A(r \mid \mathbf{f}^{*-}, i \in \mathcal{S}^R) \sum_{r'} P_A(r' \mid \mathbf{f}^{*-}, i \in \mathcal{S}^R) P_{\mathbf{f}^{*-} \cup \{r'\}}(T' = t \mid i \in \mathcal{S}^R)} \\
&\geq \frac{\sum_{\mathbf{f}^{*-}} P_A(\mathbf{f}^{*-} \mid i \in \mathcal{S}^R) P_A(r \mid \mathbf{f}^{*-}, i \in \mathcal{S}^R) P_{\mathbf{f}^{*-} \cup \{r\}}(T' = t \mid i \in \mathcal{S}^R)}{\sum_{\mathbf{f}^{*-}} P_A(\mathbf{f}^{*-} \mid i \in \mathcal{S}^R) P_A(r \mid \mathbf{f}^{*-}, i \in \mathcal{S}^R) e^\varepsilon P_{\mathbf{f}^{*-} \cup \{r\}}(T' = t \mid i \in \mathcal{S}^R)} \\
&= e^{-\varepsilon}.
\end{aligned}$$

Whenever the mechanism T is optimal (i.e. it achieves the bound $P_{\mathfrak{s}}(T = t)/P_{\mathfrak{s}'}(T = t) = \varepsilon$ for some $\mathfrak{s}, \mathfrak{s}'$ with $d_{\mathcal{D}_0}(\mathfrak{s}, \mathfrak{s}') = 1$), the above inequality is achieved for some choice r and i . Using a similar argument, the upper bound of the ratio is e^ε . Thus, while the data release mechanism T' satisfies ε' -DP for $\varepsilon' < \varepsilon$, the posterior-to-posterior protection provided by T' when the attacker knows $i \in \mathcal{U}^R(\mathfrak{S})$ is not bounded within the interval $[e^{-\varepsilon'}, e^{\varepsilon'}]$ but only in the interval $[e^{-\varepsilon}, e^\varepsilon]$. That is, the protection due to privacy amplification from sampling is lost: T' provides the same level of protection as T when the attacker knows the unit i is in the sample.

THE JOURNALIST AND SAMPLING AMPLIFICATION

In this section, we show that privacy amplification by sampling is not possible when the attacker does not have a particular target unit in mind, but instead wishes to learn about an arbitrary record. In the statistical disclosure control literature, this is often referred to as the “journalist” scenario, since a journalist often wants to expose the vulnerability of a data-release mechanism by learning any record, rather than focusing on attacking a particular record (e.g. the record belonging to their neighbor). In this situation, it makes sense for the journalist to focus on a record that is in the sample, since these records have the most influence on the data-release mechanism’s output. As in the previous subsection, let T be an ε -DP mechanism, $\mathcal{S}(\cdot)$ be the sampling function and $T' = T \circ \mathcal{S}$, so that T' is ε' -DP with $\varepsilon' < \varepsilon$. As is common convention, let us identify the units of \mathfrak{S} as $i = 1, \dots, n$, where $n = |\mathcal{U}^R(\mathfrak{S})|$. Then

$$\begin{aligned}
& \frac{\mathbb{P}_A(\mathfrak{S}_i = r \mid T'(\mathfrak{F}^*) = t)}{\mathbb{P}_A(\mathfrak{S}_i = r \mid T'(\textit{psample}[\mathfrak{F}^*]) = t)} \\
&= \frac{\sum_{\mathfrak{s}^-} \mathbb{P}_A(\mathfrak{s}^-) \mathbb{P}_A(\mathfrak{S}_i = r \mid \mathfrak{s}^-) \mathbb{P}_A(T'(\mathfrak{F}^*) = t \mid \mathfrak{S} = \mathfrak{s}^- \cup \{r\})}{\sum_{\mathfrak{s}^-} \mathbb{P}_A(\mathfrak{s}^-) \mathbb{P}_A(\mathfrak{S}_i = r \mid \mathfrak{s}^-) \sum_{r'} \mathbb{P}_A(\mathfrak{S}_i = r' \mid \mathfrak{s}^-) \mathbb{P}_A(T'(\mathfrak{F}^*) = t \mid \mathfrak{S} = \mathfrak{s}^- \cup \{r'\})} \\
&= \frac{\sum_{\mathfrak{s}^-} \mathbb{P}_A(\mathfrak{s}^-) \mathbb{P}_A(\mathfrak{S}_i = r \mid \mathfrak{s}^-) \mathbb{P}_{\mathfrak{s}^- \cup \{r\}}(T = t)}{\sum_{\mathfrak{s}^-} \mathbb{P}_A(\mathfrak{s}^-) \mathbb{P}_A(\mathfrak{S}_i = r \mid \mathfrak{s}^-) \sum_{r'} \mathbb{P}_A(\mathfrak{S}_i = r' \mid \mathfrak{s}^-) \mathbb{P}_{\mathfrak{s}^- \cup \{r'\}}(T = t)} \\
&\geq \frac{\sum_{\mathfrak{s}^-} \mathbb{P}_A(\mathfrak{s}^-) \mathbb{P}_A(\mathfrak{S}_i = r \mid \mathfrak{s}^-) \mathbb{P}_{\mathfrak{s}^- \cup \{r\}}(T = t)}{\sum_{\mathfrak{s}^-} \mathbb{P}_A(\mathfrak{s}^-) \mathbb{P}_A(\mathfrak{S}_i = r \mid \mathfrak{s}^-) \sum_{r'} \mathbb{P}_A(\mathfrak{S}_i = r' \mid \mathfrak{s}^-) e^\varepsilon \mathbb{P}_{\mathfrak{s}^- \cup \{r'\}}(T = t)} \\
&= e^{-\varepsilon}.
\end{aligned}$$

As in the previous subsection, if T is optimal then the above inequality is achieved for some choice of t , r and i . Analogous working shows that this posterior-to-posterior ratio is bounded above by e^ε , and moreover, this bound is achieved when T is optimal. Hence, as in the previous subsection, the additional privacy protection due to amplification from sampling is lost when the attacker targets an arbitrary record

in the sample. That is, a sample record is not protected by the mechanism T' at the nominal privacy level ε' of T' , but only at the privacy level ε .

Note that this result and the accompanying discussion applies more generally beyond the context of survey sampling. They holds for any DP mechanism T' which employs amplification by sampling. Such mechanisms are frequently used as modules in sanitized (i.e. privacy-protected) machine learning and neural networks as amplification by sampling is key to sanitized stochastic gradient descent algorithms (Abadi et al., 2016; Bu et al., 2020).

6.5.2 AMPLIFICATION BY SAMPLING AND COMPOSITION

An important consideration when discussing the benefits of privacy amplification from sampling is whether the composition property of DP still hold. Composition refers to the fact that the total privacy loss of two DP mechanisms with privacy loss ε_1 and ε_2 , respectively is upper bounded by the sum $\varepsilon_1 + \varepsilon_2$ of the two losses. This is an important property as it helps to track the privacy loss over multiple data releases. This property is lost, however, in the context of privacy amplification through sampling as the following example illustrates: Consider two pure ε -DP mechanisms T_1 and T_2 with privacy loss $\varepsilon = 1$ and $\varepsilon = 2$ respectively. Suppose that they are two outputs from the same sample survey (i.e. they always use the same sample). For example, T_1 is the Laplace mechanism for querying the number of males in the sample and T_2 is the Laplace mechanism for querying the number of people in the sample with incomes over \$100,000. Suppose for simplicity that the sampling mechanism for the survey was simple random sampling without replacement (SRSWOR) with sampling fraction $f = n/N = 0.1$. Let T'_1 and T'_2 be the mechanisms which apply the sampling step and then run T_1 or T_2 respectively. These mechanisms have privacy loss 0.16 and 0.49 respectively (by amplification by sampling results given in Theorem 6.4.1). A naïve interpretation

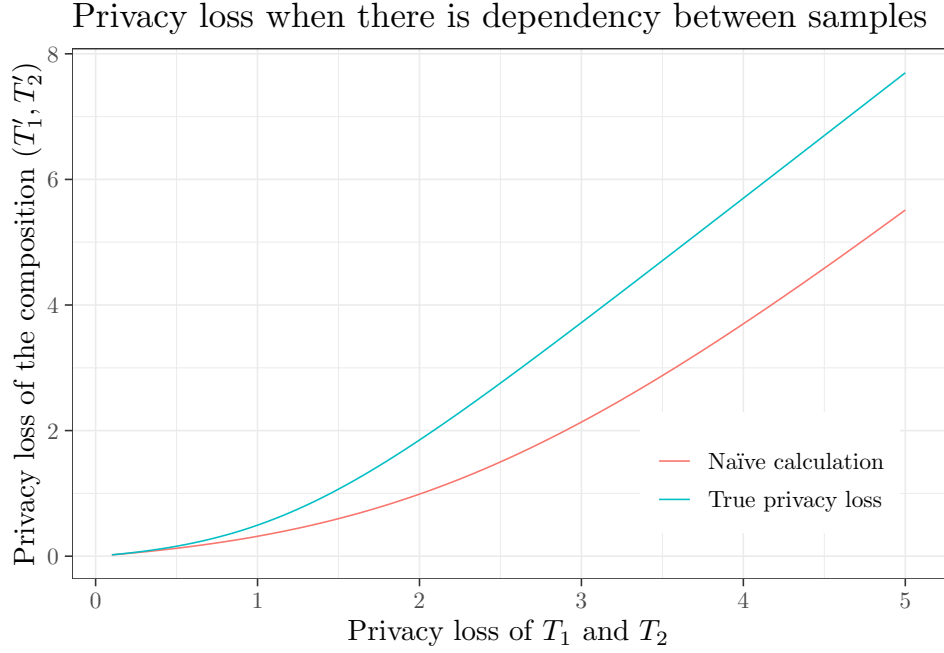


Figure 6.1: The total privacy loss over two mechanisms T'_1 and T'_2 which share the same sampling step. Here, $T'_i = T_i \circ S$ where S is simple random sampling without replacement (with sampling fraction $f = n/N = 0.1$). Both T_1 and T_2 satisfy ϵ -DP with privacy loss ϵ given on the x -axis. The total privacy loss of the composition of the two mechanisms T'_1 and T'_2 is given on the y -axis. The naïve calculation (in red) is given by the standard composition result of ϵ -DP which states that the privacy loss of (T'_1, T'_2) is the sum of privacy losses of T'_1 and T'_2 . That is, the red line is $2 \log(1 + f[\exp(\text{pl}(T_1)) - 1])$, where $\text{pl}(T_1)$ is the privacy loss of T_1 . (We assume $\text{pl}(T_1) = \text{pl}(T_2)$.) The true total privacy loss (in blue) is given by first composing T_1 and T_2 and then applying privacy amplification (Theorem 6.4.1): $\text{pl}(T'_1, T'_2) = \log(1 + f[\exp(2\text{pl}(T_1)) - 1])$.

of the composition theorem implies that their composition (T'_1, T'_2) has privacy loss 0.65. However, the correct calculus would consider the composition (T_1, T_2) – which has privacy loss $\varepsilon = 3$ – and then apply the amplification by sampling result to get a privacy loss for the composition (T'_1, T'_2) of 1.07. We note that for small sampling rates f and small values ($\ll 1$) for both ε_1 and ε_2 , the composition properties based on the amplified privacy guarantees would still hold approximately since these conditions would imply that the privacy loss of T'_i is approximately $\varepsilon'_i \approx n/N\varepsilon_i$, and thus $\varepsilon'_1 + \varepsilon'_2 \approx n/N\varepsilon_1 + n/N\varepsilon_2 = n/N(\varepsilon_1 + \varepsilon_2)$. However, for larger f or ε_i , the gap between the true privacy loss and the naïve calculation can be substantial, as illustrated by Figure 6.1.

The source of this apparent contradiction is the composition theorem’s implicit assumption that the seeds ω'_1 and ω'_2 of T'_1 and T'_2 are independent. This assumption does not hold when T'_1 and T'_2 always select the same sample. More generally, suppose that T'_1 and T'_2 are mechanisms which include sample procedures with designs τ_1 and τ_2 respectively. Then the composition theorem’s assumption is violated whenever the sample designs τ_1 and τ_2 are dependent. In such cases, the calculation of the total privacy loss across T'_1 and T'_2 cannot rely on applying the composition theorem to T'_1 and T'_2 . Instead, this calculation requires analyzing the privacy amplification of the sample designs τ_1 and τ_2 *jointly*, which will be difficult in general.

Dependency between sample designs is unfortunately a common occurrence at many NSOs. Beyond the above example where T'_1 and T'_2 use the same sample, there are (at least) two other common scenarios which lead to violations of the composition theorem’s independence assumption. Firstly, because NSOs run many different survey collections concurrently, modern sample designs aim to reduce respondent burden by controlling the overlap between the samples of different surveys. (For example, if a unit was selected

for one survey, they will have a lower (or zero) probability of being selected in the near future for a different survey.) This introduces dependence between the sample designs of the NSO's different surveys. Secondly, sample rotation – which is a common feature in the collection of time series data, such as labor force statistics – introduces dependency between the sample designs across time for the same survey.

In all three of these scenarios, frame- (or population-)level DP mechanisms will not have independent seeds and hence the standard composition theorem does not apply to these mechanisms. This is an important consideration in determining the total privacy loss of an NSO across their multiple surveys. In situations traditionally encountered in the DP literature, the composition theorem allows for modular privacy analyses, but – without a generalized composition theorem which can account for dependency between seeds – an NSO will be forced to resort to a joint privacy analysis which must simultaneously analyze all the NSO's surveys. Therefore, an important (and novel, as far as we are aware) future research is to understand the composition property of DP under varying levels of seed dependency. Such an understanding will enable modular privacy analyses of dependent DP mechanisms to be combined into an overall privacy loss – as the standard composition theorem currently enables for independent DP mechanisms.

We conclude this subsection with the general comment that the composition of multiple mechanisms becomes more complex when these mechanisms share data-processing steps in common. Sampling is an example of one such data-processing step, but it is by no means the only example. Population-level DP mechanisms will also share the same process of frame construction (even if they use different frames, it is likely that there are dependencies between the construction of the two frames), which must be accounted for when determining the overall privacy loss.

6.6 DISCUSSION

This paper develops theory for understanding and implementing differential privacy in the context of survey statistics. By recognizing the major phases in the survey-data pipeline, we identified ten different settings of DP. These settings correspond to different choices for 1) where the DP data-release mechanism starts in this pipeline; and for 2) which of the previous phases are taken as invariant. Section 6.3 formalized these ten settings into ten different conditions on the DP flavor.

Sections 6.4 and 6.5 show that the choice of the setting has significant impacts in terms of both privacy and utility. Therefore, while DP is invariant to post-processing, pre-processing steps matter. Moreover, the data custodian must necessarily choose a setting – they cannot implement DP without first deciding (perhaps implicitly) where the DP mechanism starts and which pre-processing steps are taken as invariant. Hence, contrary to commonly-held beliefs, DP does make important assumptions on the data and on the attacker, because the data custodian’s decision impacts both the utility and privacy semantics of the DP-outputted data.

Based on the discussions in the previous sections, we can offer some recommendations on the settings a data custodian might want to choose. Firstly, we advice against the population-level setting (i.e. using the domain $\mathcal{D}_0 = \mathcal{D}_0^{\text{PP}}$). Compared to the frame-level setting ($\mathcal{D}_0 = \mathcal{D}_0^{\text{fr}}$), the only advantage of the population-level setting would be potential amplification gains because the frame could be treated as a random subset of the pseudo-population. However, quantifying the resulting privacy amplification effects seems difficult, if not impossible, in practice. Moreover, for high quality frames the amplification effect should be small since the fraction of the pseudo-population on the frame would be high. On the other hand, using $\mathcal{D}_0^{\text{PP}}$ would always require the DP flavor to be frame-agnostic, implying that the design weights

could no longer be treated as fixed. This would potentially increase the sensitivity of the output of interest and would make the computation of the sensitivity challenging in most cases.

Secondly, opting for the frame-level setting ($\mathcal{D}_0 = \mathcal{D}_0^{\text{fr}}$) offers amplification from sampling, but requires a frame agnostic DP flavor, implying that the sampling weights still cannot be treated as fixed. Since previous research has shown that amplification effects tend to be small for many complex sampling designs (Bun et al., 2022) and privacy amplification is only achievable if the nosy neighbor and the journalist scenario discussed in Sections 6.5.1 and 6.5.1 are unrealistic threat models, it seems that the benefits of amplifications are outweighed by the disadvantages of this DP setting.

Thirdly, when using one of the sample-level settings, it seems preferable to work under $\mathcal{D}_{\text{sampl}|\text{fr}}$, i.e., treating the frame as invariant, as this would allow the design weights to be treated as fixed. These benefits should outweigh the fact that treating the frame as invariant will increase the risks by limiting the space of neighboring datasets. These constraints on the possible values of a record \mathfrak{s}_i may be small in practice, although more research is needed to verify this.⁴ In principle, the sample-level setting would also offer amplification from nonresponse. However, as discussed previously, quantifying these amplification effects would require knowledge of the true response mechanism.

Finally, we do not see any benefits from starting the data release mechanism only at the responding sample. If the data custodian still prefers to choose this option, we would recommend using $\mathcal{D}_{\text{resp}|\text{fr}}$ and not $\mathcal{D}_{\text{resp}|\text{sampl}}$. Our concern is that treating the target sample as fixed might enforce strong constraints on the possible values of a record \mathfrak{r}_i in some circumstances. Whether one can find examples where this is really the case would be an interesting area for future research.

⁴In the final version of this paper, we will address this question in further detail.

7

The Complexities of Differential Privacy for Survey Data^I

7.1 INTRODUCTION

DIFFERENTIAL PRIVACY (DP) (Dwork et al., 2006b) HAS BECOME THE QUASI-GOLD STANDARD in recent years for data collection and dissemination whenever privacy or confidentiality is a concern. It offers formal (that is, mathematically quantifiable) privacy guarantees by bounding the influence that any single record of the database can have on the computed outputs. The fundamental difference to earlier privacy frameworks such as k -anonymity is that the guarantees are a property of the mechanism generating the output and not a property of the data. DP specifies how much noise the mechanism needs to introduce to ensure that the probability of obtaining a specific result does not change substantially, if one record in the database is changed. In simple examples where we are interested in creating a DP version of an unpro-

^IBased on work coauthored with Jörg Drechsler.

tected statistic such as the sample mean, the required amount of noise depends on two components. The first component is the privacy-loss parameter ε , which determines how much the probability of obtaining a specific result is allowed to change.² The smaller the privacy-loss parameter, the more noise needs to be added and the better the level of protection offered. The second component is the sensitivity of the unprotected statistic of interest, which is measured as the maximum possible change of the statistic when changing one record in the database. The higher the sensitivity, the more noise needs to be added.

To illustrate, we can look at one of the classical DP mechanisms that is often used as a building block in more complex algorithms: the Laplace mechanism, which, for any univariate statistic f , ensures ε -DP by adding a random draw from a Laplace distribution centered at zero with scale parameter $b = \Delta f / \varepsilon$. The parameter Δf is the sensitivity of f measured as the maximum absolute distance (the L_1 norm) of the statistic computed over two neighboring datasets, i.e., two datasets that differ only in a single record. With this mechanism, the dependence on the two parameters is obvious: More noise is added for outputs with higher sensitivity and smaller values of ε .

This is one of the attractive properties of DP. The concept is very intuitive and requires only three steps, which in principle seem straightforward to apply: (i) define the maximum privacy loss that is still considered acceptable and select a value for ε accordingly; (ii) identify the sensitivity of the statistic of interest (for example, the sensitivity of a proportion under bounded ε -DP is simply $1/n$, where n is the number of records in the database); and (iii) choose a DP mechanism that infuses the right amount of noise into the reported output based on the parameters from steps (i) and (ii). Of course, in practice all three

²For simplicity we limit our exposition to the classical bounded ε -DP setting, where ‘bounded’ means that neighboring datasets are defined as datasets that can be obtained by replacing a single record with another record without changing the size of the database. Similar arguments would apply for other variants, such as (ε, δ) -DP, ρ -zero-concentrated DP, or f -DP, and for other definitions of neighboring datasets, such as unbounded DP for which a neighboring database is obtained by adding or removing a single record.

steps have their challenges. The discussion on how to choose and interpret the privacy loss parameter(s) shows no signs of abating (Hsu et al., 2014; Dwork et al., 2019; Abowd and Schmutte, 2019; Tschantz et al., 2020; Nanayakkara et al., 2023; Drechsler, 2023; Bailie et al., 2025b); the sensitivity of the output is not always easy to compute and can be unbounded without further assumptions (Casacuberta et al., 2022); and finding a suitable DP mechanism can be challenging. Besides, there are often some hidden complications to DP in practice beyond what this three-step process makes apparent (Abowd et al., 2022a; Seeman and Susser, 2023; Cummings et al., 2024). (For example, for the same research question there can be multiple choices for which statistic is used in step (ii), and it can be difficult to determine which one leads to the most efficient DP mechanism.) Still, the three components remain the same across applications and at least the general setup is well defined.

However, when working with survey data, there are additional complexities which typically do not arise in other settings. Moreover, the implications of using DP in the context of surveys have received little attention in the DP literature until recently. This led the U.S. Census Bureau to conclude in 2022 that “the science does not yet exist” to implement DP in its American Community Survey (US Census Bureau, 2022c). An expert panel convened by the National Academies of Sciences, Engineering, and Medicine reached a similar conclusion with respect to the Survey of Income and Program Participation (National Academies of Sciences, Engineering, and Medicine, 2024).

Given its commitment to formal privacy for all its data products, including its surveys (US Census Bureau, 2018), the U.S. Census Bureau launched a research project in 2020 (which is currently still ongoing) to better understand the complexities that arise when adopting DP in the survey context. In this paper, we will summarize some of the key findings of this project so far and also discuss some of the challenges

that still need to be addressed. Overall, we identify (at least) five aspects that need to be considered when implementing DP in the survey context:

- Data production is a multistage process. As such, there are various options for how and where to integrate DP in this pipeline, each of which come with their own advantages and disadvantages.
- Previous studies found that sampling can amplify DP's privacy guarantees. However, these amplification effects do not necessarily hold for the complex sampling designs used by statistical agencies.
- These complex sampling designs need to be incorporated into any survey statistic and hence must also be incorporated into any DP mechanism.
- Weighting adjustments are routinely used to account for unit nonresponse and to benchmark to known population totals. As these adjustments can substantially increase the sensitivity of the survey statistic, there is a need to develop robust adjustment strategies which are congenial to DP.
- Item nonresponse is often addressed using imputation, but similar to weighting adjustments some standard imputation techniques can greatly inflate the sensitivity of the resulting statistic. Ongoing research is currently investigating the feasibility of differentially private imputation techniques.

We will discuss each of these aspects in the remainder of this paper.

7.2 DP AND THE MULTISTAGE PROCESS OF DATA PRODUCTION

7.2.1 THE SURVEY PIPELINE

The production of survey data is a complex multistage process (Figure 7.1). The design of a survey typically begins by conceiving the *target population*: the set of units that one wants to study. Usually, the target

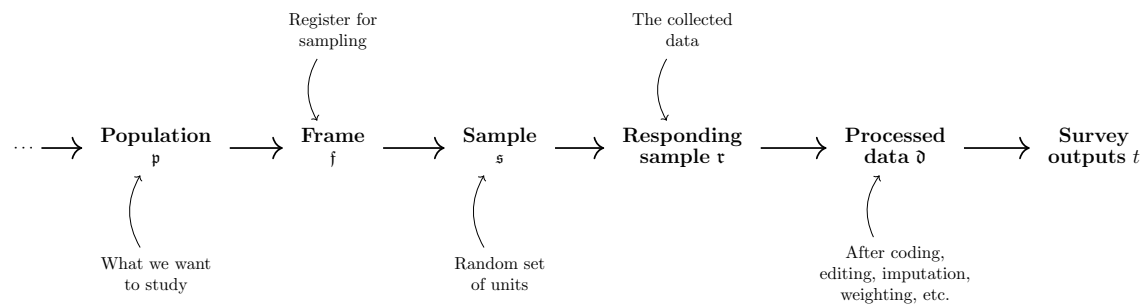


Figure 7.1: A *survey pipeline* consists of multiple steps, of which some of the most important are: determining the target population to be studied; constructing the frame; drawing the sample; collecting survey data from the responding units; processing the data (including coding free-form responses; editing inconsistent or improbable data; imputing missing records or variables; calculating the survey weights; and injecting privacy-protecting noise); and computing the survey outputs. There are of course additional steps to a survey pipeline after the survey outputs are released (such as data analysis) but, as they are not important to this paper’s subject, we exclude these steps from discussion. While not shown in this figure, it should be noted that data from previous stages of a pipeline are often used in later stages. (For example, the frame is usually used in computing the survey weights during the production of the processed data.)

population is not actually specified as a concrete list of units. Instead, it is defined conceptually: “all adults in Massachusetts” or “all businesses in Hawaii.” Once the target population has been defined, the *frame* is sourced. The frame is a register of units from which the sample will eventually be drawn. It must include sufficient contact information so that the sampled units can be surveyed. The frame should align with the target population as much as possible. However, perfect alignment is not possible in most cases, even when the target population and the frame have the same inclusion criteria, because errors will typically be made in the frame’s construction. These errors will result in overcoverage (including units which are not in the target population) and undercoverage (not including units which are in the target population).

A *sample* is randomly drawn from the frame according to the survey’s *sampling design*: the probability distribution which specifies for every potential sample the chance that that sample is selected. After sample selection, the statistical agency will solicit survey data from the sampled units. Most surveys, especially modern ones, suffer from nonresponse. This means only a subset of the sampled units will respond and

the agency will not obtain survey data from the other units. Data collected from the responding sample, along with the frame and some auxiliary information (such as data from administrative records or from previous censuses or surveys), are passed through a number of complex data processing steps before the survey outputs are computed and released. These data processing steps often include editing survey responses to correct errors in data recording; coding each free-form answer into a categorical variable; imputing missing answers to individual survey questions (“item nonresponse”) or to the entire survey questionnaire (“unit nonresponse”); and calculating multiple sets of survey weights for each record—to account for unequal probabilities of selection in the sampling design, to mitigate bias due to nonresponse patterns, and to calibrate survey data to auxiliary sources of information. Finally, we note that data may be deliberately injected with artificial noise at any point in the survey pipeline, so that releasing the survey outputs does not breach the privacy of the data subjects.

7.2.2 DP IN THE SURVEY PIPELINE

DP is a criterion applied to *data-release mechanisms*: algorithms that take data as input and produce a set of outputs which will then be published (that is, “released”). Implementing DP involves both designing a data-release mechanism which is compliant with DP, as well as integrating that mechanism into the relevant data pipeline. Both tasks are crucial for successfully producing outputs with high accuracy and good privacy protection.

There are two important considerations when integrating a DP mechanism into a data pipeline. Firstly, at what point in the pipeline should the DP mechanism start? And secondly, which of the earlier stages of the data pipeline should be considered invariant – i.e., should be treated as fixed – by DP? With survey pipelines, there are a number of possible options with respect to both considerations. In the option most

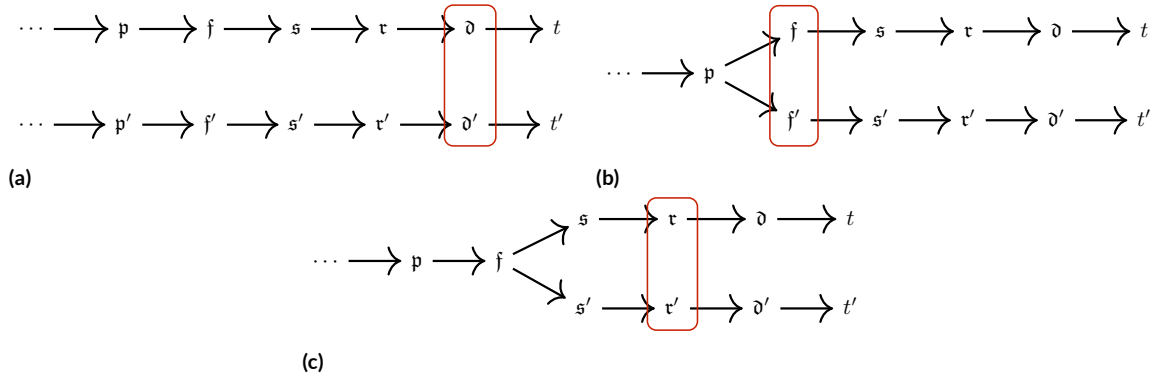


Figure 7.2: Three examples of where to start the data-release mechanism (circled in red) in the survey pipeline and which of the previous stages to take as invariant (those stages before the pipeline branches). Recall from Figure 7.1 that p denotes the population, f the frame, s the sample, r the responding sample, d the processed data and t the survey outputs. The apostrophe $'$ indicates an alternative realisation of the associated variable. Figure (a) illustrates the standard approach in which there are no invariants and the data-release mechanism only executes the final step of the survey pipeline—transforming the processed data into the survey outputs. In Figure (b), the mechanism begins with the frame and includes the sampling, responding and processing steps. The population is considered invariant. In Figure (c), the mechanism takes as input the responding sample. Both the population and the frame are taken as invariant, so that DP only compares samples from the same frame. This reduces the sensitivity of weighted estimators at the expense of reduced privacy (Section 7.4).

commonly seen in the DP literature, the data-release mechanism starts at the end of the pipeline and performs just the last step – computing the survey outputs from the processed data – and none of the previous steps are taken as invariant (Figure 7.2a). However, a mechanism could conceivably start at any point of the survey pipeline and incorporate all the steps that follow. For example, it could take as input the frame, execute the sampling step, process the data and finally compute the survey outputs (Figure 7.2b). Furthermore, any of the steps before the mechanism starts could conceivably be taken as invariant. In the rest of this subsection, we will explore these two considerations in turn.

Throughout this paper, we assume that the data-release mechanism always includes the final step of the survey pipeline, the computation of the survey outputs.³ Under this assumption, a survey pipeline

³Technically, a data-release mechanism is simply an algorithm that takes data as input and outputs some (possibly noisy) transformation of that data. So, in principle, a data-release mechanism could be incorporated into a survey pipeline even if it ends before the final step of the pipeline. (And such a mechanism could still be compliant with DP.) In this case, the survey pipeline includes additional post-processing steps after the data-release mechanism ends

can be split into those steps which are executed before the data-release mechanism starts and those steps which are executed by the mechanism. Yet choosing where to make this split is not a simple matter. In fact, there are a number of complexities associated with starting the data-release mechanism earlier or later in the pipeline. We identify five.

Firstly, starting the DP mechanism earlier can complicate the computation of the cumulative privacy loss across multiple data-release mechanisms because DP’s composition theorems⁴ are not applicable when there is dependence between the mechanisms’ noise terms (which can happen, for example, when their sampling designs are dependent) (Bailie and Drechsler, 2024).

Secondly, as we will describe in Section 7.3, including the sampling step within the data-release mechanism can amplify DP’s privacy guarantees without degrading data utility. However, this privacy amplification can be nullified if the attacker knows that the record they are attacking is in the sample (Bailie and Drechsler, 2024). More generally, if the attacker has knowledge about information intermediary to the DP mechanism (that is, information which is conditionally dependent on confidential data, or on the artificial noise introduced by the mechanism, conditioning on the output of the mechanism), the privacy guarantees afforded by DP can be weakened. For this reason, DP prohibits the direct release of such infor-

but before the computation of the outputs which will be published. Such post-processing steps are usually included to improve the utility, usability or accessibility of the survey outputs. On the other hand, any data-release mechanism can always be extended to one which ends with the final step of the survey pipeline, and any DP guarantees afforded to the original mechanism automatically carry over to the extended one by the post-processing theorem. (The post-processing theorem states that any function of a DP mechanism’s output – i.e. any “post-processing” – also satisfies DP with (at most) the same privacy loss.) Therefore, we do not gain anything by considering DP mechanisms that end before the survey pipeline’s final step.

⁴A composition theorem describes how to bound the total privacy loss incurred by multiple DP data releases which are all based on the same confidential dataset. For example, the composition theorem for pure ϵ -DP states that: if there are K mechanisms M_1, \dots, M_K , which all satisfy pure ϵ -DP and all have the same input dataset, then the total privacy loss – that is, the privacy loss of the mechanism that publishes all the outputs of M_1, \dots, M_K together – is bounded by the sum $\sum_k \epsilon_k$ over the privacy losses ϵ_k of each mechanism M_k . Existing composition theorems assume that the noise added by each mechanism is “fresh”, i.e., independent of everything else.

mation. Therefore, because the choice of the sampling design is often dependent on data in the frame, the sampling design cannot be directly made public but instead can only be released by including it in the set of DP-protected survey outputs.

However, defining a data-release mechanism – let alone one that satisfies DP – which releases the sampling design is challenging due to the third complexity we identify: Incorporating existing steps of a survey pipeline into a data-release mechanism can be difficult. A data-release mechanism is an algorithm which must be fully specified in order to be analysed by DP; hence any stage of the survey pipeline must first be fully “algorithmized” (that is, the process by which each of the stage’s possible inputs is transformed into one of its outputs must be completely and programmatically specified) before it can be included in a mechanism.⁵ A survey pipeline often includes a number of complex, ill-defined and human-intensive tasks, such as building the frame, choosing a sampling design, coding and editing. Because these tasks all usually require a degree of human judgment, they would be difficult to algorithmize. Moreover, including these procedures – or other procedures often found in a survey pipeline – in a data-release mechanism can add difficulties to making the mechanism compliant with DP. (In later sections, we will discuss some such difficulties as they relate to the weighting and imputation procedures.)

Fourthly, even if a data-release mechanism begins later in the survey pipeline so that some steps of the pipeline do not have to be incorporated in the mechanism, implementing DP still requires understanding those steps’ effect on the mechanism’s input data. For example, some imputation techniques replace

⁵The post-processing theorem provides an exception to this general rule. If the preliminary steps of a data-release mechanism, taken on their own, satisfy DP, then the later steps of the mechanism do not need to be algorithmized, because the post-processing theorem ensures that the mechanism as a whole always satisfies DP regardless of what the later steps do. All that must be checked is that the later steps only use the DP outputs from the preliminary steps, and not some other data. However, this exception does not apply to the survey pipeline steps under discussion (choosing a sampling design, coding and editing) because these steps are typically applied before – not after – privacy protection.

missing records with copies of non-missing donor records. This means an individual survey respondent can contribute to multiple records in the post-imputation dataset. This complicates the appropriate definition of neighboring datasets, since there is no longer an exact correspondence between the dataset's records and the real-world entities (the individual respondents) that should be protected: In the post-imputation dataset, changing a single record does not correspond to changing the data of one entity. Hence, naïvely applying DP to the post-imputation dataset will not provide a donor record with the expected level of protection; that is, the privacy guarantees for a donor record will be weaker than those for a post-imputed record. In general, the later the DP mechanism begins, the more difficult it is to determine an appropriate notion of neighbors since steps earlier in the pipeline may introduce dependencies between dataset records, thereby complicating the relationship between records and data subjects.

Fifthly, and most fundamentally, the starting point of the data-release mechanism determines what form of the data is protected by that mechanism. For example, if a DP mechanism begins after data processing, then it is the processed data – and not, for example, the raw responses from the data providers – which are protected by that mechanism. That is to say, DP guarantees implicitly assume that the attacker is interested in inferring the data that is input into a DP mechanism. Measures of protection are in terms of the attacker's ability to learn this input data – and not the data at other points in the pipeline. If the DP mechanism takes the processed data as input, then the DP guarantees apply to the processed data and do not necessarily carry over to the responding sample data. In order to have guarantees for the responding sample, the statistical agency must show that the pipeline from the responding sample to the survey outputs (considered as a data-release mechanism) also satisfies DP.

These five complexities demonstrate that there can be conflicting demands in deciding where a DP

mechanism should start within the survey pipeline. For example, suppose a statistical agency wants to protect the unprocessed survey responses. Then either the coding and editing steps will need to be included in the agency's mechanism (which may be difficult because these steps could be hard to algorithmize) or these steps will need to be removed from the survey pipeline (which could decrease the quality of the survey outputs).

We now return to the question of which steps of the survey pipeline should DP take as invariant. DP assesses the privacy of a data-release mechanism by comparing the survey outputs' distribution under pairs of counterfactual input datasets. These input datasets are generated by counterfactual runs of the initial steps of the pipeline, up until the data-release mechanism begins. By taking some of these steps as invariant, DP's counterfactual comparisons are reduced to only those pairs of input datasets which share the same realization of the invariant steps. For example, suppose the steps in the survey pipeline which generate the population and the frame are taken as invariant and the data-release mechanism starts with the responding sample (Figure 7.2c). Then DP only compares those responding samples (i.e. those counterfactual input datasets) which could have come from the same frame. Adding invariants will weaken the privacy guarantees provided by DP (Kifer et al., 2022; Abowd et al., 2022a; Bailie et al., 2025b). In general, the later the stage of the pipeline that is kept invariant, the greater the reduction in privacy. However, invariants may be justifiable when the output of the invariant steps can be considered as public knowledge (such as if the frame was sourced commercially rather than constructed from confidential information). Moreover, constraining some steps to be invariant has the advantage of reducing the sensitivity of weighted estimators and thereby decreasing the noise which must be added for privacy protection (Section 7.4).

7.3 DP WITH COMPLEX SAMPLING DESIGNS

Statistical agencies have been aware for decades that sampling can be a simple and effective strategy to reduce disclosure risks simply because an attacker can no longer be sure whether a specific target record is included in the sample or not. This is the main motivation why most statistical agencies only release samples from their censuses as public use micro datasets (they typically also apply additional measures to further increase the level of protection). This idea has been formalized in several papers in the context of DP (Kasiviswanathan et al., 2011; Wang et al., 2016; Bun et al., 2015; Balle et al., 2018; Wang et al., 2019). The authors show that the level of privacy is amplified through sampling, i.e., the actual privacy guarantees are higher than those implied by the chosen privacy loss parameters when protecting the sample output. Specifically, for small sampling rates r and small privacy loss parameters ϵ , applying certain simple sampling designs (simple random sampling with and without replacement, and Poisson sampling) before running an ϵ -DP mechanism reduces the privacy loss to approximately $r\epsilon$.

However, most surveys conducted by statistical agencies use complex multistage sampling designs, potentially with different sampling strategies at the different stages. These designs are primarily used to increase the accuracy of the survey outputs or to reduce the survey's operational costs. For example, the Current Population Survey (CPS), one of the flagship surveys of the U.S. Census Bureau, uses a two-stage sampling design in which stratified cluster sampling with probability proportional to size (PPS) is used to select clusters at the first stage and systematic sampling is used to sample households within clusters at the second stage (US Bureau of Labor Statistics, 2018b). There is no reason to believe that amplification effects for these complex designs are comparable to those obtainable for the simple designs discussed above. Bun et al. (2022) study the amplification effects for complex designs and find that amplification is small

for most of the sampling designs used in practice. Their findings can be summarized as follows:

- Cluster sampling using simple random sampling without replacement to draw the clusters offers negligible amplification in practice except for small ϵ and very small cluster sizes.
- With minor adjustments, stratified sampling using proportional allocation can provide privacy amplification. For small ϵ , the amplification is still linear in the sampling rate up to a constant factor.
- Data dependent allocation functions such as Neyman allocation for stratified sampling will likely result in privacy degradation. (The effects will depend on the sensitivity of the allocation function.)
- With PPS sampling at the individual level, the privacy amplification will linearly depend on the maximum probability of inclusion (for small ϵ).
- Systematic sampling will only offer amplification if the ordering of the population is truly random. In all other cases, systematic sampling will suffer from the same effects as cluster sampling, leading to no amplification (assuming the ordering is known to the attacker).

In practice this implies that for many multistage sampling designs, which typically start with (multiple stages of) stratified cluster sampling, amplification effects can generally only be expected from those stages at which individual units or households are selected (typically the last stage of selection).

7.4 DP FOR SURVEY WEIGHTED ESTIMATES

As discussed in the introduction, the amount of noise that needs to be added to achieve a specific privacy loss ϵ directly depends on the sensitivity of the statistic of interest. Intuitively, this makes sense. If the statistic changes substantially when one record is changed in the data it will be easier to infer that record's

value from observing the statistic and thus more noise will need to be added to sufficiently protect that record. From a utility perspective, this implies that more reliable (less noisy) DP outputs can be expected from statistics with low sensitivity. Thus, a common strategy with DP is to identify estimation strategies with low sensitivity and replace very sensitive estimates with less sensitive alternatives, for example by using robust statistics (Dwork and Lei, 2009; Avella-Medina, 2021).

When analyzing survey data, it is generally important to take the sampling design into account since the probabilities of selection typically vary between the units included in the sample. Unweighted estimates, especially those for descriptive statistics such as means and totals, will be biased whenever there are varying selection probabilities. To obtain unbiased estimates, each observation needs to be weighted by the inverse of its probability of selection. Hence, statistical agencies typically provide survey weights to enable researchers to take the survey design into account. In practice, these survey weights will also account for nonresponse and other data deficiencies such as undercoverage. (We will address this extra layer of complexity in the next section.)

Using survey weighted estimates raises the question: how (if at all) does the sensitivity of a statistic change when the survey design is taken into account? To illustrate the possible impacts, let us assume the analyst is interested in estimating the mean of some variable Y in the population using the sampled values $y_i, i = 1, \dots, n$, where n denotes the sample size. If the probabilities of selection were equal for all units, the sample mean would be an unbiased estimate for the population mean and its sensitivity would be R/n , where $R = \max(y_i) - \min(y_i)$ is the range of all possible values for y_i .⁶ When dealing with unequal probabilities of selection, a popular estimator for the population mean is the Horvitz-Thompson estimator

⁶Throughout this section, we consider the bounded ε -DP setting. Similar arguments (with slightly different values for the sensitivity of a statistic) would apply for other settings.

(Horvitz and Thompson (1952)): $\hat{\mu}_Y^{HT} = \sum w_i y_i / N$, where w_i is the weight of unit i , for $i = 1, \dots, n$ and N is the size of the population. Note that we assume for simplicity that N is known and does not need to be protected and w_i is the design weight, i.e., it only accounts for the sampling design.

If we can treat the weights as fixed, the sensitivity of $\hat{\mu}_Y^{HT}$ is $\max(w_i)R/N$. Whether the maximum is over all units in the frame, over all units in the population, or over all possible counterfactual units, depends on which stages of the survey pipeline are treated as invariant as discussed in Subsection 7.2.2. Note that for equal-probability designs all $w_i = N/n$ and thus the sensitivity of the Horvitz-Thompson estimator is the same as for the unweighted estimator. If $\max(w_i) > N/n$, the Horvitz-Thompson estimator will have larger sensitivity than the unweighted estimator.

However, these discussions assume that the weights can be treated as fixed, that is, they do not change if a record changes in the database. For most sampling designs used in practice, such an assumption is unrealistic. For example, with sampling proportional to size (PPS), the i th record's probability of inclusion is given by $\pi_i = (n \cdot x_i) / N \cdot \bar{x}$, where x_i is the value for unit i of the measure-of-size variable X that is used to improve the efficiency of the sampling design, and $\bar{x} = \sum_N x_i / N$ is the population mean of X . Changing the value of X for a single record will change the probabilities of inclusion and thus the survey weights for all other records in the sampling frame. Therefore, the sensitivity will be larger compared to the setting with fixed weights as we no longer only need to consider the maximum possible change in a single record's value for Y . We also need to consider the impact of the weight change for all the other records even if their values for Y don't change.

A recently-proposed strategy to mitigate this potentially-substantial increase in sensitivity is to regularize the weights, as explored by Seeman et al. (2024). (An extreme version of this strategy would set all weights to be equal; this could be justifiable if the increase in the privacy noise due to the weights dwarfs

the bias introduced by ignoring the sampling design.) Another possible strategy is to treat the frame as invariant as discussed in Figure 7.2c. Frame invariance assumes any two neighboring datasets must always originate from the same frame and so can only differ at the sample level (or later). Thus, the probabilities of inclusion will be constant between neighboring datasets. However, treating the frame as invariant has two additional implications that need to be considered. First, fixing the frame implies that privacy amplification from sampling is no longer possible (we would need to have neighboring datasets at the frame level in order to achieve amplification). However, given the results of [Bun et al. \(2022\)](#), this amplification is likely small in practice and thus the positive effects of reducing the sensitivity will tend to outweigh the negative effects of losing the amplification effect. On the other hand, fixing the frame will restrict the possible counterfactual input datasets to those which are consistent with the realized frame. Because this restriction will fix the survey weights, it might introduce strong constraints on the possible neighboring datasets, depending on the sampling design. As a consequence, the actual privacy guarantees for a frame invariant setting could be significantly weaker than the guarantees under a non-frame-invariant setting even for the same privacy loss parameter. How problematic this reduction in privacy is in real settings is currently an open question for research.

7.5 DP AND WEIGHTING ADJUSTMENTS

In practice, two adjustment steps are commonly applied to the design weights to correct for unit nonresponse and other data deficiencies such as over- or undercoverage in the sampling frame: nonresponse adjustments and calibration. Nonresponse is typically taken into account by modeling each survey unit's probability to respond and then multiplying the design weights with the inverse of the estimated response propensities. Calibration techniques rely on benchmarks known from other sources such as census data

or large scale surveys such as the American Community Survey (US Census Bureau, 2022b). These techniques can be used to adjust the survey weights in such a way that the survey weighted estimates will match the known benchmarks exactly. How these adjustment steps interfere with differential privacy has not been studied so far. (We are currently at an early stage of trying to address this problem.) However, both steps are data dependent, that is, they use information from the survey units for the adjustments. This implies that these steps cannot be ignored from a privacy perspective as the adjusted weights leak some personal information. Looking at the impacts on the sensitivity of the final statistic of interest (which uses the adjusted weights), similar problems as those discussed in the previous section will arise: changing one record in the database can potentially change the weight-adjustment factors for all other units in the survey. Thus, it seems imperative not to only account for these adjustment steps at the analysis stage. Better results in terms of the privacy-accuracy trade-off might be achieved if the weight-adjustment steps would be carried out in a differentially private way. More research is needed to better understand this trade-off. For example, it seems beneficial to identify robust adjustment strategies as less noise would be required to satisfy DP for these strategies.

In the particular case of post-stratification (which is a simple type of calibration), one such robust adjustment strategy has been proposed by Clifton et al. (2023). Another strategy would be to regularize the nonresponse and calibration weight adjustments. (This would be similar to the survey weight regularization strategy of Seeman et al. (2024) discussed in the previous section.)

7.6 DP AND IMPUTATION

All survey data are plagued by item nonresponse as survey respondents are often unwilling or unable to respond to all survey questions especially if they request sensitive information. A common strategy to

deal with this problem is to impute the missing values before analyzing the data. Imputation is especially helpful if the response process is selective, that is if it is not missing completely at random as defined by Rubin (1976). In this case, using only the fully observed cases for the analysis would give biased results. However, imputations are always data dependent as they typically build a model based on the observed data and use this model to impute the missing values. As a consequence, the implications of imputation on the DP guarantees need to be considered regardless of whether or not the imputation procedure is included inside the data-release mechanism. Some preliminary results for this problem are discussed in Das et al. (2022).

Similar to the problem of weighting adjustments, there are two possible strategies to account for imputation under DP. The first strategy only considers the effects when analyzing the imputed data. The second strategy modifies the imputation routines to ensure that the imputations already satisfy DP. As Das et al. (2022) have shown, the first strategy implies that in the worst case the sensitivity increases linearly with the number of imputed observations. This substantial increase of the sensitivity arises because changing one record in the database can potentially impact all of the imputed values. Whether the worst case applies depends on the analysis of interest and on the selected imputation procedure. Still, for statistical agencies offering pre-imputed datasets for accredited researchers, this strategy is not an option since they cannot anticipate which analyses might be performed on the imputed data.

The second strategy can break the dependence on the number of imputed records at least for certain imputation strategies. The key requirement for breaking the dependence is that the imputation model m can be written as $D_{imp}^{(i)} \sim m(D_{obs}^{(i)}, \hat{\theta})$, where $D_{imp}^{(i)}$ and $D_{obs}^{(i)}$ contain the imputed and observed variables for record i and $\hat{\theta}$ denotes the model parameters estimated on the complete data. The model implies that, given

$\hat{\theta}$, the imputed values of record i only depend on the observed values of that record and not on any other record. If these requirements are met and the parameters θ of the imputation model are estimated using any suitable differentially private mechanism with privacy loss parameter ε_1 , then, given any ε_2 differentially private mechanism used for analyzing the data, the overall privacy loss is given by $\varepsilon_1 + \varepsilon_2$.

We note that the conditional independence assumption of the imputation model holds for many imputation methods, for example, parametric imputation models based on linear regression. However, it does not hold for hot-deck imputation, an imputation method commonly applied at statistical agencies.

7.7 DISCUSSION

DP is theoretically intuitive and elegant. It provides quantifiable and composable guarantees of privacy protection (although these guarantees have been subject to some confusion and misinterpretation (Tschantz et al., 2020)). By putting data privacy on a mathematical basis, it has supplied a calculus for reasoning about the protection offered by sophisticated data-release mechanisms.

Yet implementing DP mechanisms in practice often entails unforeseen complexities. In this paper we have focused on some of the complexities which arise in the context of survey data. Many of the same complexities can also emerge in settings with data preprocessing steps or with multistage data collection (such as national censuses). The goals of this paper are to draw attention to these complexities, review the current progress on addressing them, and spur renewed research activity to resolve those that remain outstanding.

Having identified a multiplicity of challenges in obtaining DP – some of which may be unduly constraining – we suggest that future research investigates pragmatic modifications to “completely-by-the-books” implementations of DP. The goals of such modifications should be: to provide a solution which

is feasible to implement; to retain the essence of DP even while not strictly satisfying DP; and to not unduly sacrifice the accuracy of the released data, nor the privacy of the data subjects, nor the resources of the statistical agency (in implementing the solution). Of course, any such modifications should be principled, in the sense that the associated risks to privacy are properly quantified and are outweighed by gains in data utility or implementability. Assessing the privacy risks of these modifications will likely involve a combination of theoretical and empirical analyses, and require measures of data privacy which lie outside the framework of DP.

An example of one possible modification is the non-DP publication of a data-dependent sampling design. A description of a survey's sampling design is crucial information for data users. Yet, as outlined in Subsection 7.2.2, if the sampling design was chosen with reference to the frame (as is often the case), then DP requires noise to be added to it before it can be published. Moreover, designing a DP mechanism to publish a sampling design will likely be difficult. On the other hand, it defies intuition that a simple description of a survey's sampling design should be disclosive of private information. This suggests it may be reasonable to modify the DP data-release mechanism, allowing the sampling design to be released exactly (i.e. without noise) even while the other outputs are protected in line with the exact requirements of DP. But to justify this pragmatic violation of DP, the statistical agency should first address the questions: Can the risks associated with publishing a sampling design be quantified (without resorting to DP)? And when is it principled (in the sense given in the previous paragraph) to publish a sampling design as is, without privacy protection?

PART IV

BROADER PERSPECTIVES ON STATISTICAL DATA PRIVACY

This page intentionally left blank.

8

The Five Safes as a Privacy Context¹

8.1 INTRODUCTION

AS A SUPPLIER OF OFFICIAL DATA, a national statistical office (NSO) is an integral part of a well-functioning democratic state. Its data are essential for informing government policy, business strategy and academic research, thereby advancing society and driving economic growth ([Lateral Economics, 2019](#)). Yet an NSO's ongoing value depends upon maintaining its social license to collect and share data. It is therefore necessary that NSOs balance their social and economic utility with the privacy of their data providers. With the recent increase in resources available to malicious actors and the growth of competing data vendors, this trade-off is increasingly difficult to manage ([Bailie, 2020](#)).

The Five Safes is one tool that assists NSOs in balancing privacy and utility. It is a conceptual framework for designing and assessing modes of statistical data sharing under privacy and confidentiality constraints.

¹Based on work coauthored with Ruobin Gong.

Originally developed in 2003 to enable researchers' access to Office of National Statistics (ONS) microdata (Desai et al., 2016), its use has since expanded to all forms of statistical dissemination (Australian Bureau of Statistics, 2021a). It has been employed by NSOs in the UK, Australia, and New Zealand to guide design decisions and risk assessments for statistical disclosure control (SDC) (Stokes, 2017; UK Data Service, 2023; Australian Bureau of Statistics, 2021a; Statistics New Zealand, 2022). In the USA, the Five Safes have been used by the Coleridge Initiative in the context of data sharing within and across states, government agencies and researchers (Foster et al., 2021). Further, the Advisory Committee on Data for Evidence recently recommended that the use of the Five Safes in the US federal bureaucracy be expanded (Advisory Committee on Data for Evidence Building, 2022).

We make two main points in this work. Firstly, the Five Safes is a reparametrization of contextual integrity (CI) in the situation where the information flow is a statistical dissemination. Section 8.3 describes the translation between the Five Safe parameters (people, projects, settings, data and outputs) and the CI parameters (sender, recipient, subject, information type and transmission principles). Therefore, the Five Safes provides specialized guidance as to how NSOs can satisfy contextual information norms. Moreover, by placing Five Safes within the CI theory, NSOs benefit from the extensive CI literature in justifying and understanding the Five Safes.

Secondly, as a framework for controlling the disclosure risk of statistical dissemination, the Five Safes provides a natural context for differential privacy (DP). As described in Section 8.4, DP is a broad collection of technical standards which all measure (in various ways) how a statistical dissemination can depend on the response of a data provider – or, in other words, how a data provider can influence the data being shared. Importantly, we argue that DP measures some – but, crucially, not all – of the dimensions relevant

for assessing the contextual integrity of statistical data sharing. We use the Five Safes, as a holistic risk assessment tool, to situate DP within the dimensions it partially measures (safe data and safe outputs) and to explicate the dimensions to which DP is agnostic.

Our contextualization of DP within the Five Safes is important for two reasons. Any implementation of DP requires choosing its various components (see Section 8.4). This choice depends on the broader context of the implementation, which the Five Safes explicate. Moreover, by placing DP within the Five Safes, NSOs can see how DP could be used to partially control safe data and safe outputs, and how DP can be traded-off against the other safes.

To recap, this work seeks to, firstly, situate the Five Safes framework within the broader concept of CI and, secondly, to contextualise DP via the Five Safes. This explains the dual meaning of this paper’s title: the Five Safes is a context for narrow, technical notions of privacy and security, like DP; yet it is also a specific context within the theory of CI.

8.2 THE FIVE SAFES AND THE INFORMATION FLOWS THEY GOVERN

In this Section, we give a brief review of the Five Safes framework, concentrating on the two information flows with which it is concerned. The major thesis of the Five Safes is that five dimensions of data access – *people*, *projects*, *settings*, *data*, and *outputs* – collectively determine the disclosure risk in statistical dissemination. These dimensions are related yet independent to one another. The safety of each of these dimensions can be measured on a continuous scale. In the design of a statistical dissemination paradigm, a data custodian strives to ensure data confidentiality by promoting safety of these five dimensions. However, doing so usually entails more work on the custodian’s part, including vetting, supervision, and higher degrees of infrastructure security and compliance monitoring. Viewing these five dimensions under a joint

framework allows a data custodian with a fixed amount of resources to focus on the safety of a subset of these dimensions, while maintaining control of the overall disclosure risk.

To explicate the meaning of the Five Safes, we begin with an examination of the two types of information flow that it is concerned with:

$$\text{data} \rightarrow \text{people (researcher)}, \quad (8.1)$$

$$\text{outputs} \rightarrow \text{people (general public)}. \quad (8.2)$$

The two types of information flow are neither independent nor mutually exclusive to one another. Flow (8.1) is the process through which the researcher learns from the data in the possession of the data custodian. Typically, the researcher takes the initiative to access the data, conducts analyses based on the data, and publishes a set of scientifically significant findings. These published results, alongside any open information required to support the verification of these results, make up the outputs that reach the general public as captured in Flow (8.2). Alternatively, Flow (8.2) may occur when the data custodian directly share information derived from their database with the general public, without involving researchers as an intermediary.

The term ‘researcher’ here refers to a person or an entity whose identity has been subject to some degree of vetting by the data custodian. This distinguishes a researcher from a member of the general public in our current discussion, even though in practice the two identities are not well separated. As such, a ‘safe’ researcher is someone who has demonstrated a good scientific standing as well as a commitment to data confidentiality and research ethics.

The ‘safe projects’ dimension concerns whether the intended use of the data is appropriate, ethical, and compliant with relevant legislation or regulations. The use of certain sensitive data may be restricted by

law to support independent scientific research only (Foster et al., 2021, Section 12.3). The data custodian would often also ascertain that their data is used toward the advancement of science, with clear and positive social benefits and in a manner consistent with modern scientific norms, including standards of reproducibility and knowledge sharing. In addition, the ‘safe settings’ refers to the security of the environment in which data access and sharing takes place, be they physical or virtual.

We illustrate how the two information flows interact under the Five Safes framework with three examples of data dissemination paradigms.

Example 8.2.1 (Public use data files/Open data). Statistical agencies publish public use data files for access by the general public and researchers alike. The agencies do not vet people who seek access because, by design, any person or entity without abusive intentions should be able to access the resource. A high level of scrutiny is placed on the data, which doubles as the output, to ensure that they are safe. On the other hand, since the data custodian cannot supervise the use of the data once it is made open access, no scrutiny is possible regarding the safety of the projects nor the settings in which the projects will be conducted.

Public use data files are frequently in the form of tabular data, which are highly aggregated from an underlying microdata to ensure adequate confidentiality. Public use microdata exist too, but they are often heavily subsampled. The U.S. Census Bureau curates the Public Use Microdata Sample (PUMS) based on a small sample (1% and 5%) of responses from the American Community Survey (US Census Bureau, 2023k). The PUMS files are available on the Census Bureau’s website and may be accessed via the file transfer protocol (FTP), a microdata analysis tool, or through an API provided by the Bureau.

Example 8.2.2 (Data Enclaves). Data enclaves are secure access environments through which authorized researchers can query the custodian’s database. Data enclaves provide a highly secure setting for data ac-

cess. Both the people and the projects seeking access are heavily vetted: only researchers who demonstrate legitimate scientific purposes of their inquiry and compliance with research ethics are allowed access. The outputs that the researcher is allowed to obtain and bring to outside of the data enclave is subject to various degrees of scrutiny. As a result, the data accessible through data enclaves can be detailed and comprehensive.

Data enclaves may be physical or virtual. A physical enclave is synonymous with a research data center (RDC), such as the Federal Statistical Research Data Center (FSRDC) of the U.S. Census Bureau. A virtual data enclave allows authorized researchers to access restricted-use data by logging into a secure, remote server. The DataLab of the Australian Bureau of Statistics (ABS) is an example of a virtual data enclave.² The reader is referred to (Australian Bureau of Statistics, 2021a) for further illustrations of dissemination paradigms discussed in Examples 8.2.1 and 8.2.2 and an analysis of important safety considerations that pertain to them.

Example 8.2.3 (Synthetic data with validation servers). The data custodian releases a synthetic dataset that resembles the underlying confidential dataset. Researchers who hold permission to access the synthetic dataset may use it to compose their desired statistical analysis including its code implementation. Then, they may *validate* the results with the data custodian who will run the analysis on the restricted-use dataset, and release the results to the researcher if they are deemed safe.

A prominent case of Example 8.2.3 is the Survey of Income and Program Participation (SIPP) Synthetic Beta (SSB) (US Census Bureau, 2022d). The SSB is synthesized by the U.S. Census Bureau through integrating nine annual SIPP panels between 1984 and 2008, together with the W-2 records from the Social

²Due to a lack of full oversight on the data access setting compared to physical data enclaves, statistical agencies debate the safety of virtual data enclaves (see e.g. Russell, 2022).

Security Administration (SSA) and Internal Revenue Service (IRS). Researchers whose proposed analysis is deemed appropriate and feasible by the Census Bureau are permitted to access the SSB. Prior to October 2022, access could be obtained via the Synthetic Data Server (SDS) hosted by Cornell University.³ Once the researcher composes a functional and correct statistical analysis program, they submit the code to the Bureau, who in turn performs the validation on the researcher's behalf on the Gold Standard File (GSF) which is internal to the Bureau and confidential. The output is subject to a stringent level of disclosure review similar to those applicable to the FSRDCs.

Example 8.2.3 is an interesting mode of statistical dissemination, through which we see a juxtaposition of safety levels pertaining to the two information flows. On the level of Flow (8.2) where the relevant people are the general public, the outputs are strictly scrutinized according to a high safety standard, rendering this setting similar to the open data mode discussed in Example 8.2.1. On the level of Flow (8.1), data consists of two distinct components, the SSB and the GSF, where the former commands a level of safety higher than the latter due to its synthetic nature. The people, here referring to the researchers, are placed under a moderate level of scrutiny. The setting, the Cornell SDS, is effectively a virtual enclave. These elements render this setting analogous to the data enclave mode discussed in Example 8.2.2.

8.3 THE FIVE SAFES AS A PRIVACY CONTEXT FOR STATISTICAL DISSEMINATION

Contextual integrity (CI) defines an information flow as private if it conforms with contextual informational norms (Nissenbaum, 2010). There are five parameters that define contextual informational norms: the *sender*, the *recipient*, the *subject*, the *information type*, and the *transmission principles* (Nissenbaum,

³Unfortunately, the Cornell server was shut down on September 30, 2022.

Privacy norm parameters	Their meanings in statistical dissemination
sender	statistical agencies/NSOs/data custodians
recipient	people : researchers (8.1) and general public (8.2)
subject	is a component of data (8.1)
information type	is a component of data (8.1) and outputs (8.2)
transmission principles	encompass projects, settings , and more

Table 8.1: The five contextual integrity parameters and their meanings in statistical dissemination, with a reference mapping to the Five Safes (in bold).

2019). Of the five parameters, the first two are straightforwardly understood: the sender is the person or entity who is sending the information, and the recipient is who is receiving the information. The latter three parameters require that we take a tailored approach to their explication by first situating this discussion in the context of statistical dissemination.

Claim 1. In the context of statistical dissemination, the Five Safes is an instantiation of a set of informational norms that govern privacy protection.

Table 8.1 outlines the meanings of the privacy contextual informational norm parameters as they apply to statistical dissemination. These meanings are explicated with reference to the five elements of the Five Safes framework. As the Table illustrates, the Five Safes is a *reparametrization* of CI – faithful albeit imperfect – when the information flow in question is a statistical dissemination. In particular, we note that the notion of ‘subject’ is a component of ‘data’ under the Five Safes, and the notion of ‘information type’ is a component of both data and outputs. On the other hand, the ‘transmission principles’ encompass multiple dimensions of the Five Safes, including projects, settings, and more. The nuances of these over- and under-inclusive mappings will be further explicated in future work.

8.4 DIFFERENTIAL PRIVACY IN THE CONTEXT OF THE FIVE SAFES

Differential privacy is a state-of-the-art technical formulation of privacy associated with statistical and data dissemination. It has been adopted by data agencies and intermediaries. Since the proposal of ϵ -DP (or pure DP) (Dwork et al., 2006b), a multiplicity of flavors of differential privacy has emerged, including probabilistic DP (Dwork et al., 2006a), approximate DP (Machanavajjhala et al., 2008), zero-Concentrated DP (Bun and Steinke, 2016), f -DP (Dong et al., 2022), to name a few. There are *bounded* versus *unbounded* versions of DP to suit the scenarios with known versus unknown dataset sizes. The TopDown algorithm (Abowd et al., 2022a), the U.S. Census Bureau’s differentially private disclosure avoidance system (DAS) for the 2020 Decennial Census Redistricting Data (P.L. 94-171) Summary and Demographic and Housing Characteristics Files, introduced the concept of *invariants* which are exact statistics of the confidential data that are mandated for release. To gain conceptual clarity amidst such plurality of choices, we employ the unified construction proposed by (Bailie et al., 2025d), which explicitly spells out the necessary elements of a *differential privacy definition*, some of which are often overlooked:

- *What* can be protected: \mathcal{D} , the data multiverse (consisting of multiple data universes \mathcal{D});
- *Who* are protected: $d_{\mathcal{X}}$, the input divergence;
- *How* to measure protection: $d_{\mathcal{T}}$, the output divergence;
- *How much* protection is afforded: ϵ , a privacy loss budget.

A data release mechanism can be most-generally defined as differentially private as follows.

Definition 8.4.1 (Definition 3.4 of (Bailie et al., 2025d)). A data-release mechanism $T: \mathcal{X} \times [0, 1] \rightarrow \mathcal{T}$

satisfies a differential privacy definition $(\mathcal{D}, d_{\mathcal{X}}, d_{\mathcal{T}})$ with privacy budget $\varepsilon_{\mathcal{D}} \geq 0$ if

$$d_{\mathcal{T}}[P_{\mathbf{X}}(T \in \cdot), P_{\mathbf{X}'}(T \in \cdot)] \leq \varepsilon_{\mathcal{D}} d_{\mathcal{X}}(\mathbf{X}, \mathbf{X}'),$$

for all \mathbf{X}, \mathbf{X}' in every data universe $\mathcal{D} \in \mathcal{D}$.

In this work we will not explicate Definition 8.4.1 in further detail, other than remark that it illustrates how the aforementioned elements of a differential privacy definition come together. Notably, the trio $(\mathcal{D}, d_{\mathcal{X}}, d_{\mathcal{T}})$ confer a quantitative description of the privacy guarantee (its *flavor*), whereas the privacy loss budget $\varepsilon_{\mathcal{D}}$ serves as a quantitative measurement (its *strength*). We will return to this definition later in this section.

In statistical dissemination, the Five Safes delineate a context in which differential privacy can be understood. What we mean is the following.

Claim 2. Differential privacy is a quantitative standard of safety pertaining to aspects of the *outputs* and the *data* in the Five Safes.

A quantified measurement of safety levels for aspects of the outputs and the data is helpful in the Five Safes framework, because it allows for the modulation of the various elements that collectively contribute to disclosure risk. The modulation may be achieved in multiple ways, two of which we discuss here.

First, differential privacy acts as a *screen* between the data and the people who access the data. By construction, differential privacy constrains the probabilistic properties of any output by a certified mechanism. This may be employed by the statistical agency to compose open data as well as to restrict researcher release from data enclaves or validation servers. Differential privacy may also directly restrict how the researcher may interact with the data in the first place. One example is differentially private synthetic data

(see e.g. Bowen et al., 2022). In the case of *local* privacy, the measurements taken from individual data contributors are privatized as soon as they leave the end device prior to arriving at the data custodian. In each of these cases, the differential privacy guarantee ascertains, in a mathematically rigorous way, a level of difficulty for adversarial agents to deduce the value of the confidential data. This enables, at least in a heuristic way, less scrutiny to be placed on the *people*.

Second, differential privacy enables precise privacy accounting by statistical agencies. For privacy definitions of the same *flavor* (i.e. the same \mathcal{D} , $d_{\mathcal{X}}$ and $d_{\mathcal{T}}$ choices), the privacy loss budgets of two mechanisms may be *composed* – often the case added – to yield the overall budget pertaining to both.⁴ What this means is that the data custodian with a fixed amount of privacy loss budget may choose to divide the budget across a number of *projects*, evenly or unevenly according to their significance, modulating the quantity-quality tradeoff in ways that the custodian sees fit.

We also observe two limitations of differential privacy as a quantitative standard of safety.

Remark 8.4.2. Differential privacy is silent on the safety of certain aspects of the outputs and the data.

Differential privacy specifies the flavor, as it does the strength, of the privacy protection. It is, however, agnostic to the *nature* of the data at hand. A differential privacy mechanism will treat two datasets identically so long as they possess identical mathematical structures, even though one may be highly sensitive in nature (e.g. records of patients suffering from a socially stereotyped disease) while the other is not (e.g. a log of dairy preferences of customers at a coffee shop). Therefore, differential privacy should not be taken as a comprehensive quantification for the safety of the outputs and the data.

⁴Composition is not possible within approaches to privacy that are not formal. It may well be the case that the combined disclosure risk of two data products is infinite, even though either carries a finite risk.

Remark 8.4.3. Differential privacy does not purport an assessment of safety for people, projects, or settings.

Differential privacy is a property of the output rather than the process through which it is generated. As such, it is agnostic to the settings in which privatization and data sharing takes place. Indeed, one of the celebrated feature of differential privacy is that the privacy mechanism can be entirely transparent without sabotaging the privacy guarantee.

For similar reasons, differential privacy does not directly measure the safety of the people and the projects. When used as a standard to quantify the safety of aspects of the data and the outputs, however, it may serve as an indirect guidance on the safety tuning for the people and for the projects. In fact, such tuning often constitute a balancing act between privacy and *utility*.

8.5 ONGOING INQUIRIES

What we have presented so far is work in progress. In this last Section, we briefly discuss three important questions that remain unanswered.

First, we view the Five Safes as a reparametrization of contextual integrity when the information flow in question is statistical dissemination. As Section 8.3 discusses, this reparametrization is faithful but *imperfect*, in the sense that the mappings between the two frameworks can seem either narrow or capacious, as summarized in Table 8.1. In future work, we aim to supplant the meanings that are lost in this translation, in particular by spelling out aspects of data and outputs (from the Five Safes) that go beyond the subject and the information type (from the privacy norms), as well as aspects of transmission principles that go beyond the dimensions of the Five Safes.

Second, as we have already argued, differential privacy offers a strong technical notion for assessing cer-

tain aspects of the ‘safe data’ and ‘safe output’ dimensions. Does there exist more comprehensive technical notions to these dimensions, and does there exist technical notions for the other dimensions of Five Safes? We surmise that a pursuit towards technicalization may not make sense universally. After all, some of the safety dimensions (such as safe people) may be too complex to allow mathematical tractability. However, others – such as safe settings – may benefit from advances in fields such as information security.

Third, we look to further explore the interaction between the Five Safes and the legal frameworks governing the operations of the NSOs. For starters we ask: what configurations of the Five Safes best correspond to the current legal framework for a specific NSO? And how can the Five Safes be used to update the legal framework in the future?

APPENDICES

This page intentionally left blank.



Appendices to Chapter 2

A.1 THE MEASURE-THEORETIC DEFINITION OF A DATA-RELEASE MECHANISM T

In this section, we formally define a data-release mechanism and the distribution of its output. Throughout this section, fix a DP flavor $(\mathcal{X}, \mathcal{D}, d_{\mathcal{X}}, D_{\text{Pr}})$.

Definition A.1.1. Two protection objects $x, x' \in \mathcal{X}$ are *(explicitly)-comparable* if 1) $x \neq x'$; 2) $d_{\mathcal{X}}(x, x') < \infty$ or $d_{\mathcal{X}}(x', x) < \infty$; and 3) there exists a universe $\mathcal{D} \in \mathcal{D}$ such that $x, x' \in \mathcal{D}$.

Comparability defines a symmetric relation on \mathcal{X} , which can be understood as an undirected graph. Let $[x]$ denote the connected component of $x \in \mathcal{X}$ in this graph. Two protection objects $x, x' \in \mathcal{X}$ are *implicitly-comparable* if they belong to the same connected component – that is, if $x' \in [x]$.

Definition A.1.2. A *data-release mechanism* is a function $T: \mathcal{X} \times \mathcal{U} \rightarrow \mathcal{T}$, together with a probability space $(\mathcal{U}, \mathcal{F}_{\mathcal{U}}, \mathbb{P}(U \in \cdot))$ for the random seed U and a σ -algebra $\mathcal{F}_{\mathcal{T}}^{[x]}$ on \mathcal{T} for each connected component

$[x]$, such that the map $T(x, \cdot)$ given by

$$(\mathcal{U}, \mathcal{F}_{\mathcal{U}}) \rightarrow (\mathcal{T}, \mathcal{F}_{\mathcal{T}}^{[x]})$$

$$u \mapsto T(x, u),$$

is $(\mathcal{F}_{\mathcal{U}}, \mathcal{F}_{\mathcal{T}}^{[x]})$ -measurable, for all $x \in \mathcal{X}$.

Hence a data-release mechanism is a three-tuple $\left(T : \mathcal{X} \times \mathcal{U} \rightarrow \mathcal{T}, (\mathcal{U}, \mathcal{F}_{\mathcal{U}}, \mathbf{P}(U \in \cdot)), \{\mathcal{F}_{\mathcal{T}}^{[x]}\}_{x \in \mathcal{X}}\right)$.

Usually we will refer to a data-release mechanism by the function T alone, and take the other two components as implicitly given.¹

The random seed U induces a probability on the output of a data-release mechanism in the standard (push-forward) way:

Definition A.1.3. Given $x \in \mathcal{X}$, the *distribution \mathbf{P}_x of a data-release mechanism $T(x, U)$* is the probability on the measurable space $(\mathcal{T}, \mathcal{F}_{\mathcal{T}}^{[x]})$ given by

$$\mathbf{P}_x(T(x, U) \in S) = \mathbf{P}(U \in \{u \in \mathcal{U} : T(x, u) \in S\}),$$

for all $S \in \mathcal{F}_{\mathcal{T}}^{[x]}$.

The σ -algebras $\mathcal{F}_{\mathcal{T}}^{[x]}$ should not be freely chosen, since they can be manipulated to artificially reduce privacy loss. For an extreme example, if $\mathcal{F}_{\mathcal{T}}^{[x]}$ is the trivial σ -algebra $\{\emptyset, \mathcal{T}\}$, then $T(x, \cdot)$ is always measurable and \mathbf{P}_x does not vary with x (or with T). Hence $D_{\text{Pr}}(\mathbf{P}_x, \mathbf{P}_{x'}) = 0$ for all $x, x' \in [x]$. This implies that perfect privacy ($\varepsilon = 0$) can be achieved, regardless of the behavior of the data-release mechanism T . More

¹A data-release mechanism can be considered as a generalisation of a Markov kernel, since we do not require that the σ -algebra on \mathcal{T} is fixed and that $x \mapsto \mathbf{P}_x(B)$ is measurable for all $B \in \mathcal{F}_{\mathcal{T}}^{[x]}$.

generally, one can coarsen $\mathcal{F}_{\mathcal{T}}^{[x]}$ to reduce $D_{\text{Pr}}(\mathbf{P}_x, \mathbf{P}_{x'})$ by removing sets S from $\mathcal{F}_{\mathcal{T}}^{[x]}$ on which \mathbf{P}_x and $\mathbf{P}_{x'}$ differ.

To avoid the possibility of such manipulation, we recommend setting $\mathcal{F}_{\mathcal{T}}^{[x]}$ to be the σ -algebra induced by the data-release mechanism's output:

Definition A.1.4. Given a function $T : \mathcal{X} \times \mathcal{U} \rightarrow \mathcal{T}$ and a probability space $(\mathcal{U}, \mathcal{F}_{\mathcal{U}}, \mathbf{P}(U \in \cdot))$, define the σ -algebra

$$\mathcal{F}_{\mathcal{T}}^x = \{S \subset \mathcal{T} : \{u \in \mathcal{U} : T(x, u) \in S\} \in \mathcal{F}_{\mathcal{U}}\}.$$

The σ -algebra induced by T and $[x]$ is given by

$$\mathcal{F}_{\mathcal{T}}^{T, [x]} = \bigcap_{x' \in [x]} \mathcal{F}_{\mathcal{T}}^{x'}.$$

In the above definition, the set $\{u \in \mathcal{U} : T(x, u) \in S\}$ is the inverse image of S under the map $T(x, \cdot)$. This means $\mathcal{F}_{\mathcal{T}}^x$ is the largest σ -algebra such that $T(x, \cdot)$ is measurable and thus $\mathcal{F}_{\mathcal{T}}^{T, [x]}$ is the largest σ -algebra such that $T(x', \cdot)$ is measurable for all $x' \in [x]$.

Larger $\mathcal{F}_{\mathcal{T}}^{[x]}$ will typically result in larger values of $D_{\text{Pr}}(\mathbf{P}_x, \mathbf{P}_{x'})$. This is why we recommend using $\mathcal{F}_{\mathcal{T}}^{T, [x]}$. Since $\mathcal{F}_{\mathcal{T}}^{T, [x]}$ depends upon the choice of $\mathcal{F}_{\mathcal{U}}$ – with smaller $\mathcal{F}_{\mathcal{U}}$ resulting in smaller $\mathcal{F}_{\mathcal{T}}^{T, [x]}$ – the choice of $\mathcal{F}_{\mathcal{U}}$ is also important. Typically the random seed U is uniformly distributed on $[0, 1]$, equipped with the Borel σ -algebra. (Theoretically this is sufficient for most purposes, since such U can generate countably many independent real-valued random variables of arbitrary distributions.) If $T(x', \cdot)$ is continuous (for all $x' \in [x]$) and surjective (for some $x' \in [x]$), then $\mathcal{F}_{\mathcal{T}}^{T, [x]}$ is the Borel σ -algebra on \mathcal{T} .

A.2 WHAT CAN WE SAY ABOUT THE BUDGET?

Proposition A.2.1. *Suppose that $D_{\text{Pr}}(\mathbf{P}, \mathbf{Q}) = \infty$ whenever $d_{\text{TV}}(\mathbf{P}, \mathbf{Q}) = 1$. Then, for any (non-random)*

data-release mechanism $f: \mathcal{X} \rightarrow \mathcal{T}$, the following statements are equivalent:

- (A) *f satisfies $\varepsilon_{\mathcal{D}}\text{-DP}(\mathcal{X}, \mathcal{D}, d_{\mathcal{X}}, D_{\text{Pr}})$ with finite budget $\varepsilon_{\mathcal{D}} < \infty$; and*
- (B) *f is a function of $[x]$.*

Moreover, if these statements hold then f has zero privacy loss.

The assumption in Proposition A.2.1 (that $d_{\text{TV}}(\mathbf{P}, \mathbf{Q}) = 1$ implies $D_{\text{Pr}}(\mathbf{P}, \mathbf{Q}) = \infty$) is only used in proving (A) implies (B). The other direction of Proposition A.2.1 provides the following corollary.

Corollary A.2.2. *For any DP specification $\varepsilon_{\mathcal{D}}\text{-DP}(\mathcal{X}, \mathcal{D}, d_{\mathcal{X}}, D_{\text{Pr}})$, the (non-random) data-release mechanism*

$$b(x) = \{\varepsilon_{\mathcal{D}} : \mathcal{D} \in \mathcal{D} \text{ such that } x' \in \mathcal{D} \text{ for some } x' \in [x]\},$$

satisfies $\varepsilon'_{\mathcal{D}}\text{-DP}(\mathcal{X}, \mathcal{D}, d_{\mathcal{X}}, D_{\text{Pr}})$ with $\varepsilon'_{\mathcal{D}} = 0$.

The function $\varepsilon_{\mathcal{D}}$ can always be released without privacy loss, since the constant data-release mechanism $x \mapsto \varepsilon_{\mathcal{D}}$ is trivially a function of $[x]$. (In the literature, the function $\varepsilon_{\mathcal{D}}$ has been called the “policy” to distinguish it from the actual privacy loss budget associated to a particular universe \mathcal{D} (Seeman et al., 2023).) The above corollary describes what the data custodian can say about the privacy loss budgets associated to the true confidential dataset x^* . The data custodian can release the set of budgets for all universes containing some x' which is implicitly comparable to x^* (Definition A.1.1). Furthermore, under the assumption of Proposition A.2.1, the data custodian cannot be more specific about the budgets associated to x^* without adding some noise to the answer. For example, the data-release mechanism

$$b'(x) = \{\varepsilon_{\mathcal{D}} : \mathcal{D} \in \mathcal{D} \text{ such that } x \in \mathcal{D}\},$$

has infinite privacy loss unless $b(x) = b'(x)$ for all $x \in \mathcal{X}$. However, $b(x)$ has cardinality at most one, and hence $b(x) = b'(x)$, for all $x \in \mathcal{X}$ if either of the following two conditions hold: I) $\mathcal{D} \cap \mathcal{D}' = \emptyset$ for all distinct universes $\mathcal{D}, \mathcal{D}' \in \mathcal{D}$; or, more generally, II) for all distinct $x, x' \in \mathcal{X}$, if $d_{\mathcal{X}}(x, x')$ or $d_{\mathcal{X}}(x', x)$ is finite, then x and x' cannot both be in two different universes. (To prove this, observe that condition II implies $[x] \subset \mathcal{D}$ for some $\mathcal{D} \in \mathcal{D}$, assuming $\mathcal{D} \neq \emptyset$.) In particular, for an invariant-induced multiverse \mathcal{D}_c (defined below (2.4)), condition I holds and so the data custodian can release the realised budget $\varepsilon_{\mathcal{D}_c(x^*)}$ without incurring additional privacy loss.

A.3 CONNECTIONS BETWEEN THE INPUT PREMETERIC $d_{\mathcal{X}}$ AND THE MULTIVERSE \mathcal{D}

If the privacy loss budget $\varepsilon_{\mathcal{D}}$ is constant in \mathcal{D} , the effects of the multiverse \mathcal{D} can be encoded into a DP specification using only the input premetric $d_{\mathcal{X}}$. More exactly, by redefining $d_{\mathcal{X}}$ we can remove the multiverse \mathcal{D} without affecting the DP guarantee, as the following proposition demonstrates.

Proposition A.3.1. *Given a DP flavor $(\mathcal{X}, \mathcal{D}, d_{\mathcal{X}}, D_{\text{Pr}})$, define*

$$d'_{\mathcal{X}}(x, x') = \begin{cases} 0 & \text{if } x = x', \\ d_{\mathcal{X}}(x, x') & \text{else if there exists } \mathcal{D} \in \mathcal{D} \text{ such that } x, x' \in \mathcal{D}, \\ \infty & \text{otherwise.} \end{cases}$$

Then, for any constant privacy loss budget $\varepsilon_{\mathcal{D}} = \varepsilon$,

$$\mathcal{M}(\mathcal{X}, \mathcal{D}, d_{\mathcal{X}}, D_{\text{Pr}}, \varepsilon_{\mathcal{D}}) = \mathcal{M}(\mathcal{X}, \{\mathcal{X}\}, d'_{\mathcal{X}}, D_{\text{Pr}}, \varepsilon_{\mathcal{D}}).$$

The notation $\varepsilon_{\mathcal{D}}$ is slightly overloaded in the above proposition. We use it denote both the constant

function with domain \mathcal{D} in the context of $\mathcal{M}(\mathcal{X}, \mathcal{D}, d_{\mathcal{X}}, D_{\text{Pr}}, \varepsilon_{\mathcal{D}})$ and also the constant function with domain $\{\mathcal{X}\}$ in the context of $\mathcal{M}(\mathcal{X}, \{\mathcal{X}\}, d_{\mathcal{X}}, D_{\text{Pr}}, \varepsilon_{\mathcal{D}})$.

The above proposition shows how to translate a multiverse \mathcal{D} into an input premetric $d_{\mathcal{X}}$, provided that the privacy loss budget is constant. In the other direction, an input premetric may introduce its own universes. We can do this trivially by setting \mathcal{D} to be the set of pairs $\{x, x'\}$ with finite $d_{\mathcal{X}}(x, x')$:

Proposition A.3.2. *Given a premetric $d_{\mathcal{X}}$, define*

$$\mathcal{D} = \{\{x, x'\} : x, x' \in \mathcal{X} \text{ such that } d_{\mathcal{X}}(x, x') < \infty\}.$$

Then, for all constant privacy loss budgets $\varepsilon_{\mathcal{D}} = \varepsilon$,

$$\mathcal{M}(\mathcal{X}, \{\mathcal{X}\}, d_{\mathcal{X}}, D_{\text{Pr}}, \varepsilon_{\mathcal{D}}) = \mathcal{M}(\mathcal{X}, \mathcal{D}, d_{\mathcal{X}}, D_{\text{Pr}}, \varepsilon_{\mathcal{D}}).$$

More generally,

$$\mathcal{M}(\mathcal{X}, \mathcal{D}_0, d_{\mathcal{X}}, D_{\text{Pr}}, \varepsilon_{\mathcal{D}}) = \mathcal{M}(\mathcal{X}, \mathcal{D}_1, d_{\mathcal{X}}, D_{\text{Pr}}, \varepsilon'_{\mathcal{D}'}),$$

for any multiverse \mathcal{D}_0 and any budget $\varepsilon_{\mathcal{D}} : \mathcal{D}_0 \rightarrow [0, \infty]$, where $\mathcal{D}_1 = \{\mathcal{D}_0 \cap \mathcal{D} : \mathcal{D}_0 \in \mathcal{D}_0, \mathcal{D} \in \mathcal{D} \text{ such that } \mathcal{D}_0 \cap \mathcal{D} \neq \emptyset\}$ and $\varepsilon'_{\mathcal{D}'} = \inf\{\varepsilon_{\mathcal{D}} : \mathcal{D} \in \mathcal{D}_0 \text{ with } \mathcal{D}' \subset \mathcal{D}\}$.

The above proposition shows that we can always restrict the DP specification to universes of pairs of protection objects. However, interpreting the resulting multiverse is not easy because it is possible (and in practice, quite likely) that the universes overlap. As a result, there are implicit Lipschitz constraints which are implied by the restricted DP specification $\varepsilon'_{\mathcal{D}'}\text{-DP}(\mathcal{X}, \mathcal{D}_1, d_{\mathcal{X}}, D_{\text{Pr}})$ but not explicitly encoded in that specification. For example, suppose there exists $x_1, x_2, x_3 \in \mathcal{X}$ such that 1) $d_{\mathcal{X}}(x_1, x_3) = d_{\mathcal{X}}(x_1, x_2) + d_{\mathcal{X}}(x_2, x_3) < \infty$; 2) x_1 and x_2 are in some universe $\mathcal{D} \in \mathcal{D}_1$; 3) x_2 and x_3 are in another universe $\mathcal{D}' \in \mathcal{D}_1$;

but 4) no universe in \mathcal{D}_1 contains both x_1 and x_3 . Then

$$D_{\text{Pr}}(\mathbf{P}_{x_1}, \mathbf{P}_{x_3}) \leq \max(\varepsilon'_{\mathcal{D}}, \varepsilon'_{\mathcal{D}'}) d_{\mathcal{X}}(x_1, x_3), \quad (\text{A.1})$$

is an implicit Lipschitz condition of $\varepsilon'_{\mathcal{D}'}$ -DP($\mathcal{X}, \mathcal{D}_1, d_{\mathcal{X}}, D_{\text{Pr}}$) (assuming D_{Pr} satisfies the triangle inequality). But this condition is not one of the conditions explicitly set down by this DP specification.

Instead, it is more informative to consider the multiverse which is induced by the invariants of $d_{\mathcal{X}}$. An invariant-induced multiverse always partition \mathcal{X} , so its universes never overlap. (Assuming $d_{\mathcal{X}}$ satisfies the triangle inequality) if the universes in \mathcal{D} do not overlap, then there are no implicit Lipschitz conditions (like (A.1)) hidden in a DP specification $\varepsilon_{\mathcal{D}}$ -DP($\mathcal{X}, \mathcal{D}, d_{\mathcal{X}}, D_{\text{Pr}}$). Therefore, examining the multiverse induced by $d_{\mathcal{X}}$'s invariants provides a clearer protection of how choices of $d_{\mathcal{X}}$ can reduce the protection provided by DP. Proposition A.3.1 showed that $d_{\mathcal{X}}$ can encode arbitrary invariants. The following proposition describes the invariants encoded by an arbitrary $d_{\mathcal{X}}$. As such, we see that $d_{\mathcal{X}}$ can reduce the protection by DP in exactly the same way as invariants. The intuition is that an input premetric $d_{\mathcal{X}}$'s invariants are given by the connected components of the graph (\mathcal{X}, \sim) where \sim is the relation defined by:

$$x \sim x' \text{ if } d_{\mathcal{X}}(x, x') < \infty.$$

Proposition A.3.3. *Given a premetric $d_{\mathcal{X}}$, define the relation \sim by $x \sim x'$ if $d_{\mathcal{X}}(x, x') < \infty$. Let \sim_{cl} be the symmetric- and transitive-closure of \sim . Define the invariant $c(x) = [x]$ which sends $x \in \mathcal{X}$ to its equivalence class $[x]$ under \sim_{cl} . Then, for all constant privacy loss budgets $\varepsilon_{\mathcal{D}} = \varepsilon$,*

$$\mathcal{M}(\mathcal{X}, \{\mathcal{X}\}, d_{\mathcal{X}}, D_{\text{Pr}}, \varepsilon_{\mathcal{D}}) = \mathcal{M}(\mathcal{X}, \mathcal{D}_c, d_{\mathcal{X}}, D_{\text{Pr}}, \varepsilon_{\mathcal{D}}).$$

More generally,

$$\mathcal{M}(\mathcal{X}, \mathcal{D}_0, d_{\mathcal{X}}, D_{\text{Pr}}, \varepsilon_{\mathcal{D}}) = \mathcal{M}(\mathcal{X}, \mathcal{D}_1, d_{\mathcal{X}}, D_{\text{Pr}}, \varepsilon_{\mathcal{D}}),$$

for any multiverse \mathcal{D}_0 and budget $\varepsilon_{\mathcal{D}} : \mathcal{D}_0 \rightarrow [0, \infty]$, where $\mathcal{D}_1 = \{\mathcal{D}_0 \cap \mathcal{D} : \mathcal{D}_0 \in \mathcal{D}_0, \mathcal{D} \in \mathcal{D}_c \text{ such that } \mathcal{D}_0 \cap \mathcal{D} \neq \emptyset\}$ and $\varepsilon'_{\mathcal{D}} = \inf\{\varepsilon_{\mathcal{D}} : \mathcal{D} \in \mathcal{D}_0 \text{ with } \mathcal{D}' \subset \mathcal{D}\}$.

Moreover, suppose that there exists some $\mathcal{D}_0 \in \mathcal{D}_0$ such that one of the connected components of the graph (\mathcal{D}_0, \sim_d) is not in \mathcal{D}_0 . Then $\mathcal{D}_1 \neq \mathcal{D}_0$.

The above proposition proves that, more than just trivially inducing universes of size two, the input pre-metric $d_{\mathcal{X}}$ can implicitly induce invariants. Specifically, whenever the graph (\mathcal{X}, \sim) is not fully connected, $d_{\mathcal{X}}$ induces invariants because

$$\mathcal{M}(\mathcal{X}, \{\mathcal{X}\}, d_{\mathcal{X}}, D_{\text{Pr}}, \varepsilon_{\mathcal{D}}) = \mathcal{M}(\mathcal{X}, \mathcal{D}_c, d_{\mathcal{X}}, D_{\text{Pr}}, \varepsilon_{\mathcal{D}}),$$

with $\mathcal{D}_c \neq \{\mathcal{X}\}$. If x, x' are not in the same universe of \mathcal{D}_c then the DP specification $\varepsilon_{\mathcal{D}}\text{-DP}(\mathcal{X}, \{\mathcal{X}\}, d_{\mathcal{X}}, D_{\text{Pr}})$ places no restrictions on $D_{\text{Pr}}(P_x, P_{x'})$, explicitly or implicitly. Hence, the invariants of $d_{\mathcal{X}}$ are non-trivial limitations to the protection provided by $\varepsilon_{\mathcal{D}}\text{-DP}(\mathcal{X}, \mathcal{D}, d_{\mathcal{X}}, D_{\text{Pr}})$.

A.4 COMMON CHOICES FOR THE INPUT PREMETRIC $d_{\mathcal{X}}$

A direct encoding of many common DP formulations in the literature – including the original Definition 2.3.1 of [Dwork et al. \(2006b\)](#) – into the equivalent DP specifications would use the following premet-

ric for $d_{\mathcal{X}}$:

$$d_r(x, x') = \begin{cases} 0 & \text{if } x = x', \\ 1 & \text{else if } x \overset{r}{\sim} x', \\ \infty & \text{otherwise,} \end{cases} \quad (\text{A.2})$$

where the relation r on \mathcal{X} captures some notion of ‘neighboring’ datasets: $x \overset{r}{\sim} x'$ if and only if x and x' are ‘neighbors.’ There are multiple different relations r used in the DP literature but they are all formalizations of the following intuitive definition: Datasets x and x' are neighbors – i.e. $x \overset{r}{\sim} x'$ – if x and x' differ only by a single record. Such a definition of a neighboring relation r on \mathcal{X} requires that $\mathcal{X} \subset \bigcup_{n=0}^{\infty} \mathcal{R}^n$, where \mathcal{R} is the set of all theoretically-possible records (and \mathcal{R}^n is the n -fold cartesian product of \mathcal{R}).

Once \mathcal{R} has been fixed, there are two common choices for r :

- A) *Bounded* DP: x and x' are neighbors if x and x' have the same number of records and exactly one record is different in x as compared to x' :

$$x \overset{r_{bv}}{\sim} x' \text{ if } |x| = |x'| \text{ and } \frac{1}{2}|x \ominus x'| = 1, \quad (\text{for unordered datasets } x, x')$$

or

$$x \overset{r_{bv}}{\sim} x' \text{ if } |x| = |x'| \text{ and there exists a unique } j \in \{1, \dots, |x|\} \text{ such that } x_j \neq x'_j, \quad (\text{for ordered datasets } x, x')$$

(where \ominus is the symmetric set difference).

- B) *Unbounded* DP: x and x' are neighbors if x' can be formed by adding or subtracting a single record from x :

$$x \overset{r_{us}}{\sim} x' \text{ if } |x \ominus x'| = 1, \quad (\text{for unordered datasets } x, x')$$

or

$$x \overset{r_{us}}{\sim} x' \text{ if there exists some } j \in \{1, \dots, \max(|x|, |x'|)\} \text{ such that } x = x'_{-j} \text{ or } x_{-j} = x', \quad (\text{for ordered datasets } x, x')$$

(where the notation v_{-j} denotes the vector $(v_1, \dots, v_{j-1}, v_{j+1}, \dots, v_n)$ where the j -th element of v has been removed).

In the literature, DP definitions which use r_{bs} or r_{bv} are referred to as *bounded* DP and those which use r_{us} or r_{uv} are called *unbounded*. These terms arise from the fact that r_{us} and r_{uv} relates datasets of differing length while neighbors under r_{bs} and r_{bv} must have the same length. The distinction between bounded and unbounded DP is important because bounded DP reduces the number of implicitly-comparable datasets (Definition A.1.1.1). Reducing the set of implicitly-comparable datasets is equivalent to partitioning the universes into finer universes and affects inference for both the attacker and the legitimate analyst (Bailie and Drechsler, 2024; Bailie and Gong, 2023a).

It is also important to distinguish between the neighbor relations for unordered and ordered datasets. A DP flavor for unordered datasets is stronger than the corresponding flavor for ordered datasets – that is,

$$\mathcal{M}(\mathcal{X}, \mathcal{D}, d_{r_{bs}}, D_{\text{Pr}}, \varepsilon_{\mathcal{D}}) \subset \mathcal{M}(\mathcal{X}, \mathcal{D}, d_{r_{bv}}, D_{\text{Pr}}, \varepsilon_{\mathcal{D}})$$

and the same result holds when $d_{r_{bs}}$ and $d_{r_{bv}}$ are replaced with $d_{r_{us}}$ and $d_{r_{uv}}$ respectively. (These results follow by Proposition 2.4.15.) The other direction does not hold; for example, local DP mechanisms (such as randomized response) satisfy the ordered DP specification $\varepsilon\text{-DP}(\mathcal{X}, \{\mathcal{X}\}, d_{bv}, D_{\text{MULT}})$ (for the appropriate choice of \mathcal{X} and ε), but they do not satisfy the corresponding unordered DP specification $\varepsilon\text{-DP}(\mathcal{X}, \{\mathcal{X}\}, d_{bs}, D_{\text{MULT}})$ for any finite ε (assuming that $\mathcal{X} = \bigcup_{n=1}^{\infty} \mathcal{R}^n$). For this reason, ordered DP specifications should only be used when the ordering of the dataset is meaningful to an attacker (for example, if the indices of the dataset are pseudo-identifying, which would imply that, in the counterfactual scenario where a respondent changes their data, they keep the same index).

A premetric d_r built from a relation r as in (A.2) can be sharpened to a metric d_r^* as follows. Here $d_r^*(x, x')$ is defined as the length of a shortest path between x and x' in the graph on \mathcal{X} with edges given by r :

Definition A.4.1. Given a domain \mathcal{X} and a relation r on \mathcal{X} , let the graph $G = (V, E)$ have vertices $V = \mathcal{X}$ and edges

$$E = \{(x, x') \in \mathcal{X}^2 \mid x \overset{r}{\sim} x'\}.$$

Define $d_r^*(x, x')$ to be the length of a shortest path between x and x' in G .

More generally, for a premetric $d_{\mathcal{X}}$ on \mathcal{X} , let the weighted and directed graph $G = (V, E, w)$ have vertices $V = \mathcal{X}$ and edges

$$E = \{(x, x') \in \mathcal{X}^2 \mid d_{\mathcal{X}}(x, x') < \infty\}$$

and weights

$$w(x, x') = d_{\mathcal{X}}(x, x').$$

Define $d_{\mathcal{X}}^*(x, x')$ be the length of a shortest (weighted and directed) path between x and x' in G .

For example, the metrics for the above neighbor relations r_{bs} , r_{bv} , r_{us} and r_{uv} are given by:

A1) The sharpening of $d_{r_{bs}}$ (bounded, unordered neighbors) is the Hamming distance on unordered datasets:

$$d_{\text{HamS}}^u(x, x') = \begin{cases} \frac{1}{2}|x \ominus x'| & \text{if } |x| = |x'|, \\ \infty & \text{otherwise.} \end{cases} \quad (\text{A.3})$$

A2) The sharpening of $d_{r_{bv}}$ (bounded, ordered neighbors) is the Hamming distance

$$d_{\text{Ham}}^u(x, x') = \begin{cases} \sum_{i=1}^n \mathbb{I}\{x_i \neq x'_i\} & \text{if } |x| = |x'| = n, \\ \infty & \text{otherwise.} \end{cases}$$

B1) The sharpening of $d_{r_{us}}$ (unbounded, unordered neighbors) is the symmetric difference distance:

$$d_{\text{SymDiff}}^u(x, x') = |x \ominus x'|. \quad (\text{A.4})$$

B2) The sharpening of $d_{r_{uv}}$ (unbounded, ordered neighbors) is

$$d_{r_{uv}}^{u*}(x, x') = \min\{|I| + |J| : I \subset \{1, \dots, |x|\}, J \subset \{1, \dots, |x'|\}, x_{-I} = x'_{-J}\},$$

(where, for $I \subset \{1, 2, \dots, |v|\}$, the notation v_{-I} denotes the vector $(v_i : i \notin I)$ where, for every $i \in I$, the i -th element of v has been removed).

The superscript u emphasizes that these distances are defined with respect to a choice of resolution u . Each choice of resolution u defines a different version of the Hamming distance d_{HamS}^u (and different versions of d_{Ham}^u , d_{SymDiff}^u and $d_{r_{uv}}^{u*}$), since the resolution defines the elements of the *multi-set* x and hence the operation \ominus . (This also applies for the premetrics r_{bs} , r_{bv} , r_{us} and r_{uv} – they are only well-defined up to the choice the unit, although this is not made explicit in their notation.) In this article, this distinction is important since we use d_{HamS}^p and d_{HamS}^b for persons and household records.

Under mild assumptions, we have that

$$\mathcal{M}(\mathcal{X}, \{\mathcal{X}\}, d_r, D_{\text{Pr}}, \varepsilon) = \mathcal{M}(\mathcal{X}, \{\mathcal{X}\}, d_r^*, D_{\text{Pr}}, \varepsilon),$$

if and only if D_{Pr} is a metric (Bailie et al., 2025a). That is, one can equivalent use d_r or the associated metric d_r^* whenever D_{Pr} is a metric. Therefore, while a direct encoding of a typical DP formulation would set $d_{\mathcal{X}}$ to be the premetric d_r as in (A.2), the sharpening d_r^* could also be used (as long as D_{Pr} is a metric). (Another corollary of this result is that group privacy with linear decrease in privacy loss is equivalent to D_{Pr} being a metric (Bailie et al., 2025a).)

Another common choice of $d_{\mathcal{X}}$ is to encode invariants, or more generally, to encode the multiverse: For any DP specification $\varepsilon_{\mathcal{D}}\text{-DP}(\mathcal{X}, \mathcal{D}, d_{\mathcal{X}}, D_{\text{Pr}})$ where the budget $\varepsilon_{\mathcal{D}} = \varepsilon$ is constant in \mathcal{D} ,

$$\mathcal{M}(\mathcal{X}, \mathcal{D}, d_{\mathcal{X}}, D_{\text{Pr}}, \varepsilon_{\mathcal{D}}) = \mathcal{M}(\mathcal{X}, \{\mathcal{X}\}, d_{\mathcal{X}}^{\mathcal{D}}, D_{\text{Pr}}, \varepsilon),$$

where

$$d_{\mathcal{X}}^{\mathcal{D}}(x, x') = \begin{cases} d_{\mathcal{X}}(x, x') & \text{if there exists } \mathcal{D} \in \mathcal{D} \text{ with } x, x' \in \mathcal{D}, \\ \infty & \text{otherwise.} \end{cases}$$

Thus, the actual advantages of the multiverse \mathcal{D} are that 1) it allows for the possibility of having different privacy budgets for different universes (as in [Bun et al. \(2022\)](#) and [Seeman et al. \(2023\)](#)); 2) because \mathcal{D} and $d_{\mathcal{X}}$ perform separate functions (as the scope of protection and the unit of protection respectively), it is more explainable and interpretable to keep these functions separate. Setting aside these two advantages, one can always encode the function of the multiverse \mathcal{D} using only the input premetric $d_{\mathcal{X}}$ when $\varepsilon_{\mathcal{D}}$ is constant.

In the other direction, having unconnected protection objects $x, x' \in \mathcal{X}$ (such as in bounded DP) is equivalent to employing a non-totally-vacuous multiverse. (Given a DP flavor $(\mathcal{X}, \mathcal{D}, d_{\mathcal{X}}, D_{\text{Pr}})$, two protection objects $x, x' \in \mathcal{X}$ are *unconnected* if $d_{\mathcal{X}}^*(x, x') = \infty$ or $d_{\mathcal{X}}^*(x', x) = \infty$ (here $d_{\mathcal{X}}^*$ is given in Definition A.4.1); and a multiverse is *totally-vacuous* if for all distinct $x \neq x' \in \mathcal{X}$, there exists a universe $\mathcal{D} \in \mathcal{D}$ with $x, x' \in \mathcal{D}$.) This is demonstrated by the following proposition:

Proposition A.4.2. *Given a DP specification $\varepsilon_{\mathcal{D}}\text{-DP}(\mathcal{X}, \mathcal{D}, d_{\mathcal{X}}, D_{\text{Pr}})$,*

$$\mathcal{M}(\mathcal{X}, \mathcal{D}, d_{\mathcal{X}}, D_{\text{Pr}}, \varepsilon_{\mathcal{D}}) = \mathcal{M}(\mathcal{X}, \mathcal{D}_i, d_{\mathcal{X}}, D_{\text{Pr}}, \varepsilon_{\mathcal{D}}),$$

for $i = 1, 2$ where $\mathcal{D}_i = \{\mathcal{D} \cap [x] : \mathcal{D} \in \mathcal{D}, x \in \mathcal{X}\}$ and $\mathcal{D}_2 = \{\{x, x'\} : x, x' \in \mathcal{X} \text{ with } x' \in [x]\}$.

\mathcal{D}' is non-totally-vacuous whenever there are unconnected protection objects. In particular, when $\mathcal{D} = \{\mathcal{X}\}$ then $\mathcal{D}_1 = \{[x] : x \in \mathcal{X}\}$. In bounded DP, $[x]$ is all datasets with the same number of records as x . Hence, bounded DP is equivalent to using the number of records as an invariant.

A.5 THE POST-PROCESSING AND COMPOSITION MECHANISMS

Definition A.5.1. Given a DP flavor $(\mathcal{X}, \mathcal{D}, d_{\mathcal{X}}, D_{\text{Pr}})$, a data-release mechanism $(T : \mathcal{X} \times \mathcal{U} \rightarrow \mathcal{T}, (\mathcal{U}, \mathcal{F}_{\mathcal{U}}, \mathbf{P}(U \in \cdot)), \{\mathcal{F}_{\mathcal{T}}^{[x]}\}_{x \in \mathcal{X}})$ and a function $f : \mathcal{T} \rightarrow \mathcal{T}'$, the *post-processing mechanism* is a data-release mechanism consisting of

1. the function $f \circ T : \mathcal{X} \times \mathcal{U} \rightarrow \mathcal{T}'$;
2. the probability space $(\mathcal{U}, \mathcal{F}_{\mathcal{U}}, \mathbf{P}(U \in \cdot))$ inherited from T ; and
3. for each connected component $[x]$, the σ -algebra

$$\mathcal{F}_{\mathcal{T}'}^{[x]} = \left\{ S \subset \mathcal{T}' : f^{-1}(S) \in \mathcal{F}_T^{[x]} \right\}.$$

Note that the σ -algebra $\mathcal{F}_{\mathcal{T}'}^{[x]}$ of the post-processed mechanism is the largest σ -algebra such that f is measurable with respect to the σ -algebra $\mathcal{F}_T^{[x]}$ of the mechanism T .

For post-processing with a randomised function $f : \mathcal{T} \times \mathcal{U}' \rightarrow \mathcal{T}'$, the post-processing mechanism consists of the function $f(T(x, U), U')$; the probability space $(\mathcal{U} \times \mathcal{U}', \mathcal{F}_{\mathcal{U}} \otimes \mathcal{F}_{\mathcal{U}'}, \mathbf{P})$ where \mathbf{P} is the product of the probability measures of the seeds U and U' ; and the σ -algebras $\mathcal{F}_{\mathcal{T}'}^{[x]} = \left\{ S \subset \mathcal{T}' : f^{-1}(S) \in \mathcal{F}_T^{[x]} \otimes \mathcal{F}_{\mathcal{U}'} \right\}$.

As described in Section 2.5, immunity to randomized post-processing is implied by immunity to non-randomized post-processing, assuming that we can compose the post-processing function's random seed without additional privacy loss. We now formalize this idea.

Definition A.5.2. D_{Pr} is *invariant to extraneous noise* if $D_{\text{Pr}}(\mathbf{P}_1 \times \mathbf{Q}, \mathbf{P}_2 \times \mathbf{Q}) = D_{\text{Pr}}(\mathbf{P}_1, \mathbf{P}_2)$ for all $\mathbf{P}_1, \mathbf{P}_2 \in \mathcal{P}_{(\Omega, \mathcal{F})}$ and all $\mathbf{Q} \in \mathcal{P}_{(\Omega', \mathcal{F}')}$.

Invariance to extraneous noise can be thought of as the composition of two data-release mechanisms in the special case where one mechanism is constant in x and so has zero privacy loss.

Proposition A.5.3. *Suppose that D_{Pr} is invariant to extraneous noise. Then a DP flavor $(\mathcal{X}, \mathcal{D}, d_{\mathcal{X}}, D_{\text{Pr}})$ is immune to randomised post-processing if and only if $(\mathcal{X}, \mathcal{D}, d_{\mathcal{X}}, D_{\text{Pr}})$ is immune to non-randomised post-processing.*

Corollary A.5.4. *Fix a probability premetric D_{Pr} . The DP flavor $(\mathcal{X}, \mathcal{D}, d_{\mathcal{X}}, D_{\text{Pr}})$ is immune to randomised post-processing for all \mathcal{X}, \mathcal{D} and $d_{\mathcal{X}}$ if and only if*

$$D_{\text{Pr}}(\mathbf{P}_1, \mathbf{P}_2) \geq D_{\text{Pr}}(f_*(\mathbf{P}_1 \times \mathbf{Q}), f_*(\mathbf{P}_2 \times \mathbf{Q})), \quad (\text{A.5})$$

for all $\mathbf{P}_1, \mathbf{P}_2 \in \mathcal{P}_{(\Omega, \mathcal{F})}$, all $\mathbf{Q} \in \mathcal{P}_{(\Omega', \mathcal{F}')}$ and all measurable $f: \Omega_1 \times \Omega_2 \rightarrow \mathcal{T}$.

Now we will define the composition of two data-release mechanisms, T_1 and T_2 , which have the seeds U_1 and U_2 respectively. It is possible that U_1 and U_2 are not independent. Such behavior arises for Pufferfish data-release mechanisms, since they share the same data-generating step (Kifer and Machanavajjhala, 2014). More generally, U_1 and U_2 are not independent whenever T_1 and T_2 share some steps of the data life cycle in common – for example, when $T_1(x, U_1)$ and $T_2(x, U_2)$ are outputs of the same sample drawn from the population x (Bailie and Drechsler, 2024).

In these situations, the distribution of the composition $T = (T_1, T_2)$ does not factor as the product of T_1 and T_2 's distributions. To specify the distribution of T , we need to know the joint distribution of U_1 and U_2 . It is not sufficient to know only the data-release mechanisms T_1, T_2 and their (marginal) probability distributions $\mathbf{P}_x(T_1 \in \cdot), \mathbf{P}_x(T_2 \in \cdot)$, we need additional information specifying how they interact. This discussion motivates the following definition:

Definition A.5.5. Given a DP flavor $(\mathcal{X}, \mathcal{D}, d_{\mathcal{X}}, D_{\text{Pr}})$, two data release mechanisms

$$\left(T_1 : \mathcal{X} \times \mathcal{U}_1 \rightarrow \mathcal{T}_1, (\mathcal{U}_1, \mathcal{F}_{\mathcal{U}_1}, \mathbf{P}(U_1 \in \cdot)), \{\mathcal{F}_{\mathcal{T}_1}^{[x]}\}_{x \in \mathcal{X}} \right),$$

$$\left(T_2 : \mathcal{X} \times \mathcal{U}_2 \rightarrow \mathcal{T}_2, (\mathcal{U}_2, \mathcal{F}_{\mathcal{U}_2}, \mathbf{P}(U_2 \in \cdot)), \{\mathcal{F}_{T_2}^{[x]}\}_{x \in \mathcal{X}} \right),$$

and the joint distribution $\mathbf{P}(U_1 \in \cdot, U_2 \in \cdot)$ of the seeds associated with T_1 and T_2 , the *composition mechanism* is the data-release mechanism consisting of

1. the function $T = (T_1, T_2)$ defined as

$$\begin{aligned} \mathcal{X}_1 \times (\mathcal{U}_1 \times \mathcal{U}_2) &\rightarrow \mathcal{T}_1 \times \mathcal{T}_2 \\ (x, (u_1, u_2)) &\mapsto (T_1(x, u_1), T_2(x, u_2)), \end{aligned}$$

2. the probability space

$$(\mathcal{U}_1 \times \mathcal{U}_2, \mathcal{F}_{\mathcal{U}_1} \otimes \mathcal{F}_{\mathcal{U}_2}, \mathbf{P}(U_1 \in \cdot, U_2 \in \cdot)),$$

and

3. for each connected component $[x]$, the product σ -algebra $\mathcal{F}_{T_1}^{[x]} \otimes \mathcal{F}_{T_2}^{[x]}$.

A.6 BLACKWELL'S THEOREM AND POST-PROCESSING

This appendix briefly reviews and builds on some of the results in [Dong et al. \(2022\)](#) and [Su \(2024\)](#).

Given two probabilities $\mathbf{P}, \mathbf{Q} \in \mathcal{P}_{(\Omega, \mathcal{F})}$ and some data $X \in \Omega$, consider testing $H_0 : X \sim \mathbf{P}$ versus $H_1 : X \sim \mathbf{Q}$. A *decision rule* for this hypothesis test is a measurable function $\varphi : \Omega \rightarrow [0, 1]$ which specifies the probability $\varphi(x)$ of rejecting the null upon observing $X = x$. Its size and power are given by $\mathbb{E}_{X \sim \mathbf{P}}[\varphi(X)]$ and $\mathbb{E}_{X \sim \mathbf{Q}}[\varphi(X)]$ respectively.

Definition A.6.1. For $\mathbf{P}, \mathbf{Q} \in \mathcal{P}_{(\Omega, \mathcal{F})}$, the *tradeoff function* $Tr(\mathbf{P}, \mathbf{Q}) : [0, 1] \rightarrow [0, 1]$ is defined as

$$Tr(\mathbf{P}, \mathbf{Q})(\alpha) = \sup_{\varphi} \{ \mathbb{E}_{X \sim \mathbf{Q}}[\varphi(X)] : \mathbb{E}_{X \sim \mathbf{P}}[\varphi(X)] \leq \alpha \},$$

where the supremum is over all level- α decision rules φ for the hypothesis test $H_0 : X \sim P$ versus $H_1 : X \sim Q$.

Let \mathcal{F}_{Tr} be the set of functions $f : [0, 1] \rightarrow [0, 1]$ which are continuous, concave, non-decreasing and which satisfy $f(\alpha) \geq \alpha$ for all $\alpha \in [0, 1]$. Proposition 1 of [Dong et al. \(2022\)](#) proves that \mathcal{F}_{Tr} is the set of all tradeoff functions:

$$\mathcal{F}_{Tr} = \{Tr(P, Q) : P, Q \in \mathcal{P}_{(\Omega, \mathcal{F})} \text{ for any } (\Omega, \mathcal{F})\}.$$

Given $f, g \in \mathcal{F}_{Tr}$, we write $f \preceq g$ if $f(\alpha) \leq g(\alpha)$ for all $\alpha \in [0, 1]$.

Theorem A.6.2 ([Blackwell \(1953\)](#); [Kairouz et al. \(2017\)](#); [Dong et al. \(2022\)](#)). *For $P, Q \in \mathcal{P}_{(\Omega, \mathcal{F})}$ and $P', Q' \in \mathcal{P}_{(\Omega', \mathcal{F}')}$, the following statements are equivalent:*

1. $Tr(P, Q) \succcurlyeq Tr(P', Q')$.
2. *There exists a Markov kernel κ such that $P' = \kappa \circ P$ and $Q' = \kappa \circ Q$.*

In the language of DP, the Markov kernel κ is a (generalisation of a) randomised post-processing function.²

We can also use this idea to create new probability premetrics which satisfy post-processing:

²Formally, given a probability $P \in \mathcal{P}_{(\Omega, \mathcal{F})}$, a measurable space (Ω', \mathcal{F}') and a Markov kernel $\kappa : \mathcal{F}' \times \Omega \rightarrow [0, 1]$, define the probability measure $\kappa \circ P \in \mathcal{P}_{(\Omega', \mathcal{F}')}$ by

$$(\kappa \circ P)(S) = \int_{\Omega} \kappa(S, \omega) dP(\omega),$$

for all $S \in \mathcal{F}'$.

Proposition A.6.3. *Let $\lambda : \mathcal{F}_{Tr} \rightarrow \mathbb{R}$ be a function of tradeoff functions. Define $\lambda_2 : \mathcal{F}_{Tr} \rightarrow \mathbb{R}$ by $\lambda_2(g) = \lambda(g) - \lambda(f)$ where $f(\alpha) = \alpha$. Then $\lambda_2 \circ Tr$ is a probability premetric satisfying invariant to extraneous noise and 5.2 if and only if λ is non-decreasing.*

A.7 PROOFS

Proof of Proposition 2.4.11. T is constant within the universes \mathcal{D} . Therefore $D_{Pr}(P_x, P_{x'}) = 0$ for all $x, x' \in \mathcal{D}$. This proves the first half of the Proposition. To prove the second half, observe that $D_{Pr}(P_{x_1}, P_{x_2}) = \infty$ but $d_{\mathcal{X}}(x_1, x_2) < \infty$. \square

Proof of Proposition 2.4.12. This proposition relies on the metric axiom $D_{Pr}(P, Q) > 0$ if $P \neq Q$. This implies $D_{Pr}(P_{x_1}, P_{x_2}) > 0$. \square

Lemma A.7.1. *Given two DP specifications $\varepsilon_{\mathcal{D}}\text{-DP}(\mathcal{X}, \mathcal{D}, d_{\mathcal{X}}, D_{Pr})$ and $\varepsilon'_{\mathcal{D}'}\text{-DP}(\mathcal{X}, \mathcal{D}', d_{\mathcal{X}}, D_{Pr})$, suppose that, for all $\mathcal{D}' \in \mathcal{D}'$ and all $\delta > 0$, there exists $\mathcal{D} \in \mathcal{D}$ such that $\mathcal{D}' \subset \mathcal{D}$ and $\varepsilon_{\mathcal{D}} \leq \varepsilon'_{\mathcal{D}'} + \delta$. Then*

$$\mathcal{M}(\mathcal{X}, \mathcal{D}, d_{\mathcal{X}}, D_{Pr}, \varepsilon_{\mathcal{D}}) \subset \mathcal{M}(\mathcal{X}, \mathcal{D}', d_{\mathcal{X}}, D_{Pr}, \varepsilon'_{\mathcal{D}'}).$$

Proof. Suppose that $T \in \mathcal{M}(\mathcal{X}, \mathcal{D}, d_{\mathcal{X}}, D_{Pr}, \varepsilon_{\mathcal{D}})$. Let $\mathcal{D}' \in \mathcal{D}'$ and $\delta > 0$. Suppose $x, x' \in \mathcal{D}'$. By assumption, there exists $\mathcal{D} \in \mathcal{D}$ such that $\mathcal{D}' \subset \mathcal{D}$ and

$$D_{Pr}(P_x, P_{x'}) \leq \varepsilon_{\mathcal{D}} d_{\mathcal{X}}(x, x') \leq (\varepsilon'_{\mathcal{D}'} + \delta) d_{\mathcal{X}}(x, x').$$

Since $D_{Pr}(P_x, P_{x'}) \leq (\varepsilon'_{\mathcal{D}'} + \delta) d_{\mathcal{X}}(x, x')$ holds for all $\delta > 0$, it follows that $D_{Pr}(P_x, P_{x'}) \leq \varepsilon'_{\mathcal{D}'} d_{\mathcal{X}}(x, x')$.

This proves $T \in \mathcal{M}(\mathcal{X}, \mathcal{D}', d_{\mathcal{X}}, D_{Pr}, \varepsilon'_{\mathcal{D}'}).$ \square

Proof of Proposition 2.4.13. Because \mathcal{D}' is a refinement of \mathcal{D} , the assumption of Lemma A.7.1 holds for both choices of the budgets $\varepsilon_{\mathcal{D}}$ and $\varepsilon'_{\mathcal{D}'}$. The results then follow by this lemma. \square

Proof of Proposition 2.4.15. Suppose $T \in \mathcal{M}(\mathcal{X}, \mathcal{D}, d_{\mathcal{X}}, D_{\text{Pr}}, \varepsilon_{\mathcal{D}})$. Let $x, x' \in \mathcal{D} \in \mathcal{D}$. Then

$$D_{\text{Pr}}(\mathbf{P}_x, \mathbf{P}_{x'}) \leq \varepsilon_{\mathcal{D}} d_{\mathcal{X}}(x, x') \leq \varepsilon_{\mathcal{D}} d'_{\mathcal{X}}(x, x')/l.$$

The second half of the proposition follows analogously. \square

Proof of Proposition 2.5.3. “ \Leftarrow ”: Suppose T_1 and T_2 satisfy $(\mathcal{X}, \mathcal{D}, d_{\mathcal{X}}, D_{\text{Pr}})$ with budgets $\varepsilon_{\mathcal{D}}^{(1)}$ and $\varepsilon_{\mathcal{D}}^{(2)}$ respectively. Let U_1 and U_2 be the seeds of T_1 and T_2 respectively. We may assume that U_1 and U_2 are independent. Then, for any $x, x' \in \mathcal{D}$ and any $\mathcal{D} \in \mathcal{D}$,

$$\begin{aligned} D_{\text{Pr}}[\mathbf{P}_x((T_1, T_2) \in \cdot), \mathbf{P}_{x'}((T_1, T_2) \in \cdot)] &= D_{\text{Pr}}[\mathbf{P}_x(T_1 \in \cdot) \times \mathbf{P}_x(T_2 \in \cdot), \mathbf{P}_{x'}(T_1 \in \cdot) \times \mathbf{P}_{x'}(T_2 \in \cdot)] \\ &\leq D_{\text{Pr}}[\mathbf{P}_x(T_1 \in \cdot), \mathbf{P}_{x'}(T_1 \in \cdot)] + D_{\text{Pr}}[\mathbf{P}_x(T_2 \in \cdot), \mathbf{P}_{x'}(T_2 \in \cdot)] \\ &\leq (\varepsilon_{\mathcal{D}}^{(1)} + \varepsilon_{\mathcal{D}}^{(2)}) d_{\mathcal{X}}(x, x'), \end{aligned}$$

where the first line follows since U_1 and U_2 are independent; the second by (2.10); and the third because T_1 and T_2 are DP.

“ \Rightarrow ”: Set $\mathcal{X} = \{x_1, x_2\}$ and $\mathcal{D} = \{\mathcal{X}\}$. Define the premetric $d_{\mathcal{X}}$ by $d_{\mathcal{X}}(x_1, x_2) = 1$ and $d_{\mathcal{X}}(x_2, x_1) = \infty$. Define the mechanism $T_1 : \mathcal{X} \times \mathcal{U} \rightarrow \Omega$ by $\mathbf{P}_{x_1}(T_1 \in \cdot) = \mathbf{P}$ and $\mathbf{P}_{x_2}(T_1 \in \cdot) = \mathbf{Q}$. Similarly define $T_2 : \mathcal{X} \times \mathcal{U} \rightarrow \Omega'$ by $\mathbf{P}_{x_1}(T_2 \in \cdot) = \mathbf{P}'$ and $\mathbf{P}_{x_2}(T_2 \in \cdot) = \mathbf{Q}'$, with the seed of T_2 independent of the seed of T_1 .

T_1 and T_2 satisfy $(\mathcal{X}, \mathcal{D}, d_{\mathcal{X}}, D_{\text{Pr}})$ with $\varepsilon_{\mathcal{X}}^{(1)} = D_{\text{Pr}}(\mathbf{P}, \mathbf{Q})$ and $\varepsilon_{\mathcal{X}}^{(2)} = D_{\text{Pr}}(\mathbf{P}', \mathbf{Q}')$ respectively. By

assumption (T_1, T_2) also satisfies $(\mathcal{X}, \mathcal{D}, d_{\mathcal{X}}, D_{\text{Pr}})$ with $\varepsilon_{\mathcal{X}}^{(1)} + \varepsilon_{\mathcal{X}}^{(2)}$. Then

$$\begin{aligned} D_{\text{Pr}}(\mathbf{P} \times \mathbf{P}', \mathbf{Q} \times \mathbf{Q}') &= D_{\text{Pr}}[\mathbf{P}_x((T_1, T_2) \in \cdot), \mathbf{P}_{x'}((T_1, T_2) \in \cdot)] \\ &\leq \varepsilon_{\mathcal{X}}^{(1)} + \varepsilon_{\mathcal{X}}^{(2)} \\ &= D_{\text{Pr}}(\mathbf{P}, \mathbf{Q}) + D_{\text{Pr}}(\mathbf{P}', \mathbf{Q}'). \end{aligned} \quad \square$$

Proof of Proposition 2.5.4. “ \Leftarrow ”: Suppose that (2.11) holds for all $\mathbf{P}, \mathbf{Q} \in \mathcal{P}_{(\Omega, \mathcal{F})}$ and all measurable $f: (\Omega, \mathcal{F}) \rightarrow (\Omega', \mathcal{F}')$. Let $T: \mathcal{X} \times \mathcal{U} \rightarrow \Omega$ satisfy $\varepsilon_{\mathcal{D}}\text{-DP}(\mathcal{X}, \mathcal{D}, d_{\mathcal{X}}, D_{\text{Pr}})$. Since $\mathbf{P}_x(f \circ T \in S) = f_{\star}[\mathbf{P}_x(T \in \cdot)](S)$,

$$D_{\text{Pr}}[\mathbf{P}_x(f \circ T \in \cdot), \mathbf{P}_{x'}(f \circ T \in \cdot)] \leq D_{\text{Pr}}[\mathbf{P}_x(T \in \cdot), \mathbf{P}_{x'}(T \in \cdot)] \leq \varepsilon_{\mathcal{D}} d_{\mathcal{X}}(x, x'),$$

for all $x, x' \in \mathcal{D}$ and all $\mathcal{D} \in \mathcal{D}$. Hence $f \circ T$ satisfies $\varepsilon_{\mathcal{D}}\text{-DP}(\mathcal{X}, \mathcal{D}, d_{\mathcal{X}}, D_{\text{Pr}})$.

“ \Rightarrow ”: Set $\mathcal{X} = \{x_1, x_2\}$ and $\mathcal{D} = \{\mathcal{X}\}$. Define the premetric $d_{\mathcal{X}}$ by $d_{\mathcal{X}}(x_1, x_2) = 1$ and $d_{\mathcal{X}}(x_2, x_1) = \infty$. Define the mechanism $T: \mathcal{X} \times \mathcal{U} \rightarrow \Omega$ by $\mathbf{P}_{x_1}(T \in \cdot) = \mathbf{P}$ and $\mathbf{P}_{x_2}(T \in \cdot) = \mathbf{Q}$. Then T satisfies $\varepsilon_{\mathcal{D}}\text{-DP}(\mathcal{X}, \mathcal{D}, d_{\mathcal{X}}, D_{\text{Pr}})$ with $\varepsilon_{\mathcal{X}} = D_{\text{Pr}}(\mathbf{P}, \mathbf{Q})$ and hence

$$D_{\text{Pr}}[f_{\star}(\mathbf{P}), f_{\star}(\mathbf{Q})] = D_{\text{Pr}}[\mathbf{P}_{x_1}(f \circ T), \mathbf{P}_{x_2}(f \circ T)] \leq \varepsilon_{\mathcal{X}} d_{\mathcal{X}}(x_1, x_2) = D_{\text{Pr}}(\mathbf{P}, \mathbf{Q}). \quad \square$$

Proof of Proposition A.2.1. Suppose that $f(x)$ is a function of $[x]$. Let $\mathcal{D} \in \mathcal{D}$ and $x, x' \in \mathcal{D}$. If $d_{\mathcal{X}}(x, x') = \infty$ then the Lipschitz condition (2.3) is satisfied with $\varepsilon_{\mathcal{D}} = 0$. Otherwise, $x' \in [x]$ so that $f(x') = f(x)$. This implies $D_{\text{Pr}}(\mathbf{P}_x, \mathbf{P}_{x'}) = 0$ so that (2.3) is again satisfied with $\varepsilon_{\mathcal{D}} = 0$. This proves f satisfies $\varepsilon_{\mathcal{D}}\text{-DP}(\mathcal{X}, \mathcal{D}, d_{\mathcal{X}}, D_{\text{Pr}})$ with zero privacy loss.

In the other direction, suppose that f satisfies $\varepsilon_{\mathcal{D}}\text{-DP}(\mathcal{X}, \mathcal{D}, d_{\mathcal{X}}, D_{\text{Pr}})$ with $\varepsilon_{\mathcal{D}} < \infty$. Let $x, x' \in \mathcal{X}$ with $x \in [x']$. We will show that $f(x) = f(x')$ by induction. Let G be the graph with nodes \mathcal{X} and edges

(x_1, x_2) if x_1 and x_2 are explicitly comparable (Definition A.1.1). Since $x \in [x']$, there exists a finite path $P = (x = x_1, x_2, x_3, \dots, x_{n-1}, x_n = x')$ of distinct nodes in G from x to x' .

Now we proceed with the proof by induction on n . The base case is trivial: $f(x) = f(x)$. For the step case, assume $f(x_{n-1}) = f(x)$ so that, in order to prove $f(x) = f(x')$ it suffices to show $f(x') = f(x_{n-1})$. Because x_{n-1} and $x' = x_n$ are explicitly comparable, there must exist some $\mathcal{D} \in \mathcal{D}$ such that $x_{n-1}, x' \in \mathcal{D}$. Moreover, $d_{\mathcal{X}}(x_{n-1}, x') < \infty$ or $d_{\mathcal{X}}(x', x_{n-1}) < \infty$. By the Lipschitz condition 2.3, this implies $D_{\text{Pr}}(\mathbf{P}_{x_{n-1}}, \mathbf{P}_{x'})$ or $D_{\text{Pr}}(\mathbf{P}_{x'}, \mathbf{P}_{x_{n-1}})$ is also finite. Hence $d_{\text{TV}}(\mathbf{P}_{x_{n-1}}, \mathbf{P}_{x'}) < 1$. But $\mathbf{P}_{x_{n-1}}$ is a point mass at $f(x_{n-1})$. Thus $f(x_{n-1}) = f(x')$. \square

Proof of Corollary A.2.2. Clearly $f(x)$ is a function of $[x]$. Using the reasoning in the proof of Proposition A.2.1, this implies $f(x)$ satisfies $\varepsilon'_{\mathcal{D}}\text{-DP}(\mathcal{X}, \mathcal{D}, d_{\mathcal{X}}, D_{\text{Pr}})$ with $\varepsilon'_{\mathcal{D}} = 0$. \square

Proof of Proposition A.3.1. Suppose that $T \in \mathcal{M}(\mathcal{X}, \mathcal{D}, d_{\mathcal{X}}, D_{\text{Pr}}, \varepsilon_{\mathcal{D}})$, where $\varepsilon_{\mathcal{D}} = \varepsilon$ is a constant privacy loss budget, and let $x, x' \in \mathcal{X}$. There are three cases:

1. Suppose $x = x'$. Then $D_{\text{Pr}}(\mathbf{P}_x, \mathbf{P}_{x'}) = 0 = \varepsilon d'_{\mathcal{X}}(x, x')$.
2. Suppose $x \neq x'$ but there exists some data universe $\mathcal{D} \in \mathcal{D}$ such that $x, x' \in \mathcal{D}$. Then $\varepsilon d'_{\mathcal{X}}(x, x') = \varepsilon d_{\mathcal{X}}(x, x') \geq D_{\text{Pr}}(\mathbf{P}_x, \mathbf{P}_{x'})$.
3. Suppose $x \neq x'$ and there does not exist a universe $\mathcal{D} \in \mathcal{D}$ such that $x, x' \in \mathcal{D}$. Then $\varepsilon d'_{\mathcal{X}}(x, x') = \infty \geq D_{\text{Pr}}(\mathbf{P}_x, \mathbf{P}_{x'})$.

This proves $T \in \mathcal{M}(\mathcal{X}, \{\mathcal{X}\}, d'_{\mathcal{X}}, D_{\text{Pr}}, \varepsilon_{\mathcal{D}})$.

Now suppose that $T \in \mathcal{M}(\mathcal{X}, \{\mathcal{X}\}, d'_{\mathcal{X}}, D_{\text{Pr}}, \varepsilon_{\mathcal{D}})$, again with a constant privacy loss budget $\varepsilon_{\mathcal{D}} = \varepsilon$.

Let $\mathcal{D} \in \mathcal{D}$ and $x, x' \in \mathcal{D}$. Then

$$\varepsilon d_{\mathcal{X}}(x, x') = \varepsilon d'_{\mathcal{X}}(x, x') \geq D_{\text{Pr}}(\mathbf{P}_x, \mathbf{P}_{x'}).$$

This proves $T \in \mathcal{M}(\mathcal{X}, \mathcal{D}, d_{\mathcal{X}}, D_{\text{Pr}}, \varepsilon_{\mathcal{D}})$. \square

Proof of Proposition A.3.2. Since \mathcal{D}_1 is a refinement of \mathcal{D}_0 , one direction

$$\mathcal{M}(\mathcal{X}, \mathcal{D}_0, d_{\mathcal{X}}, D_{\text{Pr}}, \varepsilon_{\mathcal{D}}) \subset \mathcal{M}(\mathcal{X}, \mathcal{D}_1, d_{\mathcal{X}}, D_{\text{Pr}}, \varepsilon'_{\mathcal{D}'})$$

is immediate by Proposition 2.4.13. In the other direction, suppose $T \in \mathcal{M}(\mathcal{X}, \mathcal{D}_1, d_{\mathcal{X}}, D_{\text{Pr}}, \varepsilon'_{\mathcal{D}'})$ and let $x, x' \in \mathcal{D}_0$ for some $\mathcal{D}_0 \in \mathcal{D}_0$. If $d_{\mathcal{X}}(x, x') = \infty$, then the Lipschitz condition (2.3) holds trivially.

Otherwise, $\{x, x'\} \in \mathcal{D}$ and hence $\{x, x'\} \cap \mathcal{D}_0 \in \mathcal{D}_1$. Thus,

$$D_{\text{Pr}}(P_x, P_{x'}) \leq \varepsilon'_{\{x, x'\} \cap \mathcal{D}_0} d_{\mathcal{X}}(x, x') \leq \varepsilon_{\mathcal{D}_0} d_{\mathcal{X}}(x, x').$$

This proves $T \in \mathcal{M}(\mathcal{X}, \mathcal{D}_0, d_{\mathcal{X}}, D_{\text{Pr}}, \varepsilon_{\mathcal{D}})$. □

Proof of Proposition A.3.3. As in the proof of Proposition A.3.2, one direction follows by Proposition 2.4.13.

In the other direction, suppose that $T \in \mathcal{M}(\mathcal{X}, \mathcal{D}_1, d_{\mathcal{X}}, D_{\text{Pr}}, \varepsilon'_{\mathcal{D}'})$ and let $x, x' \in \mathcal{D}_0$ for some $\mathcal{D}_0 \in \mathcal{D}_0$.

If $d_{\mathcal{X}}(x, x') = \infty$, then the Lipschitz condition (2.3) holds trivially. Otherwise, $[x] = [x']$ so that $\mathcal{D}_c(x) = \mathcal{D}_c(x') \in \mathcal{D}_c$. Then

$$D_{\text{Pr}}(P_x, P_{x'}) \leq \varepsilon'_{\mathcal{D}_0 \cap \mathcal{D}_c(x)} d_{\mathcal{X}}(x, x') \leq \varepsilon_{\mathcal{D}_0} d_{\mathcal{X}}(x, x'),$$

where the first inequality follows by noting that $x, x' \in \mathcal{D}_0 \cap \mathcal{D}_c(x) \in \mathcal{D}_1$ and $T \in \mathcal{M}(\mathcal{X}, \mathcal{D}_1, d_{\mathcal{X}}, D_{\text{Pr}}, \varepsilon_{\mathcal{D}})$ and the second by the definition of $\varepsilon'_{\mathcal{D}'}$.

To prove the last part of the proposition, observe that the connected components of the graph (\mathcal{D}_0, \sim_d) are exactly the equivalence classes of \sim_d , when restricting \sim_d to \mathcal{D}_0 – that is, the equivalence classes of $\sim_d \cap (\mathcal{D}_0 \times \mathcal{D}_0)$. But by definition, the set of these equivalence classes is

$$\{\mathcal{D}_0 \cap \mathcal{D} : \mathcal{D} \in \mathcal{D}_c \text{ such that } \mathcal{D}_0 \cap \mathcal{D} \neq \emptyset\}.$$

Hence, this set is not in \mathcal{D}_0 iff the connected components of (\mathcal{D}_0, \sim_d) are not in \mathcal{D}_0 (assuming that $\mathcal{D}_0 \neq \emptyset$). Therefore, $\mathcal{D}_1 \neq \mathcal{D}_0$ iff there exists $\mathcal{D}_0 \in \mathcal{D}_0$ such that one of its connected components is not in \mathcal{D}_0 (assuming that $\emptyset \notin \mathcal{D}_0$). \square

Proof of Proposition A.5.3. One direction (from randomised post-processing to non-randomised post-processing) is trivial. In the other direction, let $T : \mathcal{X} \times \mathcal{U} \rightarrow \mathcal{T}$ be a data-release mechanism which satisfies $\varepsilon_{\mathcal{D}}\text{-DP}(\mathcal{X}, \mathcal{D}, d_{\mathcal{X}}, D_{\text{Pr}})$ for some $\varepsilon_{\mathcal{D}}$ and let $f : \mathcal{T} \times \mathcal{U}' \rightarrow \mathcal{T}'$ be a randomised function. Denote the random seeds of T and f by U and U' respectively. The law of $f(T(x, U), U')$ is equal to $f_{\star}(\mathbb{P}_x(T \in \cdot) \times \mathbb{P}(U' \in \cdot))$. Therefore,

$$\begin{aligned} D_{\text{Pr}} \left[\mathbb{P}_x(f(T(x, U), U') \in \cdot), \mathbb{P}_{x'}(f(T(x', U), U') \in \cdot) \right] \\ = D_{\text{Pr}} \left[f_{\star}(\mathbb{P}_x(T \in \cdot) \times \mathbb{P}(U' \in \cdot)), f_{\star}(\mathbb{P}_{x'}(T \in \cdot) \times \mathbb{P}(U' \in \cdot)) \right] \\ \leq D_{\text{Pr}} \left[\mathbb{P}_x(T \in \cdot) \times \mathbb{P}(U' \in \cdot), \mathbb{P}_{x'}(T \in \cdot) \times \mathbb{P}(U' \in \cdot) \right] \\ = D_{\text{Pr}}[\mathbb{P}_x(T \in \cdot), \mathbb{P}_{x'}(T \in \cdot)], \end{aligned}$$

where the first inequality follows by non-randomised post-processing and the second because D_{Pr} is invariant to extraneous noise. \square

Lemma A.7.2. *Fix a probability premetric D_{Pr} . Suppose that, for all \mathcal{X}, \mathcal{D} and $d_{\mathcal{X}}$, the DP flavor $(\mathcal{X}, \mathcal{D}, d_{\mathcal{X}}, D_{\text{Pr}})$ is immune to randomised post-processing. Then D_{Pr} is invariant to extraneous noise.*

Proof. Let $\mathbb{P}_1, \mathbb{P}_2 \in \mathcal{P}_{(\Omega, \mathcal{F})}$ and $\mathbb{Q} \in \mathcal{P}_{(\Omega', \mathcal{F}')}$. Set $\mathcal{X} = \{x_1, x_2\}$ and $\mathcal{D} = \{\mathcal{X}\}$. Define the premetric $d_{\mathcal{X}}$ by $d_{\mathcal{X}}(x_1, x_2) = 1$ and $d_{\mathcal{X}}(x_2, x_1) = \infty$. Define the mechanism $T : \mathcal{X} \times \mathcal{U} \rightarrow \Omega \times \Omega'$ by $\mathbb{P}_{x_1}(T \in \cdot) = \mathbb{P}_1 \times \mathbb{Q}$ and $\mathbb{P}_{x_2}(T \in \cdot) = \mathbb{P}_2 \times \mathbb{Q}$. Then T satisfies $\varepsilon_{\mathcal{D}}\text{-DP}(\mathcal{X}, \mathcal{D}, d_{\mathcal{X}}, D_{\text{Pr}})$ with $\varepsilon_{\mathcal{X}} = D_{\text{Pr}}(\mathbb{P}_1 \times \mathbb{Q}, \mathbb{P}_2 \times \mathbb{Q})$. Define the projection $f(\omega, \omega') = \omega$. Since f is (non-random) post-processing,

$f \circ T$ must also satisfy $\varepsilon_{\mathcal{D}}\text{-DP}(\mathcal{X}, \mathcal{D}, d_{\mathcal{X}}, D_{\text{Pr}})$ with the same privacy loss budget. Hence

$$D_{\text{Pr}}(\mathbf{P}_1, \mathbf{P}_2) = D_{\text{Pr}}(\mathbf{P}_{x_1}(f \circ T \in \cdot), \mathbf{P}_{x_2}(f \circ T \in \cdot)) \leq \varepsilon_{\mathcal{X}}.$$

This proves $D_{\text{Pr}}(\mathbf{P}_1, \mathbf{P}_2) \leq D_{\text{Pr}}(\mathbf{P}_1 \times \mathbf{Q}, \mathbf{P}_2 \times \mathbf{Q})$. For the opposite inequality, define the mechanism $T' : \mathcal{X} \times \mathcal{U} \rightarrow \Omega$ by $\mathbf{P}_{x_1}(T' \in \cdot) = \mathbf{P}_1$ and $\mathbf{P}_{x_2}(T' \in \cdot) = \mathbf{P}_2$. Then T' satisfies $\varepsilon_{\mathcal{D}}\text{-DP}(\mathcal{X}, \mathcal{D}, d_{\mathcal{X}}, D_{\text{Pr}})$ with $\varepsilon_{\mathcal{X}} = D_{\text{Pr}}(\mathbf{P}_1, \mathbf{P}_2)$. Let $f(t, U') = (t, U')$ be the identity function on $\mathcal{T} \times \Omega'$ where $U' \sim \mathbf{Q}$. Now f is random post-processing, so $f[T(x, U), U']$ must again satisfy $\varepsilon_{\mathcal{D}}\text{-DP}(\mathcal{X}, \mathcal{D}, d_{\mathcal{X}}, D_{\text{Pr}})$ where $\varepsilon_{\mathcal{X}} = D_{\text{Pr}}(\mathbf{P}_1, \mathbf{P}_2)$. Hence

$$D_{\text{Pr}}(\mathbf{P}_1 \times \mathbf{Q}, \mathbf{P}_2 \times \mathbf{Q}) = D_{\text{Pr}}\left[\mathbf{P}_{x_1}(f[T(x_1, U), U'] \in \cdot), \mathbf{P}_{x_2}(f[T(x_2, U), U'] \in \cdot)\right] \leq \varepsilon_{\mathcal{X}}.$$

This proves $D_{\text{Pr}}(\mathbf{P}_1 \times \mathbf{Q}, \mathbf{P}_2 \times \mathbf{Q}) \leq D_{\text{Pr}}(\mathbf{P}_1, \mathbf{P}_2)$. □

Proof of Corollary A.5.4. By Lemma A.7.2 and Propositions A.5.3 and 2.5.4, the DP flavor $(\mathcal{X}, \mathcal{D}, d_{\mathcal{X}}, D_{\text{Pr}})$ is immune to randomised post-processing for all \mathcal{X}, \mathcal{D} and $d_{\mathcal{X}}$ if and only if the following two conditions hold:

1. D_{Pr} is invariant to extraneous noise; and
2. D_{Pr} satisfies (2.11) (for all \mathbf{P}, \mathbf{Q} and f).

Thus, all we must prove is that D_{Pr} satisfies (A.5) (for all $\mathbf{P}_1, \mathbf{P}_2$, all \mathbf{Q} and all f) if and only if the above two conditions hold. In one direction, clearly (A.5) implies (2.11). Further, by setting f in (A.5) to be the identity map, we get

$$D_{\text{Pr}}(\mathbf{P}_1, \mathbf{P}_2) \geq D_{\text{Pr}}(\mathbf{P}_1 \times \mathbf{Q}, \mathbf{P}_2 \times \mathbf{Q}),$$

and by setting f to be the projection map onto the first coordinate

$$D_{\text{Pr}}(\mathbf{P}_1 \times \mathbf{Q}, \mathbf{P}_2 \times \mathbf{Q}) \geq D_{\text{Pr}}(f_{\star}(\mathbf{P}_1 \times \mathbf{Q} \times \mathbf{Q}'), f_{\star}(\mathbf{P}_2 \times \mathbf{Q} \times \mathbf{Q}')) = D_{\text{Pr}}(\mathbf{P}_1, \mathbf{P}_2).$$

This proves that (A.5) implies D_{Pr} is invariant to extraneous noise.

In the other direction, suppose that D_{Pr} is invariant to extraneous noise and satisfies (2.11). Then

$$D_{\text{Pr}}(P_1, P_2) = D_{\text{Pr}}(P_1 \times Q, P_2 \times Q) \geq D_{\text{Pr}}[f_{\star}(P_1 \times Q), f_{\star}(P_2 \times Q)].$$

This proves that D_{Pr} satisfies A.5. □

B

Appendices to Chapter 3

B.1 BACKGROUND ON DATA SWAPPING

Invented by Dalenius and Reiss [1978; 1982] and further expanded upon by Fienberg and McIntyre [2004], data swapping (also called record swapping, particularly in Europe) refers to a family of SDC methods which select some subset of records and permute the values these records take for a subset of variables. These methods differ on which variables are swapped, how records are selected to be swapped, and how the interchanging of the values of the swapping variables between the selected records is conducted. (See (DePersio et al., 2012; Kim, 2015; Shlomo et al., 2010; Fienberg and McIntyre, 2004) for examples of different data swapping methods.) Traditionally, claims of SDC protection provided by swapping methods have been based on the intuition that a successful disclosure requires linking inferred information about a *sensitive variable* to an individual entity using some *quasi-identifying variables*. By sensitive variable, we mean a variable that is plausibly of interest to an attacker – for example, a person’s race or a household’s income.

Learning the value of a sensitive variable for an individual record may not be problematic on its own since the attacker does not know to whom the record belongs. Thus, an attacker has two goals: 1) to infer the value of a sensitive variable for an individual record and 2) to determine, using quasi-identifying variables, the individual entity associated with that record. Since the sensitive variable and the quasi-identifiers must belong to the same record, the attacker needs to infer them jointly. The idea behind data swapping is to hinder such joint inference by randomly permuting the records' quasi-identifiers while keeping the sensitive variables fixed (or visa versa). In this way, there are multiple plausible values for the original dataset which are compatible with the swapped dataset – thereby adding uncertainty to the relationship between any record's sensitive variables and its quasi-identifiers.

It is important to emphasize that the above discussion is only an intuitive justification for data swapping. A major motivation for this paper is to supplement such intuitive arguments with mathematical SDC guarantees. Some such guarantees are provided by the PSA's DP specification. In fact, Theorem 3.2.4 can be interpreted as a formalization of the above intuitive argument because it provides a bound on how plausible the true confidential dataset is compared to other compatible datasets. This bound ensures a degree of uncertainty in the relationship between V_{Swap} and $V_{\text{Hold}} \setminus V_{\text{Swap}}$. Taking V_{Swap} to be the quasi-identifiers and $V_{\text{Hold}} \setminus V_{\text{Swap}}$ the sensitive variables (or visa versa), this recovers the above argument. However, theoretically any set of variables can function as quasi-identifiers, depending on the attacker's auxiliary knowledge and the context of the data collection (see e.g. [Sweeney \(2000, 2002\)](#); [Machanavajjhala et al. \(2007\)](#); [Cohen \(2022\)](#)). As such, arguments that rely on knowing what variables are quasi-identifiers may have limited utility outside the scope of context-specific SDC analyses.

Data swapping is widely utilized – typically in combination with other SDC methods – by statistical

offices across the globe. As we remark in the main body of this paper, it has been used and studied extensively by the USCB (McKenna and Haubach, 2019; Steel and Zayatz, 2003; Zayatz et al., 2010; Zayatz, 2007; Lauger et al., 2014; Lemons et al., 2015). The Office for National Statistics (ONS) of the United Kingdom (UK) has employed it for their 2001, 2011 and 2021 Censuses (Spicer, 2020; Office for National Statistics, 2023; Shlomo et al., 2010). It is one of the two protection methods recommended by Eurostat’s *Centre of Excellence on Statistical Disclosure Control* (Glessing and Schulte Nordholt, 2017) and was used (or is intended to be used) for protecting census data by 15 of 30 European Union states surveyed by de Vries et al. (2023). The Australian Bureau of Statistics uses it as one of their primary SDC methods for releasing microdata (Australian Bureau of Statistics, 2021b). And it has been explored as a method for protecting the Japanese Population Census (Ito and Hoshino, 2014).

While we largely focus on the swapping procedure used in the 2010 US Census, much of this paper also applies to other statistical agencies, especially when their swapping mechanisms are similar to the US 2010 Census DAS. In particular, the ONS’s Targeted Record Swapping (UK Statistics Authority, 2021) closely aligns with the procedure used in the 2010 US Census and hence this work is also relevant for the 2021 UK Census.

B.2 OTHER RELATED WORK

In this appendix, we briefly review some related work which was not covered in the main body of this paper. Firstly, there is existing literature examining DP under invariants. One branch of this literature develops DP mechanisms which report invariants without noise. In addition to the USCB’s work on the TDA (Abowd et al., 2022a), other papers in this branch include Gong and Meng (2020); Gao et al. (2022) and Dharangutte et al. (2023). As invariants can be viewed from an attacker’s perspective as background knowl-

edge, work addressing how to incorporating this knowledge into DP (Kifer and Machanavajjhala, 2014; He et al., 2014; Kifer and Machanavajjhala, 2011; Desfontaines et al., 2020) is also relevant. In particular, Seeman et al. (2022) applies the Pufferfish privacy framework to construct a DP formulation which can handle invariants, although with the additional complication that the data must be modeled. Although not specifically addressing invariant-respecting DP, Protivash et al. (2022) demonstrates that related DP formulations – which, like invariants, also restrict the data universes – may not provide sufficient SDC. More recent work on invariants includes Cho and Awan (2024).

There is also related work studying SDC for the US Decennial Census. Ashmead et al. (2019) and Kifer et al. (2022) describe DP semantics for the 2020 Census, with the former focusing on the impact of invariants. Abowd et al. (2023) examines the 2010 DAS, using a reconstruction attack to demonstrate that aggregation did not provide SDC, as has traditionally been assumed. Christ et al. (2022) compares data swapping with standard ε -DP mechanisms. And the paper which first proposed data swapping (Dalenius and Reiss, 1982) includes theoretical justification for the SDC provided by data swapping, which was reviewed by Fienberg and McIntyre (2004).

For a review of literature related to the system of DP specifications introduced in Part I, see Subsection 2.3.2.

B.3 PROOF OF THEOREM 3.2.4

In this appendix, we prove that Algorithm 3.2.1 (the PSA) satisfies $\varepsilon_{\mathcal{D}}\text{-DP}(\mathcal{X}, \mathcal{D}_{\text{cSwap}}, d_{\text{HamS}}^r, D_{\text{MULT}})$ for the value of $\varepsilon_{\mathcal{D}}$ given in Theorem 3.2.4. Assume throughout this appendix the conditions of Theorem 3.2.4: that all $x \in \mathcal{X}$ share a common set of variables, partitioned into subsets V_{Swap} and V_{Hold} ; and that the PSA swaps records at the same resolution as d_{HamS}^r .

By Proposition 1, we may also assume that there is exactly one matching variable, one non-matching holding variable and one swapping variable. For ease of exposition, we assume that each of these variables can take on a finite number of values, which we denote by $m = 1, \dots, \mathcal{M}$ and $b = 1, \dots, \mathcal{H}$ and $s = 1, \dots, \mathcal{S}$ respectively, although the proof immediately generalizes beyond this assumption. Recall that n_{mbs}^x is the count of records in x which take the value (m, b, s) . Replacing a category m, b, s with \cdot denotes a marginal count – for example, $n_{m\cdot s}^x = \sum_{b=1}^{\mathcal{H}} n_{mbs}^x$. We will drop x in the superscript and \cdot in the subscript when this does not cause ambiguity.

Write (M_i^x, H_i^x, S_i^x) for the i -th record in x , so that we can write x as the vector $[(M_i, H_i, S_i)]_{i=1}^n$, where $n = n_{\cdot\cdot\cdot}^x = |x|$ is the number of records in x . With this notation, $n_{mbs}^x = \sum_{i=1}^n 1_{M_i^x=m} 1_{H_i^x=b} 1_{S_i^x=s}$.

Let $\ell_1^r(x, x')$ be the ℓ_1 -distance on the interior cells of the fully-saturated contingency table

$$\ell_1^r(x, x') := \sum_{m,b,s} |n_{mbs}^x - n_{mbs}^{x'}|. \quad (\text{B.1})$$

Lemma B.3.1. $\ell_1^r(x, x') = 2d_{\text{HamS}}^r(x, x')$ if $|x| = |x'|$.

Lemma B.3.2. D_{MULT} is a metric on the space of a.e. equal random variables (over the same probability space \mathcal{T}).

Proof. It is easy to see that D_{MULT} is symmetric and $D_{\text{MULT}}(X, Y) = 0$ if and only if $X = Y$ a.e. All that remains is to verify the triangle inequality. Let $\{E_n\} \subset \mathcal{F}$ such that

$$\left| \ln \frac{\mathbb{P}(X \in E_n)}{\mathbb{P}(Z \in E_n)} \right| \rightarrow D_{\text{MULT}}(X, Z),$$

as $n \rightarrow \infty$. Then

$$\left| \ln \frac{\mathbb{P}(X \in E_n)}{\mathbb{P}(Z \in E_n)} \right| \leq |\ln[\mathbb{P}(X \in E_n)] - \ln[\mathbb{P}(Y \in E_n)]| + |\ln[\mathbb{P}(Y \in E_n)] - \ln[\mathbb{P}(Z \in E_n)]|$$

$$\leq D_{\text{MULT}}(X, Y) + D_{\text{MULT}}(Y, Z).$$

□

Recall that σ_m is the random perturbation sampled by the PSA which deranges the selected records in matching stratum m . Let σ be the permutation defined by $\sigma(i) = \sigma_{M_i}(i)$. Since σ_m fixes i whenever $M_i \neq m$, it is the case that $\sigma = \sigma_{M_1} \circ \dots \circ \sigma_1$. (Note that σ is a random function of the input dataset x , although we leave this dependence implicit.) For a permutation g , write $g(x)$ as shorthand for the dataset in which the values of the swapping variables have been permuted according to g . That is, if $x = [(M_i, H_i, S_i)]_{i=1}^n$ then $g(x) = [(M_i, H_i, S_{g(i)})]_{i=1}^n$. Given an input dataset x , the swapped dataset $\sigma(x)$ generated by the PSA is denoted by Z .

Let P_x denote the probability induced by the randomness in the PSA (i.e. the randomness in selecting records and in sampling the permutation σ), taking the input dataset x as fixed. Recall that the output of the PSA is the fully saturated contingency table $C(Z) = [n_{jkl}^Z]$.

Lemma B.3.3. *If x and x' differ only by reordering of rows (i.e. $d_{\text{HamS}}^r(x, x') = 0$), then*

$$D_{\text{MULT}}[P_x(C(Z)), P_{x'}(C(Z))] = 0.$$

Proof. The contingency table $[n_{mhs}^Z]$ is invariant to reordering of rows of Z . Thus $P_x(C(Z)) = P_{x'}(C(Z))$.

□

Lemma B.3.4. *Fix some data universe $\mathcal{D} \in \mathcal{D}_{\text{cswap}}$ and some $x, x' \in \mathcal{D}$ with $d_{\text{HamS}}^r(x, x') = \Delta$. Then there exists a permutation ρ which fixes exactly $n - \Delta$ records such that $C(\rho(x)) = C(x')$.*

Proof. We have that $\Delta < \infty$ since the invariants c_{swap} imply that all datasets in \mathcal{D} have the same number of records. Hence the symmetric difference $x \ominus x'$ contain 2Δ records, with Δ records from x and Δ records

from x' . Denote the records in $x \ominus x'$ which come from x by x_0 and the records from x' by x'_0 , so that $x \ominus x'$ is the disjoint union of x_0 and x'_0 .

Without loss of generality, we may assume that there is a single matching category ($\mathcal{M} = 1$). (If there is more than one matching category, apply the following argument to each category separately.) Then the dataset x (disregarding the order of the records) can be represented as the matrix $C(x) = [n_{bs}^x]$.

We will need the following result (*) whose proof is straightforward: For any $x'', x''' \in \mathcal{X}$, the matrix $C(x'') - C(x''') = [n_{bs}^{x''} - n_{bs}^{x'''}]$ has zero row- and column-sums if and only if $x'' \in \mathcal{D}_{\text{cswap}}(x''')$. Moreover, $x'' \in \mathcal{D}_{\text{cswap}}(x''')$ implies $x''_0 \in \mathcal{D}_{\text{cswap}}(x'''_0)$ and $C(x'') - C(x''') = C(x''_0) - C(x'''_0)$.

By the above result (*), the marginal counts of x_0 and x'_0 agree: $n_b^{x_0} = n_b^{x'_0}$ and $n_s^{x_0} = n_s^{x'_0}$ for all b and s . But the interior cells disagree: if $n_{bs}^{x_0} > 0$ then $n_{bs}^{x'_0} = 0$ (and visa versa, swapping x_0 and x'_0). Further $C(x_0) - C(x'_0)$ has positive entries which sum to Δ and negative entries which sum to $-\Delta$, and zero row- and column-sums.

By construction of x_0 and x'_0 , if we can permute x_0 to produce x'_0 then we can use the same permutation to produce x' from x (up to reordering of records). Critically, permutations of x_0 can only derange Δ records (since there are only Δ records in x_0) and indeed must derange Δ records to produce x'_0 (since there are no records in common between x_0 and x'_0). Therefore we have reduced the problem: we need to find a permutation ρ (regardless of the number of records it fixes) such that $C(\rho(x_0)) = C(x'_0)$.

We construct this permutation ρ by induction on $\Delta = d_{\text{HamS}}^r(x, x') = d_{\text{HamS}}^r(x_0, x'_0)$. There are two base cases: The case $\Delta = 1$ is vacuous since $d_{\text{HamS}}^r(x, x') = 1$ implies that x, x' are not in the same data universe. Why? If $\ell_1^r(x, x') = 2$ then $C(x) - C(x')$ only has one or two non-zero cells. But this implies $C(x) - C(x')$ has a row or column with non-zero sum.

For $\Delta = 2$, the result (*) implies that the 2×2 top-left submatrix of $A = C(x_0) - C(x'_0)$ looks like

$$A_{1:2,1:2} = \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix},$$

(up to re-ordering of rows and columns). Therefore, (up to reordering of records) x_0 and x'_0 differ by a single swap: if b, b', s, s' are indices such that $A_{bs} = A_{b's'} = 1$ then define ρ to be the swap of the records (k, l) and (k', l') in x_0 . We have $C(\rho(x_0)) = C(x'_0)$ as desired.

This completes the base cases. Now we will prove the induction step. By (*), we can always re-order the rows and columns of $A = C(x_0) - C(x'_0)$ such that the 2×2 top-left submatrix looks like

$$A_{1:2,1:2} = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix},$$

with $A_{11}, A_{22} > 0$ and $A_{21} < 0$. Define x_1 by swapping the records $(1, 1)$ and $(2, 2)$ in x_0 . Then the top-left submatrix of $A' = C(x_1) - C(x'_0)$ looks like

$$A'_{1:2,1:2} = \begin{bmatrix} A_{11} - 1 & A_{12} + 1 \\ A_{21} + 1 & A_{22} - 1 \end{bmatrix},$$

and the rest of A' is the same as A . If $A_{12} < 0$ then $\ell_1^r(x_1, x'_0) = \ell_1(A') = \ell_1(A) - 4$. If $A_{12} \geq 0$ then $\ell_1(A') = \ell_1(A) - 2$. In both cases, we can use the induction hypothesis to give us a permutation ρ_1 of x_1 which produces x'_0 (up to reordering of records). Define the permutation ρ as the composition of ρ_1 with the swap of $(1, 1)$ and $(2, 2)$. Then $C(\rho(x_0)) = C(x'_0)$ as desired. \square

Proof of Theorem 3.2.4. Fix x and x' in the same data universe $\mathcal{D} \in \mathcal{D}_{\text{cswap}}$. Let $\Delta = d_{\text{HamS}}^r(x, x')$. We

need to prove that $D_{\text{MULT}}[P_x(C(Z)), P_{x'}(C(Z))] \leq \Delta \varepsilon_{\mathcal{D}}$ or equivalently

$$P_x[C(\sigma(x)) = C(z)] \leq \exp(\Delta \varepsilon_{\mathcal{D}}) P_{x'}[C(\sigma(x')) = C(z)],$$

for all possible swapped datasets z , where the probability is over the random permutation σ sampled by the PSA. Since the output $C(Z)$ does not depend on the ordering of the records in the input x , we may without loss of generality reorder the records in x' . Hence, there exists a permutation ρ which fixes exactly $n - \Delta$ records such that $\rho(x') = x$ by Lemma B.3.4.

Since

$$P_x[C(\sigma(x)) = C(z)] = \sum_{z'} P_x[\sigma(x) = z'],$$

where the sum is over datasets z' with $d_{\text{HamS}}^r(z, z') = 0$, it suffices to show

$$P_x[\sigma(x) = z] \leq \exp(\Delta \varepsilon_{\mathcal{D}}) P_{x'}[\sigma(x') = z], \quad (\text{B.2})$$

for all possible swapped datasets z .

Recall

$$b = \max\{0, n_{m..} \mid \text{there are two records with different values in matching stratum } m\}.$$

If $b = 0$, then x and x' only differ by reordering of rows and hence $\varepsilon_{\mathcal{D}} = 0$ satisfies the DP condition (B.2) by Lemma B.3.3. Having taken care of the case $b = 0$, from herein we may assume $b \geq 2$. (The case $b = 1$ is not possible.)

If $p \in \{0, 1\}$ then $\varepsilon_{\mathcal{D}} = \infty$ and the DP condition (B.2) holds vacuously.

All that remains is to prove (B.2) holds in the case where $0 < p < 1$. Since x and x' themselves differ by the permutation ρ , we can permute x to produce z if and only if we can permute x' to produce z . Thus,

either $P_x(\sigma(x) = z)$ and $P_{x'}(\sigma(x') = z)$ are both zero, or they are both non-zero. We need only focus on the case where both probabilities are non-zero.

Recall that any permutation σ selected with non-zero probability by the PSA can be decomposed as $\sigma = \sigma_{\mathcal{M}} \circ \dots \circ \sigma_1$, where σ_m will leave any unit i with matching category $\mathcal{M}_i \neq m$ fixed. Write x_m for the records of x with $\mathcal{M}_i = m$. Because we perform random selection and permutation independently for each stratum m ,

$$\frac{P_x(\sigma(x) = z)}{P_{x'}(\sigma(x') = z)} = \frac{\prod_{m=1}^{\mathcal{M}} P_x(\sigma_m(x_m) = z_m)}{\prod_{m=1}^{\mathcal{M}} P_{x'}(\sigma_m(x'_m) = z_m)}.$$

Thus, to prove (B.2) it suffices to show

$$\frac{P_x(\sigma_m(x_m) = z_m)}{P_{x'}(\sigma_m(x'_m) = z_m)} \leq \exp(\Delta_m \varepsilon_{\mathcal{D}}), \quad (\text{B.3})$$

for all m where $\Delta_m = d_{\text{HamS}}^r(x_m, x'_m)$.

Fix some m . For notation simplicity, whenever it is not essential to indicate the role of m , we will drop the subscript m from herein (until the end when we need to optimize over m). (This is the same as assuming V_{Match} is empty.)

Let $G_{x \rightarrow z} = \{\text{permutation } g : g(x) = z\}$. We use the notation g instead of σ to emphasise that g is not random, while the permutation σ chosen by Algorithm 3.2.1 is random. There is a bijection between $G_{x \rightarrow z}$ and $G_{x' \rightarrow z}$ given by $g \mapsto g \circ \rho$. Since

$$P_x(\sigma(x) = z) = \sum_{g \in G_{x \rightarrow z}} P_x(\sigma = g),$$

we will prove (B.3) by showing

$$P_x(\sigma = g) \leq \exp(\Delta \varepsilon_{\mathcal{D}}) P_{x'}(\sigma = g \circ \rho),$$

for all $g \in G_{x \rightarrow z}$. (Note that this may not obtain the best possible bound for specific x and x' , but it is mathematically easier to bound $P_x(\sigma = g)/P_{x'}(\sigma = g \circ \rho)$ than bound the desired ratio

$$\frac{\sum_{g \in G_{x \rightarrow z}} P_x(\sigma = g)}{\sum_{g \in G_{x \rightarrow z}} P_{x'}(\sigma = g \circ \rho)}$$

directly. Yet in the case where $G_{x \rightarrow z}$ and $G_{x' \rightarrow z}$ are singletons, this approach gives tight bounds.)

Let k_g be the number of records (in category m) which were deranged (i.e. not fixed) by g and let $d(k)$ denote the k -th derangement number (i.e. the number of derangements of size k):

$$\begin{aligned} d(k) &= k! \sum_{j=0}^k \frac{(-1)^j}{j!} \\ &= kd(k-1) + (-1)^k \quad \text{for } k \geq 0. \end{aligned} \tag{B.4}$$

Fix $g \in G_{x \rightarrow z}$ and $g' = g \circ \rho$. We now compute $P_x(\sigma = g)$. The permutation g is sampled in Algorithm 3.2.1 via a two-step procedure. Firstly records are independent selected for derangement with probability p . Suppose that g deranges records $\{i_1, \dots, i_{k_g}\}$. Since we disallow the possibility of selecting only one record,

$$P_x(\text{the selected records are } \{i_1, \dots, i_{k_g}\}) = \frac{p^{k_g}(1-p)^{n-k_g}}{1 - P_x(\text{exactly 1 record selected})}.$$

Secondly we sample uniformly from the set of all derangements of k_g records. Hence, we sample g with probability $[d(k_g)]^{-1}$ and therefore,

$$P_x(\sigma = g) = \frac{p^{k_g}(1-p)^{n-k_g}}{[1 - P_x(\text{exactly 1 record selected})]d(k_g)}.$$

This gives

$$\frac{P_x(\sigma = g)}{P_{x'}(\sigma = g')} = o^\delta \frac{d(k_g - \delta)}{d(k_g)}, \tag{B.5}$$

where $o = p/(1-p)$ and $\delta = k_g - k_{g'}$.

Our aim is now to bound the RHS of (B.5) by $\exp(\Delta\varepsilon_{\mathcal{D}})$. Since g' and g differ only by the permutation ρ (which fixes $n - \Delta$ records), we must have $k_g - \Delta \leq k_{g'} \leq k_g + \Delta$. Therefore, there are at most $2\Delta + 1$ possible cases:

$$\begin{aligned} \delta \in S &= \{ \delta \in \mathbb{Z} \mid -\Delta \leq \delta \leq \Delta \text{ and } (k_g - \delta = 0 \text{ or } 2 \leq k_g - \delta \leq n) \} \\ &= \{ \delta \in \mathbb{Z} \mid \max(-\Delta, k_g - n) \leq \delta \leq \min(\Delta, k_g) \text{ and } \delta \neq k_g - 1 \}. \end{aligned}$$

Suppose $0 < p \leq 0.5$. Since $d(k)$ is non-decreasing (except at $k = 1$ which is not realizable by g or g') and $(1-p)/p \geq 1$, the RHS of (B.5) is maximised when $k_{g'_m} = n_m$ and $k_{g_m} = n_m - \Delta_m$ (i.e. $\delta = -\Delta_m$), in which case

$$\begin{aligned} \frac{P_x(\sigma = g)}{P_{x'}(\sigma = g')} &= o^{-\Delta} \prod_{m=1}^{\mathcal{M}} \frac{d(n_m)}{d(n_m - \Delta_m)} \\ &\leq o^{-\Delta} \prod_{m=1}^{\mathcal{M}} (n_m + 1)^{\Delta_m} \\ &\leq o^{-\Delta} (b+1)^{\Delta} \\ &= \exp(\Delta\varepsilon_{\mathcal{D}}), \end{aligned} \tag{B.6}$$

for $\varepsilon_{\mathcal{D}} = \ln(b+1) - \ln o$. (The second line uses Lemma B.3.5 which is given below this proof.)

Now suppose $0.5 < p < 1$. In the case of $\delta_m = \Delta_m$, the ratio (B.5) is maximised at o^{Δ_m} when $k_{g_m} = \Delta_m = 2$. Moreover, o^{Δ_m} also dominates $o^{\delta_m} \frac{d(k_{g_m} - \delta_m)}{d(k_{g_m})}$ for all $0 \leq \delta_m \leq \Delta_m$ and all possible k_{g_m} . Thus,

$$\frac{P_x(\sigma = g)}{P_{x'}(\sigma = g')} \leq \prod_{m=1}^{\mathcal{M}} \max \left\{ o^{\Delta_m}, o^{\delta_m} \frac{d(k_{g_m} - \delta_m)}{d(k_{g_m})} : \delta_m \in S_m \text{ and } \delta_m < 0 \right\}$$

$$\begin{aligned}
&\leq \prod_{m=1}^{\mathcal{M}} \max \left\{ o^{\Delta_m}, o^{\delta_m} (k_{g_m} - \delta_m + 1)^{-\delta_m} : \delta_m \in S_m \text{ and } \delta_m < 0 \right\} \\
&\leq \prod_{m=1}^{\mathcal{M}} \max \left\{ o^{\Delta_m}, o^{-\delta_m} (n_m + 1)^{\delta_m} : 0 < \delta_m \leq \Delta_m \right\} \\
&\leq \max \left\{ o^{\Delta}, o^{-\delta} (b + 1)^{\delta} : 0 < \delta \leq \Delta \right\}.
\end{aligned}$$

If $o^{-1}(b + 1) \geq 1$ then $o^{-\delta}(b + 1)^{\delta}$ is maximised at $\delta = \Delta$. Otherwise $o^{-\delta}(b + 1)^{\delta} < 1 < o^{\Delta}$. Hence

$$\frac{P_x(\sigma = g)}{P_{x'}(\sigma = g')} \leq \exp(\Delta \varepsilon_{\mathcal{D}}), \quad (\text{B.7})$$

for $\varepsilon_{\mathcal{D}} = \max \{ \ln o, \ln(b + 1) - \ln o \}$. Combining (B.6) and (B.7), we have

$$\varepsilon_{\mathcal{D}} = \begin{cases} \ln(b + 1) - \ln o & \text{if } 0 < p \leq 0.5 \text{ and } b > 0, \\ \max \{ \ln o, \ln(b + 1) - \ln o \} & \text{if } 0.5 < p < 1 \text{ and } b > 0. \end{cases} \quad (\text{B.8})$$

When $b > 0$, we have $b \geq 2$ and hence also $\max \{ \ln o, \ln(b + 1) - \ln o \} = \ln(b + 1) - \ln o$ for $0.5 < p \leq \sqrt{b + 1}/(\sqrt{b + 1} + 1)$. Thus, (B.8) simplifies to

$$\varepsilon_{\mathcal{D}} = \begin{cases} \ln(b + 1) - \ln o & \text{if } 0 < p \leq \frac{\sqrt{b+1}}{\sqrt{b+1}+1} \text{ and } b > 0, \\ \ln o & \text{if } \frac{\sqrt{b+1}}{\sqrt{b+1}+1} < p < 1 \text{ and } b > 0. \end{cases}$$

as required. □

Lemma B.3.5. *For any $k \in \mathbb{N}$ and any $a \in \mathbb{N}$ satisfying $0 \leq a \leq k$ and $a \neq k - 1$,*

$$\frac{d(k)}{d(k - a)} \leq (k + 1)^a,$$

where $d(k)$ is the number of derangements of k elements (see equation (B.4)).

Proof. We use induction on k . The base cases $k = 0, 1, 2$ are straightforward to verify since $d(0) =$

$d(2) = 1$ and $d(1) = 0$. For the induction step, we can assume $m \geq 3$ so that $d(k-1) \geq 1$ and hence

$$\begin{aligned} \frac{d(k)}{d(k-a)} &= \frac{d(k)}{d(k-1)} \frac{d(k-1)}{d(k-a)} \\ &\leq \frac{d(k)}{d(k-1)} k^{a-1} \end{aligned}$$

by the induction hypothesis. The result then follows by the identity (B.4):

$$\begin{aligned} \frac{d(k)}{d(k-1)} &= \frac{kd(k-1) + (-1)^k}{d(k-1)} \\ &\leq k + 1. \end{aligned} \quad \square$$

B.4 OPTIMALITY OF THEOREM 3.2.4

Throughout this appendix we make the following assumptions. Following Proposition 1, we may assume there is a single matching variable, a single non-matching holding variable and a single swapping variable. Let \mathcal{M} , \mathcal{H} and \mathcal{S} be the domains for the matching variable, the non-matching holding variable and the swapping variable respectively. Define $\mathcal{X}_\times = \bigcup_{k=1}^\infty (\mathcal{M} \times \mathcal{H} \times \mathcal{S})^k$. (Note $\mathcal{X}_{\text{CEF}} \subset \mathcal{X}_\times$, but we cannot assume the reverse inclusion.)

Recall that $b = \max\{0, n_{m..} \mid \text{there are two records with different values in matching stratum } m \in \mathcal{M}\}$; that $o = p/(1-p)$; and that $d(k)$ denotes the k -th derangement number (see equation (B.4)).

Theorem B.4.1. *Assume that $|\mathcal{H}|, |\mathcal{S}| \geq 2$ (so that $\mathcal{D} \in \mathcal{D}_{\text{cswap}}$ are not all singletons and swapping is not completely vacuous).*

Suppose that the PSA satisfies $\varepsilon_{\mathcal{D}}\text{-DP}(\mathcal{X}_\times, \mathcal{D}_{\text{cswap}}, d_{\text{HAMS}}^r, D_{\text{MULT}})$. Then:

(A) *If $p \in \{0, 1\}$, then there exists a universe $\mathcal{D}_0 \in \mathcal{D}_{\text{cswap}}$ such that $\varepsilon_{\mathcal{D}_0} = \infty$.*

(B) *If $0 < p < 1$, then there exists a universe $\mathcal{D}_0 \in \mathcal{D}_{\text{cswap}}$ such that $\varepsilon_{\mathcal{D}_0} \geq \ln o$.*

(C) If $0 < p < 1$, then there exists a universe $\mathcal{D}_0 \in \mathcal{D}_{\text{cSwap}}$ such that $\varepsilon_{\mathcal{D}_0} \geq 0.5 \ln[d(b)/d(b-2)] - \ln(o)$.

The above values of $\varepsilon_{\mathcal{D}_0}$ describe lower bounds on the privacy loss of the PSA; any DP specification for the PSA must have a privacy loss budget at least equal to these values. Comparing these lower bounds to the privacy loss budget $\varepsilon_{\mathcal{D}}^{(1)}$ given in Theorem 3.2.4 shows that $\varepsilon_{\mathcal{D}}^{(1)}$ is optimal in the weak sense that there exists universes \mathcal{D}_0 for which $\varepsilon_{\mathcal{D}_0}^{(1)}$ is arbitrarily close to the best possible budget $\varepsilon_{\mathcal{D}_0}^{(\text{inf})}$.

Theorem B.4.2. Assume $|\mathcal{H}|, |\mathcal{S}| \geq 4$. For each $\mathcal{D}_0 \in \mathcal{D}_{\text{cSwap}}$, define

$$\varepsilon_{\mathcal{D}_0}^{(\text{inf})} = \inf\{\varepsilon_{\mathcal{D}_0} \mid \text{the PSA satisfies } \varepsilon_{\mathcal{D}}\text{-DP}(\mathcal{X}_{\times}, \mathcal{D}_{\text{cSwap}}, d_{\text{HamS}}^r, D_{\text{MULT}})\}.$$

(That is, $\varepsilon_{\mathcal{D}}^{(\text{inf})}$ is the pointwise infimum over all privacy loss budgets $\varepsilon_{\mathcal{D}}$ satisfied by the PSA.) Then $\varepsilon_{\mathcal{D}}^{(\text{inf})}$ is the smallest budget under which the PSA satisfies the DP flavor $(\mathcal{X}_{\times}, \mathcal{D}_{\text{cSwap}}, d_{\text{HamS}}^r, D_{\text{MULT}})$.

Let $\varepsilon_{\mathcal{D}}^{(1)}$ be the privacy loss budget given in Theorem 3.2.4. There exists $\mathcal{D}_0 \in \mathcal{D}_{\text{cSwap}}$ such that

$$\varepsilon_{\mathcal{D}_0}^{(1)} - \varepsilon_{\mathcal{D}_0}^{(\text{inf})} \leq \begin{cases} f(b) & \text{if } 0 < p < \frac{\sqrt{b+1}}{\sqrt{b+1}+1} \text{ and } b > 0, \\ 0 & \text{otherwise,} \end{cases}$$

where

$$f(b) = \frac{1}{2} \ln \left[\frac{(b+1)^2}{b(b-1)} \frac{1 + \frac{e}{2(b-2)!}}{1 - \frac{e}{2b!}} \right],$$

is a positive, monotonically decreasing function for $b \geq 2$ which converges to zero, and satisfies, for example,

$$f(b) \leq 0.148 \text{ for all } b \geq 10.$$

We emphasize that this is a weak form of optimality. A budget $\varepsilon_{\mathcal{D}}$ can be tight at the level of the output (in the sense that $\frac{\mathbb{P}_x(C(\sigma)(x)=z)}{\mathbb{P}_{x'}(C(\sigma)(x')=z)} = \exp[\varepsilon_{\mathcal{D}} d_{\text{HamS}}^r(x, x')]$ for all $x, x' \in \mathcal{D}$, all z and all \mathcal{D}); or at the level of

the data (in the sense that $D_{\text{MULT}}(\mathbf{P}_x, \mathbf{P}_{x'}) = \varepsilon_{\mathcal{D}} d_{\text{HamS}}^r(x, x')$ for all $x, x' \in \mathcal{D}$ and all \mathcal{D}); or at the level of the universe (in the sense that $D_{\text{MULT}}(\mathbf{P}_x, \mathbf{P}_{x'}) = \varepsilon_{\mathcal{D}} d_{\text{HamS}}^r(x, x')$ for some $x, x' \in \mathcal{D}$, and all $\mathcal{D} \in \mathcal{D}_{\text{cSwap}}$). The optimality of Theorem 3.2.4 is weaker than any of these notions; all we have shown is that, for all $\delta > 0$, there exists some $\mathcal{D}_0 \in \mathcal{D}_{\text{cSwap}}$ and some $x, x' \in \mathcal{D}_0$ such that $\varepsilon_{\mathcal{D}_0}^{(1)} - D_{\text{MULT}}(\mathbf{P}_x, \mathbf{P}_{x'}) < \delta$. Part of the sub-optimality arises from the fact that $\varepsilon_{\mathcal{D}}^{(1)}$ is a function only of p and b . We could perform a tighter analysis of the PSA by allowing $\varepsilon_{\mathcal{D}}$ to depend on \mathcal{D} in more complex ways (i.e. by allowing $\varepsilon_{\mathcal{D}}$ to be a function of other properties of \mathcal{D} , not just b).

Proof of Theorem B.4.1. Result (A) follows from Propositions 2 and 3. Result (B) follows from Proposition 4. Result (C) follows from Propositions 5 and 6. \square

Proof of Theorem B.4.2. Because the multiverse $\mathcal{D}_{\text{cSwap}}$ partitions \mathcal{X}_{\times} , the DP constraint imposed on each universe \mathcal{D} is independent of the constraint on another universe $\mathcal{D}' \neq \mathcal{D}$. Hence the PSA does indeed satisfy $\varepsilon_{\mathcal{D}}^{(\text{inf})}$ -DP($\mathcal{X}_{\times}, \mathcal{D}_{\text{cSwap}}, d_{\text{HamS}}^r, D_{\text{MULT}}$). Clearly, $\varepsilon_{\mathcal{D}}^{(\text{inf})} \leq \varepsilon_{\mathcal{D}}$ holds for all \mathcal{D} and all budgets $\varepsilon_{\mathcal{D}}$ for which the PSA satisfies $\varepsilon_{\mathcal{D}}$ -DP($\mathcal{X}_{\times}, \mathcal{D}_{\text{cSwap}}, d_{\text{HamS}}^r, D_{\text{MULT}}$). Hence $\varepsilon_{\mathcal{D}}^{(\text{inf})}$ is the smallest budget for which the PSA satisfies the DP flavor ($\mathcal{X}_{\times}, \mathcal{D}_{\text{cSwap}}, d_{\text{HamS}}^r, D_{\text{MULT}}$).

Moving on to the second half of the theorem, we have by Theorem B.4.1 that

$$\varepsilon_{\mathcal{D}_0}^{(1)} - \varepsilon_{\mathcal{D}_0}^{(\text{inf})} = 0,$$

if $b = 0$ or $p = 0$ or $\frac{\sqrt{b+1}}{\sqrt{b+1}+1} \leq p \leq 1$. On the other hand, if $0 < p < \frac{\sqrt{b+1}}{\sqrt{b+1}+1}$ and $b > 0$, then

$$\begin{aligned} \varepsilon_{\mathcal{D}_0}^{(1)} - \varepsilon_{\mathcal{D}_0}^{(\text{inf})} &\leq \ln(b+1) - \frac{1}{2} \ln[d(b)/d(b-2)] \\ &= \frac{1}{2} \ln \left[(b+1)^2 \frac{\left\lfloor \frac{(b-2)!}{e} + \frac{1}{2} \right\rfloor}{\left\lfloor \frac{b!}{e} + \frac{1}{2} \right\rfloor} \right] \end{aligned}$$

$$\leq \frac{1}{2} \ln \left[\frac{(b+1)^2}{b(b-1)} \frac{1 + \frac{e}{2(b-2)!}}{1 - \frac{e}{2b!}} \right]$$

$$= f(b),$$

where the first line follows by Proposition 5 and the second line by the identity $d(k) = \lfloor \frac{k!}{e} + \frac{1}{2} \rfloor$. The second term inside the logarithm

$$\frac{1 + \frac{e}{2(b-2)!}}{1 - \frac{e}{2b!}}$$

has a numerator which decreases with b and a denominator which increases. Hence this term is monotonically decreasing. The first term inside the logarithm $\frac{(b+1)^2}{b(b-1)}$ has negative first derivative and hence is also decreasing. Therefore, $f(b)$ is monotonically decreasing. Moreover, $f(b)$ is positive and converges to zero because both terms inside the logarithm are greater than one and converge to one. \square

Proposition 2. *Suppose $p = 0$ and $|\mathcal{H}|, |\mathcal{S}| \geq 2$. Then there exists $\mathcal{D}_0 \in \mathcal{D}_{\text{cSwap}}$ such that $C(x) \neq C(x')$ for some $x, x' \in \mathcal{D}_0$. Hence the PSA does not satisfy $\varepsilon_{\mathcal{D}}\text{-DP}(\mathcal{X}_{\times}, \mathcal{D}_{\text{cSwap}}, d_{\text{HamS}}^r, D_{\text{MULT}})$ for any finite $\varepsilon_{\mathcal{D}_0}$ and any such \mathcal{D}_0 .*

Proof. First we show that such a universe $\mathcal{D}_0 \in \mathcal{D}_{\text{cSwap}}$ exists. Given $|\mathcal{H}|, |\mathcal{S}| \geq 2$, the datasets $[(m, b, s), (m, b', s')]$ and $[(m, b, s'), (m, b', s)]$ (for any choice of $m \in \mathcal{M}$, $b \neq b' \in \mathcal{H}$ and $s \neq s' \in \mathcal{S}$) are in the same universe $\mathcal{D}_0 \in \mathcal{D}_{\text{cSwap}}$ and satisfy $C(x) \neq C(x')$.

Let $x, x' \in \mathcal{X}_{\times}$ be datasets which are in the same universe \mathcal{D}_0 . Suppose $C(x) \neq C(x')$. If $p = 0$ then the permutation σ sampled by the PSA must be the identity. Thus, $P_{x'}(C(\sigma(x')) = C(x)) = 0$ but $P_x(C(\sigma(x)) = C(x)) = 1$. Since $d_{\text{HamS}}^r(x, x') < \infty$, the DP condition

$$P_x(C(\sigma(x)) = C(x)) \leq \exp[d_{\text{HamS}}^r(x, x')\varepsilon_{\mathcal{D}_0}]P_{x'}(C(\sigma(x')) = C(x)),$$

cannot be satisfied by a finite $\varepsilon_{\mathcal{D}_0}$. □

Proposition 3. *Suppose $p = 1$ and $|\mathcal{H}|, |\mathcal{S}| \geq 2$. Then there exists $\mathcal{D}_0 \in \mathcal{D}_{\text{cswap}}$ with $n_{m_0 b_0} = n_{m_0 b'_0} = n_{m_0 s_0} = n_{m_0 s'_0} = 1$ for some $m_0 \in \mathcal{M}$, $b_0 \neq b'_0 \in \mathcal{H}$ and $s_0 \neq s'_0 \in \mathcal{S}$. Hence the PSA does not satisfy $\varepsilon_{\mathcal{D}}\text{-DP}(\mathcal{X}_{\times}, \mathcal{D}_{\text{cswap}}, d_{\text{HamS}}^r, D_{\text{MULT}})$ for any finite $\varepsilon_{\mathcal{D}_0}$ and any such \mathcal{D}_0 .*

Proof. The universe given in the proof of Proposition 2 satisfies the property: $n_{m_0 b_0} = n_{m_0 b'_0} = n_{m_0 s_0} = n_{m_0 s'_0} = 1$ for some $m_0 \in \mathcal{M}$, $b_0 \neq b'_0 \in \mathcal{H}$ and $s_0 \neq s'_0 \in \mathcal{S}$.

Now take any $\mathcal{D}_0 \in \mathcal{D}_{\text{cswap}}$ which satisfies this property. Then there exists $x, x' \in \mathcal{D}_0$ which differ by a single swap between (m_0, b_0, s_0) and (m_0, b'_0, s'_0) – that is,

$$x = [(m_0, b_0, s_0), (m_0, b'_0, s'_0), x_{3:n}],$$

$$x' = [(m_0, b_0, s'_0), (m_0, b'_0, s_0), x_{3:n}],$$

where $x_{3:n} = [(M_i, H_i, S_i), i = 3, \dots, n]$. Then $n_{m_0 b_0 s_0}^x = n_{m_0 b'_0 s'_0}^x = 1$ and $n_{m_0 b s_0}^x = n_{m_0 b_0 s}^x = 0$ for all $b \neq b_0$ and all $s \neq s_0$. Since no records can be fixed by σ when $p = 1$, we have $n_{m_0 b_0 s_0}^{\sigma(x)} = 0$ for any possible σ and hence $\mathbb{P}_x(C(\sigma(x)) = C(x)) = 0$ but $\mathbb{P}_x(C(\sigma(x')) = C(x)) > 0$. □

Proposition 4. *Suppose that $0 < p < 1$ and $|\mathcal{H}|, |\mathcal{S}| \geq 2$. Then there exists $\mathcal{D}_0 \in \mathcal{D}_{\text{cswap}}$ and $m_0 \in \mathcal{M}$ such that $n_{m_0} \geq 2$ and $n_{m_0 b}, n_{m_0 s} \in \{0, 1\}$ for all $b \in \mathcal{H}$ and $s \in \mathcal{S}$. A necessary condition for the PSA to satisfy $\varepsilon_{\mathcal{D}}\text{-DP}(\mathcal{X}_{\times}, \mathcal{D}_{\text{cswap}}, d_{\text{HamS}}^{bb}, D_{\text{MULT}})$ is that $\varepsilon_{\mathcal{D}_0} \geq \ln o$ for any such \mathcal{D}_0 .*

Proof. Let $x, x' \in \mathcal{D}_0$ with $d_{\text{HamS}}^r(x_{m_0}, x'_{m_0}) = 2$ and $d_{\text{HamS}}^r(x_m, x'_m) = 0$ for all $m \neq m_0$. (Such a pair of datasets exist because $n_{m_0} \geq 2$.) Reorder the records in x' so that there exists a permutation ρ which deranges exactly two records and satisfies $\rho(x') = x$. (Such a permutation exists by Lemma B.3.4.)

Because $n_{m_0b}, n_{m_0s} \in \{0, 1\}$ for all $m \in \mathcal{M}$ and $s \in \mathcal{S}$, there are no vacuous swaps in the m_0 stratum.

That is, $g(x_{m_0}) \neq x_{m_0}$ for all permutations g which are not the identity id . Hence $G_{x_{m_0} \rightarrow x_{m_0}} = \{\text{id}\}$.

Thus,

$$\begin{aligned} \frac{P_x(C(\sigma(x)) = C(x))}{P_{x'}(C(\sigma(x')) = C(x))} &= \frac{P_x(C(\sigma_{m_0}(x_{m_0})) = C(x_{m_0}))}{P_{x'}(C(\sigma_{m_0}(x'_{m_0})) = C(x_{m_0}))} \\ &= \frac{P_x(\sigma_{m_0} = \text{id})}{P_{x'}(\sigma_{m_0} = \rho)} \\ &= o^{-2}. \end{aligned}$$

Hence, $P_{x'}(C(\sigma(x')) = C(x)) \leq \exp[d_{\text{HamS}}^r(x, x')\varepsilon_{\mathcal{D}_0}]P_x(C(\sigma(x)) = C(x))$ if and only if $\varepsilon_{\mathcal{D}_0} \geq \ln o$. \square

Proposition 5. *Suppose that $0 < p < 1$ and $|\mathcal{H}|, |\mathcal{S}| \geq 4$. Then there exists $\mathcal{D}_0 \in \mathcal{D}_{\text{cSwap}}$ which has the following properties:*

$$\max_b n_{m_0b} \leq \frac{b}{2} - 1 \text{ and } \max_s n_{m_0s} \leq \frac{b}{2} - 1, \quad (\text{B.9})$$

for some $m_0 \in \mathcal{M}$ with $n_{m_0} = b$, and there exists $b_1 \neq b_2 \in \mathcal{H}$ and $s_1 \neq s_2 \in \mathcal{S}$ such that

$$n_{m_0b_1} = n_{m_0b_2} = n_{m_0s_1} = n_{m_0s_2} = 1. \quad (\text{B.10})$$

A necessary condition for the PSA to satisfy $\varepsilon_{\mathcal{D}}\text{-DP}(\mathcal{X}_{\times}, \mathcal{D}_{\text{cSwap}}, d_{\text{HamS}}^r, D_{\text{MULT}})$ is that

$$\varepsilon_{\mathcal{D}_0} \geq 0.5 \ln[d(b)/d(b-2)] - \ln(o),$$

for any \mathcal{D}_0 satisfying the above properties.

We will use the following two lemmata in the proof of Proposition 5.

Lemma B.4.3. *For any x and any permutation g ,*

$$d_{\text{HamS}}^r(x, g(x)) \leq k_g,$$

where k_g is the number of records which are deranged by g .

Proof. For every record (M_i, H_i, S_i) permuted by g , the counts in the fully-saturated contingency table can change by at most 2: the count $n_{M_i H_i S_i}$ will decrease by (at most) 1 and the count $n_{M_i H_i, S_{g(i)}}$ will increase by (at most) 1. Thus, in sum, the counts n_{mbs} can change by at most $2k_g$. That is,

$$\ell_1^r(x, g(x)) = \sum_{m,b,s} \left| n_{mbs}^x - n_{mbs}^{g(x)} \right| \leq 2k_g.$$

The desired result then follows by Lemma B.3.1. □

Lemma B.4.4. *Suppose that $\mathcal{D}_0 \in \mathcal{D}_{\text{cswap}}$ satisfies (B.9). Then there exists $x \in \mathcal{D}_0$ and a derangement g of x_{m_0} such that*

$$d_{\text{HamS}}^r(x_{m_0}, g(x_{m_0})) = b.$$

(In fact, such an x and g exist if and only if \mathcal{D}_0 satisfies (B.9).)

Proof. We suppress the subscript m_0 in x_{m_0} throughout the proof.

We begin by consider the cases $b = 1$ and $b = 0$ individually. Equation (B.9) implies that $b \neq 1$. Similarly, no derangement of x exists when $b = 1$. In the case of $b = 0$, the result is also trivial. Hence we may assume throughout that $b \geq 2$.

“ \Rightarrow ”: Suppose that \mathcal{D}_0 does not satisfy (B.9). Then $n_{m_0} < b$ or there exists (WLOG) a swapping category s_0 such that $n_{m_0 s_0} \geq b/2$. In the first case, any permutation g of x deranges at most n_{m_0} records and hence $d_{\text{HamS}}^r(x, g(x)) < b$ by Lemma B.4.4. By the pigeonhole principle, the second case implies every derangement g of x_{m_0} must send a record with swapping value s_0 to a record which also has value s_0 . Yet

the counts n_{mbs} are unaffected by permutations of records within the same swapping category s . Hence

$$\sum_{b,s} \left| n_{mbs}^x - n_{mbs}^{g(x)} \right| \leq 2(k_g - 1) < 2b,$$

(where the first inequality follows by the reasoning in the proof of Lemma B.4.3). The desired result then follows by Lemma B.3.1.

“ \Leftarrow ”: Assume for now that b is even. By equation (B.9), there exists $x \in \mathcal{D}_0$ whose records are ordered so that every odd record has a different V_{Swap} and V_{Hold} compared to the subsequent record. That is, $H_i \neq H_{i+1}$ and $S_i \neq S_{i+1}$ for all odd i . (One can construct x by picking any $x' \in \mathcal{D}_0$, ordering the records of $x' \in \mathcal{D}_0$ so that the values of V_{Hold} differ between consecutive records, and then permuting V_{Swap} so that their values also differ between consecutive records.)

Construct g by swapping odd and even records:

$$g(i) = \begin{cases} i + 1 & \text{if } i \text{ odd,} \\ i - 1 & \text{if } i \text{ even.} \end{cases}$$

Then $k_g = b$ and $d_{\text{HamS}}^r(g(x), x) = b$.

Now suppose that b is odd. Then equation (B.9) implies that there exists $x \in \mathcal{D}_0$ such that

- 1) $H_i \neq H_{i+1}$ and $S_i \neq S_{i+1}$ for all odd $i < n_{m_0}$; and
- 2) $H_{n_{m_0}} \notin \{H_{n_{m_0}-1}, H_{n_{m_0}-2}\}$ and $S_{n_{m_0}} \notin \{S_{n_{m_0}-1}, S_{n_{m_0}-2}\}$.

Why is this true? We already know that 1) must be true by the proof for even b . Suppose that 2) is not true for any x . Then it must not be true for any x' which are just reorderings of the records of x . Hence, for every adjacent pair $(i, i + 1)$ (with $i < n_{m_0}$ odd), we must have $H_{n_{m_0}} \in \{H_i, H_{i+1}\}$ or $S_{n_{m_0}} \in \{S_i, S_{i+1}\}$. Yet this would contradict equation (B.9).

Construct g by swapping odd and even records, bar the final three records, which are permuted. That is,

$$g(i) = \begin{cases} i + 1 & \text{if } i < n_{m_0} \text{ odd,} \\ i - 1 & \text{if } i < n_{m_0} - 1 \text{ even,} \\ n_{m_0} & \text{if } i = n_{m_0} - 1, \\ n_{m_0} - 2 & \text{if } i = n_{m_0}. \end{cases}$$

As before, $k_g = b$ and $d_{\text{HamS}}^r(g(x), x) = b$. □

Proof of Proposition 5. Fix some $\mathcal{D}_0 \in \mathcal{D}_{\text{cSwap}}$ which satisfies the properties (B.9) and (B.10). Such a universe exists when $|\mathcal{H}|, |\mathcal{S}| \geq 4$ because, for example,

$$x = [(m_0, b_1, s_1), (m_0, b_2, s_2), (m_0, b_3, s_3), (m_0, b_3, s_3), (m_0, b_4, s_4), (m_0, b_4, s_4)],$$

satisfies these properties.

Let $\varepsilon_0 = 0.5 \ln[d(b)/d(b-2)] - \ln(o)$. We want to prove that

$$\frac{\mathbb{P}_x(C(\sigma(x)) = C(z))}{\mathbb{P}_{x'}(C(\sigma(x')) = C(z))} = \exp[d_{\text{HamS}}^r(x, x')\varepsilon_0], \quad (\text{B.11})$$

for some $x, x', z \in \mathcal{D}_0$.

We will construct x and x' so that they are identical except within the matching category m_0 . Then by independence between matching categories,

$$\frac{\mathbb{P}_x(C(\sigma(x)) = C(z))}{\mathbb{P}_{x'}(C(\sigma(x')) = C(z))} = \frac{\mathbb{P}_x(C(\sigma_{m_0}(x_{m_0})) = C(z_{m_0}))}{\mathbb{P}_{x'}(C(\sigma_{m_0}(x'_{m_0})) = C(z_{m_0}))}.$$

This justifies dropping the subscript m_0 from x_{m_0} and ignoring records with matching categories not equal

to m_0 throughout the remainder of the proof.

We construct x as follows: The first two records of x are (m_0, b_1, s_1) and (m_0, b_2, s_2) . The remainder of the records satisfy (B.9). Hence construct the remainder of x according to the procedure given in the proof of Lemma B.4.4. Let x' be the same as x , except interchange the values of the swapping variable of the first two records. That is, $x' = [(m_0, b_1, s_2), (m_0, b_2, s_1), x_{3:n}]$.

Lemma B.4.4 implies there exists a permutation g_0 which fixes the first two records and deranges the remaining records such that

$$d_{\text{HamS}}^r(x, g_0(x)) = b - 2.$$

Moreover, for $g'_0 = g_0 \circ (12)$, we have

$$d_{\text{HamS}}^r(x', g'_0(x')) = b.$$

Set $z = g_0(x) = g'_0(x')$.

Now we will prove (B.11) holds for these choices of x, x' and z . We have

$$\frac{\mathbb{P}_x(C(\sigma(x)) = C(z))}{\mathbb{P}_{x'}(C(\sigma(x')) = C(z))} = \frac{\sum_{z' \text{ re-ordering of } z} \mathbb{P}_x(\sigma(x) = z')}{\sum_{z' \text{ re-ordering of } z} \mathbb{P}_{x'}(\sigma(x') = z')}.$$

Fix some z' which is a re-ordering of z – i.e. some z' with $C(z') = C(z)$. We will show that $\frac{\mathbb{P}_x(\sigma(x)=z')}{\mathbb{P}_{x'}(\sigma(x')=z')} = \exp(2\varepsilon_0)$, when assuming that one of the numerator or the denominator is non-zero (which implies the other is also non-zero, since x and x' differ by a single swap). Since both the numerator and denominator are non-zero when $z' = z$, this result will prove (B.11).

We know that $d_{\text{HamS}}^r(z, z') = 0$ and $d_{\text{HamS}}^r(x', z) = b$. Then using the triangle inequality (twice, once for \leq and once for \geq), $d_{\text{HamS}}^r(x', z') = b$. Lemma B.4.3 implies that $k_g = b$ for all $g \in G_{x' \rightarrow z'}$.

By the same reasoning, $d_{\text{HamS}}^r(x, z') = b - 2$. This implies $k_g \geq b - 2$ for all $g \in G_{x \rightarrow z'}$ by Lemma B.4.3.

We now show that, in fact, $k_g = b - 2$. By construction,

$$n_{m_0 b_1 s_1}^x = n_{m_0 b_2 s_2}^x = 1 \text{ and } n_{m_0 b_1 s}^x = n_{m_0 b_2 s}^x = n_{m_0 b s_1}^x = n_{m_0 b s_2}^x = 0,$$

for all $b \notin \{b_1, b_2\}$ and $s \notin \{s_1, s_2\}$. These equations also hold for z and hence also for z' . Thus, all $g \in G_{x \rightarrow z'}$ must fix the first two records and hence $k_g \leq b - 2$.

In the proof of Theorem 3.2.4, we showed that $P_x(\sigma = g)$ only depends on k_g and, furthermore, that

$$\frac{P_x(\sigma = g)}{P_{x'}(\sigma = g')} = \frac{(1-p)^2 d(b)}{p^2 d(b-2)},$$

when $k_g = b - 2$ and $k_{g'} = b$. Thus,

$$\frac{P_x(\sigma(x) = z')}{P_{x'}(\sigma(x') = z')} = \frac{\sum_{g \in G_{x \rightarrow z'}} P_x(\sigma = g)}{\sum_{g' \in G_{x' \rightarrow z'}} P_{x'}(\sigma = g')} = \frac{(1-p)^2 d(b)}{p^2 d(b-2)} = \exp(2\varepsilon_0),$$

since $k_g = b - 2$ for all $g \in G_{x \rightarrow z'}$ and $k_{g'} = b$ for all $g' \in G_{x' \rightarrow z'}$. □

Proposition 6. *Suppose that $0 < p \leq 0.5$ and $|\mathcal{H}|, |\mathcal{S}| \geq 2$. Then there exists $\mathcal{D}_0 \in \mathcal{D}_{\text{cSwap}}$ such that $b = 2$ and*

$$n_{m_0 b_1} = n_{m_0 b_2} = n_{m_0 s_1} = n_{m_0 s_2} = 1, \tag{B.12}$$

for some $m_0 \in \mathcal{M}$ with $n_{m_0} = b$ and some $b_1 \neq b_2$ and $s_1 \neq s_2$.

A necessary condition for the PSA to satisfy $\varepsilon_{\mathcal{D}}\text{-DP}(\mathcal{X}_{\times}, \mathcal{D}_{\text{cSwap}}, d_{\text{HamS}}^r, D_{\text{MULT}})$ is that

$$\varepsilon_{\mathcal{D}_0} \geq 0.5 \ln[d(b)/d(b-2)] - \ln(o),$$

for any such \mathcal{D}_0 .

Proof. Because $|\mathcal{H}|, |\mathcal{S}| \geq 2$, any dataset of the form $[(m_0, b_1, s_1), (m_0, b_2, s_2)]$ satisfies (B.12). Moreover, x is in some universe \mathcal{D}_0 , thereby proving the first half of the proposition. The second half of the

proposition follows by the same reasoning as the proof of Proposition 5 applied to $x = [(m_0, b_1, s_1), (m_0, b_2, s_2)]$ and $x = [(m_0, b_1, s_2), (m_0, b_2, s_1)]$. \square

B.5 PROOF AND DISCUSSION OF THEOREM 3.3.1

Proof of Theorem 3.3.1. We first analyze the TDA for producing the PL file. [Abowd et al. \(2022a\)](#) proves that the household NMF mechanism T_b satisfies ρ -DP($\mathcal{X}_{\text{CEF}}, \{\mathcal{X}_{\text{CEF}}\}, d_{r_{bs}}^{bb}, D_{\text{NoR}}$), where $\rho^2 = 0.07$ and $d_{r_{bs}}^{bb}$ is the input premetric corresponding to bounded DP on household-records (Appendix A.4). But $(\mathcal{X}_{\text{CEF}}, \{\mathcal{X}_{\text{CEF}}\}, d_{r_{bs}}^{bb}, D_{\text{NoR}})$ and $(\mathcal{X}_{\text{CEF}}, \{\mathcal{X}_{\text{CEF}}\}, d_{\text{HamS}}^{bb}, D_{\text{NoR}})$ are equivalent DP flavors by Proposition 27 of [Bun and Steinke \(2016\)](#). Hence T_b satisfies ρ -DP($\mathcal{X}_{\text{CEF}}, \mathcal{D}_{\text{TDA}}, d_{\text{HamS}}^p, D_{\text{NoR}}$) by Propositions 2.4.15 and 2.4.13 with $\rho^2 = 0.07$. We can similarly conclude that T_p satisfies ρ -DP($\mathcal{X}_{\text{CEF}}, \mathcal{D}_{\text{TDA}}, d_{\text{HamS}}^p, D_{\text{NoR}}$) with $\rho^2 = 2.56$. Then by composition, the mechanism $T_{pb} = [T_p, T_b]$ has privacy loss budget $\rho^2 = 0.07 + 2.56 = 2.63$. Proposition 2.4.11 implies the invariants $c_{\text{TDA}}(x_p, x_{bb})$ – considered as a data-release mechanism – satisfies ρ -DP($\mathcal{X}_{\text{CEF}}, \mathcal{D}_{\text{TDA}}, d_{\text{HamS}}^p, D_{\text{NoR}}$) with $\rho^2 = 0$. Therefore, the composed mechanism $T = [T_{pb}, c_{\text{TDA}}]$ has budget $\rho^2 = 2.63$. The second step of the TDA is post-processing on T and hence has the same budget.

The argument for producing the DHC file is almost analogous. The composed mechanism $T_{pb} = [T_p, T_b]$ has budget $\rho^2 = 7.70 + 4.96 = 12.66$. Now the second step of the TDA also uses the PL file P . Hence, this second step is post-processing on the composed mechanism $[T_{pb}, P, c_{\text{TDA}}]$. This composed mechanism has budget $\rho^2 = 12.66 + 2.63 + 0 = 15.29$.

The second half of the theorem follows from Proposition 2.4.11. (Hence it can be generalized from $(\mathcal{X}_{\text{CEF}}, \mathcal{D}', d_{\text{HamS}}^p, D_{\text{NoR}})$ to any DP flavor $(\mathcal{X}_{\text{CEF}}, \mathcal{D}, d_{\mathcal{X}}, D_{\text{Pr}})$ satisfying the assumptions of this proposition.) \square

The second step of the TDA requires access to both the NMF $[T_p(x_p), T_b(x_{bb})]$ and the invariant statistics $c_{\text{TDA}}(x_p, x_{bb})$ computed on the Census Edited File. Under the DP flavor $(\mathcal{X}_{\text{CEF}}, \{\mathcal{X}_{\text{CEF}}\}, d_{\mathcal{X}}, D_{\text{Pr}})$, the invariant statistics $c_{\text{TDA}}(x_p, x_{bb})$ cannot be released with finite budget. So the second step of the TDA is not post-processing (in the sense given in Section 2.5) under this flavor – it is only post-processing when conditioning on the invariants. In fact, the second half of Theorem 3.3.1 shows that any argument which relies on TDA’s second step being post-processing must necessarily use a DP flavor which conditions on the invariants c_{TDA} .

It is also necessary to use person-records as the resolution of the Hamming distance in the TDA’s DP specification. While the household mechanism T_b satisfies $(\mathcal{X}_{\text{CEF}}, \mathcal{D}_{c_{\text{TDA}}}, d_{\text{HamS}}^{bb}, D_{\text{NoR}})$, the sensitivity of the person-level query Q_p due to a single change in a household record can be very large. (In the Census Edited File, the maximum possible household size is 99,999 (Population Reference Bureau and US Census Bureau’s 2020 Census Data Products and Dissemination Team, 2023).) This means T_p can satisfy $(\mathcal{X}_{\text{CEF}}, \mathcal{D}_{c_{\text{TDA}}}, d_{\text{HamS}}^{bb}, D_{\text{NoR}})$ only with a very large amplification in the privacy loss budget.

B.6 THE 2010 US CENSUS DISCLOSURE AVOIDANCE SYSTEM

This appendix collates information about the 2010 disclosure avoidance system (DAS) which has been made public by the US Census Bureau. Most of this information also applies to the 2000 DAS – as it was very similar to the 2010 DAS – but likely not to the 1990 DAS, which used a significantly different data swapping procedure (McKenna, 2018).

The main references are McKenna (2018); McKenna and Haubach (2019) and Abowd (2021), with additional information spread across various other USCB publications (Zayatz et al., 2010; Zayatz, 2003; Hawala, 2008; US Census Bureau, 2022a, 2021c; Hawes, 2021b; Zayatz, 2007; Lauger et al., 2014; Garfinkel,

2019; Hawes et al., 2021; Steel and Zayatz, 2003; Lemons et al., 2015). However, the publicly available documentation on the 2010 DAS is deliberately incomplete as some implementation details have been deemed confidential by the USCB due to concerns that they may allow the privacy protections of the 2010 DAS to be undermined. We are not the only researchers external to the USCB who have attempted to reproduce the 2010 DAS (Kim, 2015; Radway and Christ, 2023; Christ et al., 2022; Keyes and Flaxman, 2022); however, we believe this documentation is the most comprehensive of those that are currently publicly available.

The primary protection method of the 2010 DAS was data swapping. Special tabulations had additional rules-based protections (see McKenna (2018, Appendix A) for these rules). Synthetic data methods were used to protect the confidentiality of group quarters (GQs) since swapping was infeasible for GQs due to their sparsity and the consequent lack of matching records (Hawala, 2008). These synthetic data methods involved replacing some GQ data with predicted values from a generalized linear model (McKenna, 2018, Section 6.5).

The data swapping procedure for the 2000 and 2010 DAS had three main steps:

- Step 1: A random set S of household records was selected.
- Step 2: Each record in S was paired with a similar, nearby household.
- Step 3: The location of each household in S was swapped with the location of its pair.

We will describe each of these steps in detail below. The data swapping procedure was applied only to households (i.e. ‘occupied housing units’ (US Census Bureau, 2012)) and not to unoccupied housing units or group quarters. At the end of step 3, the DAS swapping procedure outputs a dataset – called the post-swapped dataset – which differs from its input (the Census Edited File) only on the locations of the selected households and their pairs. All publications from the 2010 Census were derived from this

post-swapped dataset (Zayatz et al., 2010; Lauger et al., 2014). (The post-swapped dataset is called the Hundred-percent Detailed File by the USCB.)

Step 1: Household records were randomly selected into the set S with a probability that, for 2010, depended on (possibly amongst other factors):

- (A) The size of the household's block (larger blocks decreased the probability of selection)
- (B) Whether the household contained individuals of a race category not found elsewhere in its block (unique race categories increased the probability of selection)
- (C) The imputation rate within the household's block (higher imputation rates decreased the probability of selection)
- (D) Whether the household was unique within their geographical area on some set of variables (such households were always included in S). (It isn't clear what geographical area was used, but we speculate that it may have been either the household's block group, tract or county.)
- (E) Whether the household record was imputed and what proportion of the record was imputed (McKenna, 2018; Abowd, 2021; McKenna and Haubach, 2019).

Note that (at least in 2000) the selection of records into S was not mutually independent. This was because the number of records in S was capped so that the proportion of swapped records (i.e. the swap rate) was controlled at pre-specified thresholds at the state level (Steel and Zayatz, 2003). (The swap rates for each state were approximately equal (Steel and Zayatz, 2003).) There may have also been other dependencies between households' selection into S .

Exactly how a household's probability of selection was calculated is not public information. However, the USCB has confirmed that in 2010, the marginal selection probability (unconditional on other selections) was zero for totally-imputed households, and was non-zero for all other households (Abowd, 2021; McKenna, 2018; Hawes et al., 2021).¹

Step 2: For each household record in S , the DAS swapping procedure found a household which

¹ However, this appears to be contradicted by another statement from the USCB: "there was a threshold value

- had the same number of adults (over 18 years of age);
- had the same number of minors (under 18 years of age);²
- had the same tenure status;³
- was located in the same state; but
- was located in a different block (Abowd, 2021; Garfinkel, 2019; US Census Bureau, 2021c).⁴

A household which satisfies these requirements is called a matching household. Each record in S was paired with a matching household. In 2000 (and hence plausibly in 2010 as well), the swapping procedure prioritized pairings where

1. the matching record was also in S ; or
2. both records were geographically close (e.g. they were in the same tract or county); or
3. the matching record had a high “disclosure risk” (Steel and Zayatz, 2003).

for not swapping in blocks with a high imputation rate” (McKenna and Haubach, 2019). Assuming that this imputation rate threshold was under 100%, there would be not-totally-imputed household records with zero selection probability.

There is a possible explanation of this contradiction. All not-totally-imputed households may have had the possibility of being swapped (in step 2) even though some of them had zero probability of being selected into S . Yet this would require that, for all the not-totally-imputed households b with zero selection probability, there was a household b' with non-zero selection probability that matched b on the five criteria in step 2 and furthermore that b and b' could possibly be matched given the DAS’s prioritization of certain matches over other matches (e.g. 1.-3. in step 2). It seems infeasible to guarantee such requirements for all possible Census Edited Files.

²As a consequence, the paired housing units also matched on the occupancy status (occupied versus unoccupied), and total number of persons.

³The 2010 Census classified households’ tenure as either owner-occupied (owned outright), owner-occupied (with a loan or mortgage), renter occupied, or occupied without payment of rent. It is unclear if the swapping procedure matched households on these categories, or only on the broader categories of A. owner-occupied vs B. renter-occupied (including without payment of rent) (US Census Bureau, 2012, 2010).

⁴The public documentation from the USCB is contradictory on whether there were additional requirements beyond the five listed here (Hawes et al., 2021; Abowd, 2021; US Census Bureau, 2022a).

It is possible that there were other criteria for deciding the pairing when there were multiple matching households. It is unclear how these criteria were ranked in their importance. (For example, how did the swapping procedure decide between I. a pair where both records were in S but were in a different counties; and II. a pair where both records were in the same block group but one record was not in S ?) However, it is likely that criteria 1. was considered the most important since it minimizes the number of swaps (Steel and Zayatz, 2003).

Step 3: Steps 1 and 2 produced pairs of household records. These pairs consisted of one record from S along with its matching record found in Step 2 (which may also be in S). In Step 3, all pairs had their locations swapped. More exactly, for each household in S , the value of its block, block group, tract, and county were swapped with the corresponding values of its paired record. (Note that a pair of records might have had the same block group, tract or county, in which case these values did not change. The paired households were always in the same state (Garfinkel, 2019), so this location variable was never swapped.)

B.6.1 COMPARING THE 2010 DAS WITH THE PSA

In this subsection, we compare the 2010 data swapping procedure with the PSA. The PSA is a general algorithm (in the sense that its parameters – such as the swapping and matching variables – are not set but must be chosen). Thus, for the purposes of this comparison, we will consider the PSA using the implementation choices which attempt to mirror the 2010 DAS, as given in Subsection 3.2.5.

There are a number of key similarities between the data swapping procedure in the 2010 DAS and the PSA from Subsection 3.2.5:

1. The *swapping units* (i.e. the records which are swapped) are household-records for both the 2010 DAS and the PSA.
2. *Swap rates:* The swap rate is defined as the fraction of records which were swapped. For the 2010 DAS, the swap rate is the fraction of records which were selected into S or were paired with a record

in S . This rate was explicitly controlled by the US Census Bureau at the state level and all states had approximately the same swap rate (Zayatz, 2003). Although the USCB has not released the value of the 2010 DAS's swap rate, at the national level it is purported to be between 2-4% (boyd and Sarathy, 2022).

In comparison, the PSA controls the expected swap rate (where the expectation is over the randomness in the PSA). An implementer of the PSA cannot precisely fix the swap rate – but only the expected swap rate (via the parameter p). However, when the number of records n is large, the swap rate is typically very close to p , since its variance is approximately $p(1 - p)/n \approx 0$.

Hence, one may set the PSA's parameter p so that the swap rates for the PSA and the 2010 DAS are similar at the state and national levels.

3. The *matching variables* of the 2010 DAS include the household's state, the number of adult occupants, the number of child occupants and the household's tenure status. There may be other matching variables (which have not been disclosed by the USCB), but Abowd (2021) implicitly suggests that this is not the case. The PSA could be implemented with exactly the same matching variables. However, the matching variables of the PSA implementation in Subsection 3.2.5 are the household's state and counts of adults and children – the household's tenancy status was not included. By excluding a matching variable, the PSA from Subsection 3.2.5 has fewer invariants and its privacy loss budget ε is a conservative estimate, compared to a PSA implementation which mirrored the 2010 DAS matching variables.
4. The *swapping variables* of the 2010 DAS and the PSA from Subsection 3.2.5 are the same: the households' county, tract, block group and block are swapped by the 2010 DAS and the PSA. (As we will discuss later in this subsection, the 2010 DAS sometimes used the households' county, tract or block group as matching variables in an adaptive matching procedure. For our purposes, they can still be considered as swapping variables; matching variables can always be swapped since swapping them does not change the data.)

There are a number of significant differences between the PSA and the 2010 swapping procedure:

1. The 2010 DAS *swapped* pairs of records, whereas the PSA *permutes* multiple records. While any permutation is equal to a sequence of multiple pairwise swaps, the 2010 DAS does not allow for such arbitrary swaps. However, permutation swapping (under the name n -Cycle swapping) was actively being investigated by the USCB (DePersio et al., 2012; Lauger et al., 2014) before this work was supplanted by their shift towards DP (McKenna and Haubach, 2019). The USCB found that permutation swapping provided both better data utility and better data protection than the pairwise swapping used in 1990-2010 Censuses; this is corroborated by our DP analysis of permutation swapping.

2. *Swap probabilities*: The probability of a given household being swapped was not constant in 2010 DAS. In fact, swapping was highly targeted to households which were “vulnerable to re-identification” (Hawes et al., 2021). Moreover, the probability of a household being swapped was dependent on whether other households were selected for swapping (for example because the absolute state-wide swap rates were controlled). In comparison, in the PSA, the probability of a household being swapped is constant and independent of other households.
3. *Adaptive matching*: The 2010 DAS paired households according to a complicated matching procedure. For example, they prioritised matching households which shared the same county or tract. (More details on their matching procedure is given in Step 2 of the 2010 DAS description given above.) In essence, this means that sometimes the household’s county or tract were included as matching variables, and sometimes they were not; and whether they were included was a function of the household as well as its matching households. The matching procedure for the PSA is much simpler by comparison: the matching variables are static and the choice of how to swap the selected matching households is made uniformly at random.
4. *Non-vacuous swaps*: A swap is vacuous if it does not change the dataset, except (possibly) by reordering the records. A pairwise swap is not vacuous if and only if the paired records have different values for both their swapping variables V_{Swap} and their holding variables V_{Hold} . It is unclear whether the 2010 DAS prohibited vacuous swaps but we suspect so. On the other hand, vacuous swaps are allowed by the PSA.

B.6.2 MODIFYING THE PSA TO FURTHER ALIGN WITH THE 2010 DAS

We discuss some possible extensions to the PSA in Part III (Bailie et al., 2025d). Those extensions aimed to reduce the PSA’s invariants without foregoing its DP guarantee. In this subsection, we propose four additional extensions to the PSA which are DP while being more faithful to the 2010 DAS. These extensions address the differences between the 2010 DAS and the PSA identified in the previous subsection. We show that these differences can be bridged without losing the guarantee of DP – at the cost of greatly complicating the calculation of the privacy loss budget. We do not attempt these calculations; we only argue why the privacy loss budgets for these extensions remain bounded away from infinity.

First, we address one aspect of the 2010 DAS which cannot be incorporated into a DP swapping mechanism. The 2010 DAS used disjoint, pairwise swapping (Lauger et al., 2014). This is not a transitive action – its orbit space does not equal the universe induced by the swapping invariants – and hence it cannot satisfy differential privacy. (A necessary condition for a mechanism T to be ϵ -DP is that $P_x(T \in \cdot)$ and $P_{x'}(T \in \cdot)$ have common support for all x and x' in the same data universe with $d_{\mathcal{X}}(x, x') < \infty$. When the SDC method is a random group action on x , as is the case for permutation swapping, this condition is equivalent to the group action being transitive.)

VARIABLE SWAPPING PROBABILITIES The PSA uses the same swapping probability p for all records. However, we can modify the PSA to use a different swapping probability p_i for each record i . As long as these probabilities are uniformly bounded away from zero and one (so that p_i^{-1} and $(1 - p_i)^{-1}$ are bounded away from infinity), this modification will satisfy the same DP flavor as the original PSA (with a finite budget). The proof would follow the same strategy as in Appendix B.3; only the final computations would change, as one would need to optimise over $o_i = p_i / (1 - p_i)$ for all i .

NON-UNIFORM PERMUTATIONS The PSA samples derangements of the selected records uniformly at random, whereas the 2010 DAS prioritizes certain swaps over others. We can mirror this aspect of the 2010 DAS by sampling from a non-uniform distribution over the derangements. This would allow for some derangements to be selected with higher probability than other derangements. The advantage here is that some derangements, which result in poor data utility (such as when geographically-distant records are swapped), can be under-sampled; while other, more desirable derangements can be over-sampled. This would mimic the adaptive matching of the 2010 DAS. By reasoning which is analogous to the previous extension, this extension will also retain the PSA's DP flavor, provided that $P_x(\sigma = g) / P_{x'}(\sigma = g')$ is

uniformly bounded by $\exp[O(|k_g - k_{g'}|)]$, for all derangements g and g' (of k_g and $k_{g'}$ records respectively).

PROHIBITING IMPUTED RECORDS FROM BEING SWAPPED The 2010 DAS never swaps records which have been completely imputed. (The rationale is that imputed records do not require privacy protection.) We can modify the PSA so that $p_i = 0$ for all records i which are imputed. Suppose that the records which are imputed are constant. If the PSA satisfies $\varepsilon_{\mathcal{D}}\text{-DP}(\mathcal{X}_{\text{CEF}}, \mathcal{D}_{\text{cSwap}}, d_{\text{HamS}}^{bb}, D_{\text{MULT}})$, then this modification would satisfy $\varepsilon'_{\mathcal{D}}\text{-DP}(\mathcal{X}_{\text{CEF}}, \mathcal{D}_{\text{cSwap}}, d'_{\mathcal{X}}, D_{\text{MULT}})$, where

$$d'_{\mathcal{X}}(x, x') = \begin{cases} d_{\text{HamS}}^{bb}(x, x') & \text{if } x, x' \text{ do not differ on any imputed record,} \\ \infty & \text{otherwise} \end{cases}$$

and $\varepsilon'_{\mathcal{D}} \leq \varepsilon_{\mathcal{D}}$ since the maximum stratum size b is reduced when the imputed records are removed.

PROHIBITING VACUOUS SWAPS A swap (or more generally a permutation) is vacuous if it does not change the dataset, except perhaps by re-ordering the records.

We assume that the 2010 DAS does not allow vacuous swaps. We can similarly prohibit the PSA from allowing vacuous swaps. Instead of sampling derangements uniformly at random, we would put zero probability on vacuous derangements. Under the action of non-vacuous derangements, the orbit space is still the entire data universe. Hence, this modification will still satisfy the same DP flavor as the PSA, however the calculation of the privacy loss budget $\varepsilon_{\mathcal{D}}$ will be difficult as one must optimise $P_x(\sigma = g)/P_{x'}(\sigma = g')$ over all permutations g and g' which are non-vacuous with respect to x and x' , over all $x, x' \in \mathcal{D}$.



Appendices to Chapter 5

C.1 DEFINITION OF $\text{supp}(x \mid t, \theta)$

We assume that \mathcal{X} is equipped with a topology $\tau_{\mathcal{X}}$ and that \mathcal{G} is the Borel σ -algebra induced by $\tau_{\mathcal{X}}$. Denote the support of P_{θ} by

$$\text{supp}(P_{\theta}) = \bigcap \{S \subset \mathcal{X} \text{ closed and Borel measurable} \mid P_{\theta}(S) = 1\}. \quad (\text{C.1})$$

Here we mean ‘closed’ with respect to $\tau_{\mathcal{X}}$, not necessarily the topology induced by the metric d . (Generally we should not use the topology induced by d : Since d is typically discrete, this topology results in $\text{supp}(P_{\theta}) = \emptyset$ whenever \mathcal{X} is uncountable¹ and then Theorems 5.4.1 and 5.6.3 would be vacuous.) A standard example would be $\mathcal{X} = \bigcup_{n \in \mathbb{N}} \mathbb{R}^n$ (i.e. the universe of datasets with a finite number of real-valued

¹If \mathcal{X} is uncountable, then $P_{\theta}(\{x\}) = 0$. Yet $\{x\}^c$ is closed under the discrete topology, and hence $x \notin \text{supp}(P_{\theta})$. This argument applies for all $x \in \mathcal{X}$; thus $\text{supp}(P_{\theta}) = \emptyset$.

records) with the topology induced by the map

$$\begin{aligned}\mathcal{X} &\rightarrow \mathbb{R}^{\mathbb{N}} \\ x &\mapsto (x, 0, 0, \dots),\end{aligned}$$

when $\mathbb{R}^{\mathbb{N}}$ is equipped with the product Euclidean topology. (We assume that \mathcal{G} is Borel to ensure that $\text{supp}(P_\theta) \in \mathcal{G}$.)

Similarly, assume that \mathcal{T} is equipped with a topology $\tau_{\mathcal{T}}$ and that \mathcal{F} is the Borel σ -algebra induced by $\tau_{\mathcal{T}}$. Analogously to (C.1), define $\text{supp}(P_x) \subset \mathcal{T}$ for each $x \in \mathcal{X}$. Write $\text{supp}_0(x \mid t) = \{x \mid t \in \text{supp}(P_x)\}$ and finally define

$$\text{supp}(x \mid t, \theta) = \text{supp}(P_\theta) \cap \text{supp}_0(x \mid t).$$

We assume that $(\mathcal{X}, \tau_{\mathcal{X}})$ and $(\mathcal{T}, \tau_{\mathcal{T}})$ are second countable (that is, there exist countable bases for $\tau_{\mathcal{X}}$ and $\tau_{\mathcal{T}}$) to ensure that

$$P_\theta(E_1) = P_x(E_2) = 0, \tag{C.2}$$

for any measurable $E_1 \subset \text{supp}(P_\theta)^c$ and $E_2 \subset \text{supp}(P_x)^c$.

C.2 THE DENSITY RATIO METRIC IS WELL-DEFINED

Proposition 7. *The density ratio metric d_{DR} is well-defined. That is, $d_{\text{DR}}(\mu, \nu)$ does not depend on the choice of f, g and τ in (5.18).*

For $\mu, \nu \in \Omega$, write $\mu \ll \nu$ to denote that μ is absolutely continuous with respect to ν .

We need the following result, which can be found in a standard probability-theory textbook (such as (Billingsley, 2012, Exercise 32.6)):

Lemma C.2.1. *Let $\mu, \nu, \tau \in \Omega$ such that $\mu \ll \nu$ and $\nu \ll \tau$. Then $\mu \ll \tau$ and the Radon-Nikodym derivative $\frac{d\mu}{d\nu}$ satisfies*

$$\frac{d\mu}{d\nu} = \frac{f}{g}, \quad \mu\text{-a.e.}, \quad (\text{C.3})$$

where f, g are the τ -densities of μ and ν respectively and, on the RHS of (C.3), $0/0 = 0$.

Proof. of Proposition 7: Let $\tau_1, \tau_2 \in \Omega$ and suppose μ and ν are both non-zero and absolutely continuous with respect to both τ_1 and τ_2 . Let f_1 and f_2 be the densities of μ with respect to τ_1 and τ_2 respectively. Similarly, define g_1 and g_2 as the densities of ν with respect to τ_1 and τ_2 .

Define $\tau = \tau_1 + \tau_2$. Then $\tau_1 \ll \tau$ and $\tau_2 \ll \tau$. By Lemma C.2.1,

$$\frac{f_1}{g_1} = \frac{\frac{d\mu}{d\tau} / \frac{d\tau_1}{d\tau}}{\frac{d\nu}{d\tau} / \frac{d\tau_1}{d\tau}} = \frac{\frac{d\mu}{d\tau}}{\frac{d\nu}{d\tau}}, \quad \tau_1\text{-a.e.}$$

Hence

$$\tau_1\text{-ess sup}_{t, t' \in \mathcal{T}^0} \frac{f_1(t) g_1(t')}{g_1(t) f_1(t')} = \tau_1\text{-ess sup}_{t, t' \in \mathcal{T}^0} \frac{\frac{d\mu}{d\tau}(t) \frac{d\nu}{d\tau}(t')}{\frac{d\nu}{d\tau}(t) \frac{d\mu}{d\tau}(t')} \leq \tau\text{-ess sup}_{t, t' \in \mathcal{T}^0} \frac{\frac{d\mu}{d\tau}(t) \frac{d\nu}{d\tau}(t')}{\frac{d\nu}{d\tau}(t) \frac{d\mu}{d\tau}(t')} \quad (\text{C.4})$$

(Note we use the notation τ -ess sup to refer to the essential supremum with respect to the measure τ .)

Now we prove the reverse inequality of (C.4). For any $E \in \mathcal{F}$ with $\tau_1(E) = 0$, we have that

$$\tau\left(\left\{\frac{d\mu}{d\tau} > 0\right\} \cap E\right) = 0. \quad (\text{C.5})$$

(Otherwise $\mu(E) = \int_E \frac{d\mu}{d\tau} d\tau > 0$ and hence $\mu \not\ll \tau_1$.) By symmetry, (C.5) also holds with $\frac{d\nu}{d\tau}$ in place of $\frac{d\mu}{d\tau}$. This implies

$$\tau_1\text{-ess sup}_{t, t' \in \mathcal{T}^0} \frac{f_1(t) g_1(t')}{g_1(t) f_1(t')} \geq \tau\text{-ess sup}_{t, t' \in \mathcal{T}^0} \frac{\frac{d\mu}{d\tau}(t) \frac{d\nu}{d\tau}(t')}{\frac{d\nu}{d\tau}(t) \frac{d\mu}{d\tau}(t')}. \quad (\text{C.6})$$

By combining (C.5) and (C.6), we get an equality between these two essential suprema. By exactly the

same reasoning, we have that

$$\tau_2\text{-ess sup}_{t,t' \in \mathcal{T}^0} \frac{f_2(t) g_2(t')}{g_2(t) f_2(t')} = \tau\text{-ess sup}_{t,t' \in \mathcal{T}^0} \frac{\frac{d\mu}{d\tau}(t) \frac{d\nu}{d\tau}(t')}{\frac{d\nu}{d\tau}(t) \frac{d\mu}{d\tau}(t')},$$

and hence

$$\tau_1\text{-ess sup}_{t,t' \in \mathcal{T}^0} \frac{f_1(t) g_1(t')}{g_1(t) f_1(t')} = \tau_2\text{-ess sup}_{t,t' \in \mathcal{T}^0} \frac{f_2(t) g_2(t')}{g_2(t) f_2(t')}.$$

Since logarithms are continuous, they are interchangeable with essential suprema:

$$\tau_1\text{-ess sup}_{t,t' \in \mathcal{T}^0} \ln \left(\frac{f_1(t) g_1(t')}{g_1(t) f_1(t')} \right) = \ln \left(\tau_1\text{-ess sup}_{t,t' \in \mathcal{T}^0} \frac{f_1(t) g_1(t')}{g_1(t) f_1(t')} \right).$$

This proves that the value of $d_{\text{DR}}(\mu, \nu)$ is the same when computed using f_1, g_1 and τ_1 , as when computed using f_2, g_2 and τ_2 . □

That the density ratio metric d_{DR} is well-defined (Proposition 7) is also an easy corollary of Proposition 12 below.

C.3 METRIC SPACES

Definition C.3.1. A *metric* d on a set S is a function $S \times S \rightarrow [0, \infty]$ that satisfies the following properties for all $x, y, z \in S$:

1. Positive definiteness: $d(x, y) = 0$ if and only if $x = y$;
2. Symmetry: $d(x, y) = d(y, x)$; and
3. Triangle inequality: $d(x, z) \leq d(x, y) + d(y, z)$.

A *premetric* d on a set S is a function $S \times S \rightarrow [0, \infty]$ that satisfies $d(x, x) = 0$ for every $x \in S$.

Note that the co-domain of a metric is the extended, non-negative real numbers; we allow a metric to take the value of positive infinity. Metrics of this kind are sometimes referred to as *extended*-metrics, or ∞ -metrics, to distinguish them from those metrics with co-domain $[0, \infty)$.

Premetrics naturally arise in the context of differential privacy: The distorting function associated with a DP flavour's distortion model is a premetric [Bailie et al. \(2025b\)](#). It is also common in many DP flavours that the ‘metric’ d on \mathcal{X} is not in fact a metric, but only a premetric.

Definition C.3.2. Two metrics d_1 and d_2 (which are defined on the same set S) are *strongly equivalent* ([Carothers, 2000](#), p. 121) if there exist constants $0 < a \leq b < \infty$ such that

$$ad_1(x, y) \leq d_2(x, y) \leq bd_1(x, y),$$

for all $x, y \in S$.

Proposition 13 below proves that the multiplicative distance D_{MULT} and the density ratio metric d_{DR} are strongly equivalent on the space of probability measures (but not on the space Ω of σ -finite measures).

Lemma C.3.3. *The multiplicative distance D_{MULT} is a metric on the collection of measures on $(\mathcal{T}, \mathcal{F})$.*

Proof. Suppose μ, ν and τ are measures on $(\mathcal{T}, \mathcal{F})$. To prove property 1. of a metric, note that $D_{\text{MULT}}(\mu, \nu) = 0$ if and only if $\mu(S) = \nu(S)$ for all $S \in \mathcal{F}$. Yet this holds if and only if $\mu = \nu$. Property 2. follows by observing

$$D_{\text{MULT}}(\mu, \nu) = \sup_{S \in \mathcal{F}} \left| \ln \frac{\mu(S)}{\nu(S)} \right| = \sup_{S \in \mathcal{F}} \left| -\ln \frac{\mu(S)}{\nu(S)} \right| = \sup_{S \in \mathcal{F}} \left| \ln \frac{\nu(S)}{\mu(S)} \right| = D_{\text{MULT}}(\nu, \mu).$$

Property 3. is implied by

$$\begin{aligned} D_{\text{MULT}}(\mu, \nu) &= \sup_{S \in \mathcal{F}} |\ln \mu(S) - \ln \nu(S)| \\ &= \sup_{S \in \mathcal{F}} |\ln \mu(S) - \ln \tau(S) + \ln \tau(S) - \ln \nu(S)| \\ &\leq \sup_{S \in \mathcal{F}} |\ln \mu(S) - \ln \tau(S)| + \sup_{S \in \mathcal{F}} |\ln \tau(S) - \ln \nu(S)| \end{aligned}$$

$$= D_{\text{MULT}}(\mu, \nu) + D_{\text{MULT}}(\tau, \nu).$$

□

Lemma C.3.4. *The density ratio metric d_{DR} is a metric on the space Ω_1 of probability measures on $(\mathcal{T}, \mathcal{F})$.*

Proof. Suppose P, Q and R are probability measures on $(\mathcal{T}, \mathcal{F})$. Proof of property 1.: Suppose that $P = Q$. Then P and Q are mutually absolutely continuous and have P -densities f and g respectively, with

$$f(t) = g(t) = 1,$$

for all $t \in \mathcal{T}$. Thus $d_{\text{DR}}(P, Q) = 0$. Now suppose that $P \neq Q$. If P and Q are not mutually absolutely continuous, then $d_{\text{DR}}(P, Q) = \infty > 0$. If they are mutually absolutely continuous, then they have P -densities f and g respectively. Moreover, because $P \neq Q$, there exists $S \in \mathcal{F}$ such that $P(S) > Q(S)$.

Hence $P(f > g) > 0$. This implies

$$0 < \text{ess sup}_{t \in \mathcal{T}^c} \ln \frac{f(t)}{g(t)}.$$

Also, $P(S^c) < Q(S^c)$ (here S^c is the complement of S in \mathcal{T}), so by exactly the same reasoning

$$0 < \text{ess sup}_{t' \in \mathcal{T}^c} \ln \frac{g(t')}{f(t')}.$$

Combining these two results gives $d_{\text{DR}}(P, Q) > 0$.

Property 2. follows by the symmetry of

$$\frac{f(t)}{f(t')} \bigg/ \frac{g(t)}{g(t')} = \frac{g(t')}{g(t)} \bigg/ \frac{f(t')}{f(t)},$$

and the fact that t and t' are interchangeable in the definition of the density ratio metric.

Finally, we prove property 3. Suppose that P and R are not mutually absolutely continuous. Then Q is

either mutually absolutely continuous with P , or with R , but not both. Hence $d_{\text{DR}}(P, Q) + d_{\text{DR}}(Q, R) = \infty$ and thus

$$d_{\text{DR}}(P, R) \leq d_{\text{DR}}(P, Q) + d_{\text{DR}}(Q, R),$$

holds vacuously. Now suppose that P and R are mutually absolutely continuous. We may also suppose that Q is mutually absolutely continuous with respect to both P and R . (When this is not the case, property 3. again holds vacuously.) Let f, g and h be P -densities of P, Q and R respectively. Define

$$\mathcal{T}_{f,g}^{\circ} = \{t \in \mathcal{T} \mid 0 < f(t), g(t) < \infty\},$$

$$\mathcal{T}_{f,b}^{\circ} = \{t \in \mathcal{T} \mid 0 < f(t), b(t) < \infty\},$$

$$\mathcal{T}_{g,b}^{\circ} = \{t \in \mathcal{T} \mid 0 < g(t), b(t) < \infty\}.$$

Then

$$\begin{aligned} d_{\text{DR}}(P, R) &= \text{ess sup}_{t, t' \in \mathcal{T}_{f,b}^{\circ}} \ln \left(\frac{f(t)}{f(t')} \right) - \ln \left(\frac{b(t)}{b(t')} \right) \\ &= \text{ess sup}_{t, t' \in \mathcal{T}_{f,b}^{\circ}} \ln \left(\frac{f(t)}{f(t')} \right) - \ln \left(\frac{g(t)}{g(t')} \right) + \ln \left(\frac{g(t)}{g(t')} \right) - \ln \left(\frac{b(t)}{b(t')} \right) \\ &\leq \text{ess sup}_{t, t' \in \mathcal{T}_{f,g}^{\circ}} \ln \left(\frac{f(t)}{f(t')} \right) - \ln \left(\frac{g(t)}{g(t')} \right) + \text{ess sup}_{t, t' \in \mathcal{T}_{g,b}^{\circ}} \ln \left(\frac{g(t)}{g(t')} \right) - \ln \left(\frac{b(t)}{b(t')} \right) \\ &= d_{\text{DR}}(P, Q) + d_{\text{DR}}(Q, R), \end{aligned}$$

where all the essential suprema are with respect to P . We can exchange $\mathcal{T}_{f,g}^{\circ}$, $\mathcal{T}_{f,b}^{\circ}$ and $\mathcal{T}_{g,b}^{\circ}$ in the above computations because

$$P(\mathcal{T}_{f,g}^{\circ}) = P(\mathcal{T}_{f,b}^{\circ}) = P(\mathcal{T}_{g,b}^{\circ}) = 1.$$

□

The density ratio metric d_{DR} is not a metric on Ω . While it is easy to verify that properties 2. and 3. hold (follow the same reasoning as in the above proof), d_{DR} does not satisfy property 1: There exists $\mu \neq \nu \in \Omega$ such that $d_{\text{DR}}(\mu, \nu) = 0$. For example, suppose that

$$\nu(S) = 2\mu(S), \quad (\text{C.7})$$

for all $S \in \mathcal{F}$. Then $f(t) = 1$ and $g(t) = 2$ are μ -densities of μ and ν respectively. Yet this implies $d_{\text{DR}}(\mu, \nu) = 0$.

Proposition 8. *The density ratio metric d_{DR} is a pseudo-metric on Ω . That is, d_{DR} is a function from $\Omega \times \Omega$ to the extended real line which satisfies the following properties for all $\mu, \nu, \tau \in \Omega$:*

1. $d_{\text{DR}}(\mu, \mu) = 0$;
2. $d_{\text{DR}}(\mu, \nu) \geq 0$;
3. $d_{\text{DR}}(\mu, \nu) = d_{\text{DR}}(\nu, \mu)$; and
4. $d_{\text{DR}}(\mu, \tau) \leq d_{\text{DR}}(\mu, \nu) + d_{\text{DR}}(\nu, \tau)$.

Proof. Properties 1., 3. and 4. follow by the same reasoning as in the proof of Lemma C.3.4 with minor adjustments to account for the fact that μ, ν or τ may be zero. Property 2. follows by applying Properties 1., 4. and 3. sequentially:

$$0 = d_{\text{DR}}(\mu, \mu) \leq d_{\text{DR}}(\mu, \nu) + d_{\text{DR}}(\nu, \mu) = 2d_{\text{DR}}(\mu, \nu).$$

□

C.4 DISTORTING FUNCTIONS AND DISTORTION MODELS

Definition C.4.1. Montes et al. (2020a) A *distorting function* d_{DIST} is a function $S \times S \rightarrow [0, \infty]$ where S is one of the following spaces:

- the set of all measures on a measurable space $(\mathcal{T}, \mathcal{F})$;
- the set Ω of all σ -finite measures on $(\mathcal{T}, \mathcal{F})$;
- the set of all finite measures on $(\mathcal{T}, \mathcal{F})$;
- the set Ω_1 of all probability measures on $(\mathcal{T}, \mathcal{F})$; or
- the set $\Omega_1^* = \{P \in \Omega_1 \mid P(\{t\}) > 0 \forall t \in \mathcal{T}\}$ of probabilities on $(\mathcal{T}, 2^{\mathcal{T}})$ with non-zero mass on every event $E \subset \mathcal{T}$ (often with the assumption that \mathcal{T} has finite cardinality).

Definition C.4.2. Montes et al. (2020a) Given a distorting function d_{DIST} , a positive constant $r > 0$ (termed the *distortion parameter*), and a probability $P_0 \in \Omega_1$ (the *nucleus*), the distortion model on P_0 associated with d_{DIST} and r is the closed d_{DIST} -ball centred at P_0 with radius r :

$$B_{d_{\text{DIST}}}^r(P_0) = \{P \in \Omega_1 \mid d_{\text{DIST}}(P, P_0) \leq r\}.$$

One may use Ω_1^* in place of Ω_1 in the above definition, as in Montes et al. (2020a).

A distortion model is one example of the more general notion of neighbourhood models, found throughout the literature on IP and robustness. In general, a neighbourhood model on P_0 is simply a set of probabilities which contains P_0 . An advantage of distortion models is that they are characterised by a small number of parameters (three). Beyond the symmetric IoM $\mathcal{I}_1(e^{-\varepsilon}P_0, e^{\varepsilon}P_0)$ and the density ratio neigh-

bourhood $N_r(P_0) \cap \Omega_1$, other examples of distortion models can be found in [Montes et al. \(2020a,b\)](#); [Montes \(2023\)](#); [Miranda et al. \(2024\)](#); [Destercke et al. \(2022\)](#); [Pelessoni et al. \(2021\)](#); [Walley \(1991\)](#).

Invariance under marginalisation, invariance under updating and immunity to dilation are three desiderata for distortion models. For symmetric IoMs and density ratio neighbourhoods, these desiderata are studied in [Wasserman and Kadane \(1992\)](#); [Wasserman \(1992\)](#); [Seidenfeld and Wasserman \(1993\)](#). [de Campos et al. \(1994\)](#) showed that, under certain restrictions, $\mathcal{I}_1(e^{-\varepsilon}P_0, e^{\varepsilon}P_0)$ is 2-monotone. To the best of our knowledge, we are not aware of studies investigating the geometry of the symmetric IoM or the density ratio neighbourhood – for example, are they polytopes and, if so, how many extremal points do they have?

In Appendix C.3, we showed that the distorting functions D_{MULT} and d_{DR} are metrics – that is, they are positive definite, symmetric and satisfy the triangle inequality on the space Ω_1 of probability measures.

Two other desirable properties of a distorting function d_{DIST} are:

1. *Quasi-Convexity* ([Montes et al., 2020a](#); [Miranda et al., 2024](#)): Given a real- or complex-vector space V , a function $d : V \times V \rightarrow [0, \infty]$ is quasi-convex (in its second argument) if

$$d(v_1, \alpha v_2 + [1 - \alpha]v_3) \leq \max\{d(v_1, v_2), d(v_1, v_3)\}, \quad (\text{C.8})$$

for all $\alpha \in [0, 1]$ and all $v_1, v_2, v_3 \in V$.

2. *Continuity* ([Montes et al., 2020a](#)):² For $\mu \in \Omega$, let $\Omega_\nu = \{\nu \in \Omega \mid \nu \ll \mu\}$ be the set of σ -finite measures ν which are absolutely continuous with respect to μ . Given $S \subset \Omega$, a function $d : S \times S \rightarrow [0, \infty]$ is continuous (in its second argument) with respect to the supremum norm if, for all $\mu \in \Omega$, all $\nu_1, \nu_2 \in S \cap \Omega_\mu$ and all $\varepsilon > 0$, there exists $\delta > 0$ such that, for all $\nu_3 \in S \cap \Omega_\mu$,

$$\|\nu_2 - \nu_3\|_\infty^\mu < \delta \Rightarrow |d(\nu_1, \nu_2) - d(\nu_1, \nu_3)| < \varepsilon,$$

where $\|\cdot\|_\infty^\mu$ denotes the supremum norm.³ (The continuity is uniform if δ does not depend on ν_1 or ν_2 .)

²Note that our definition of continuity is strictly stronger than that given in [Montes et al. \(2020a\)](#) which allows δ to depend on ν_3 .

³Given a σ -finite measure space $(\mathcal{T}, \mathcal{F}, \mu)$, the supremum norm $\|\cdot\|_\infty^\mu$ on the set Ω_ν is defined as:

$$\|\nu\|_\infty^\mu = \text{ess sup} |f(t)|,$$

Remark C.4.3. The astute reader may consider (C.8) to be a strange definition of convexity. A more standard definition of convexity (in the second argument) would be the requirement:

$$d(v_1, \alpha v_2 + [1 - \alpha]v_3) \leq \alpha d(v_1, v_2) + (1 - \alpha)d(v_1, v_3), \quad (\text{C.9})$$

for all $\alpha \in [0, 1]$ and all $v_1, v_2, v_3 \in V$. This requirement is strictly stronger than quasi-convexity (equation (C.8)). (It is straightforward to prove (C.9) is stronger than (C.8) and we provide some examples later in this remark to prove strictness.) In fact, (C.9) is often too strong a requirement, for three reasons.

Firstly, if a distorting function d_{DIST} on Ω_1^* satisfies (C.8) and continuity, then the ball $\{P \in \Omega_1^* \mid d_{\text{DIST}}(P, P_0) \leq r\}$ centred at $P_0 \in \Omega_1^*$ is equal to the credal set induced by the lower envelope of this ball (Montes et al., 2020a, Proposition 3.1). Further, under (C.8) and continuity, there exist simple necessary and sufficient conditions for this lower envelope to be a probability interval (Montes et al., 2020a, Propositions 3.3, 3.4). Hence, (C.8) is a useful requirement for a distorting function d_{DIST} as it ensures the distortion model associated with d_{DIST} is well-behaved. It would be unnecessary to require the stricter condition (C.9) solely to ensure this nice behaviour.

Secondly, some of the common distorting functions found in the IP literature do not satisfy (C.9) but do satisfy (C.8). One example is the linear vacuous distorting function:

$$d_{\text{LV}} : \Omega_1^* \times \Omega_1^* \rightarrow [0, \infty),$$

$$(P, Q) \mapsto \max_{\emptyset \neq S \subset \mathcal{T}} \frac{Q(S) - P(S)}{Q(S)}.$$

Montes et al. (2020a, Proposition 5.1) show that d_{LV} satisfies (C.8). However, the following counterexample demonstrates that d_{LV} does not satisfy (C.9): Let $\alpha = 0.5$ and $\mathcal{T} = \{1, 2\}$. Let P_1 and P_2 be the

where the essential supremum is with respect to μ and f is a μ -density of ν .

uniform probability on $(\mathcal{T}, 2^{\mathcal{T}})$ and define $P_3 \in \Omega_1^*$ by $P_3(\{1\}) = 0.7$. Then

$$d_{\text{LV}}(P_1, \alpha P_2 + (1 - \alpha)P_3) = \frac{1}{6} > \frac{1}{7} = \alpha d_{\text{LV}}(P_1, P_2) + (1 - \alpha)d_{\text{LV}}(P_1, P_3).$$

The multiplicative distance D_{MULT} also does not satisfy (C.9) (although we show below that it satisfies (C.8)), even when restricting to Ω_1^* and to \mathcal{T} with finite cardinality. To see this, let $\alpha = 0.5$ and let P_1 be uniform on $\mathcal{T} = \{1, \dots, 10\}$; define P_2 by $P_2(\{1\}) = 1/5$ and $P_2(\{t\}) = 8/90$ for all $t \neq 1$; and define P_3 by $P_3(\{1\}) = 3/10$ and $P_3(\{t\}) = 7/90$ for all $t \neq 1$. Then

$$D_{\text{MULT}}(P_1, \alpha P_2 + (1 - \alpha)P_3) = \ln 2.5 > \frac{1}{2}(\ln 2 + \ln 3) = \alpha D_{\text{MULT}}(P_1, P_2) + (1 - \alpha)D_{\text{MULT}}(P_1, P_3).$$

Thirdly and finally, (C.8) aligns with the notion of convexity which arises independently in the DP literature [Kifer et al. \(2022\)](#): Under a mild assumption, a DP flavor [Bailie et al. \(2025b\)](#) with distorting function d_{DIST} is convex in the sense of [Kifer et al. \(2022\)](#) if and only if d_{DIST} is quasi-convex in the sense of (C.8). Because most of the commonly used DP flavors are convex, it follows that most of the distorting functions used in DP are quasi-convex.

Proof. that D_{MULT} satisfies quasi-convexity: Let μ, ν and τ be measures on $(\mathcal{T}, \mathcal{F})$. For $E \in \mathcal{F}$, we have

$$\begin{aligned} \ln \frac{\mu(E)}{\alpha \nu(E) + (1 - \alpha)\tau(E)} &\leq \ln \frac{\mu(E)}{\min\{\nu(E), \tau(E)\}} \\ &\leq \sup_{S \in \mathcal{F}} \left\{ \max \left\{ \ln \frac{\mu(S)}{\nu(S)}, \ln \frac{\mu(S)}{\tau(S)} \right\} \right\} \\ &\leq \max\{D_{\text{MULT}}(\mu, \nu), D_{\text{MULT}}(\mu, \tau)\}. \end{aligned}$$

Similarly,

$$\ln \frac{\alpha \nu(E) + (1 - \alpha)\tau(E)}{\mu(E)} \leq \ln \frac{\max\{\nu(E), \tau(E)\}}{\mu(E)}$$

$$\begin{aligned} &\leq \sup_{S \in \mathcal{F}} \left\{ \max \left\{ \ln \frac{\nu(S)}{\mu(S)}, \ln \frac{\tau(S)}{\mu(S)} \right\} \right\} \\ &\leq \max \{ D_{\text{MULT}}(\mu, \nu), D_{\text{MULT}}(\mu, \tau) \}. \end{aligned}$$

□

Lemma C.4.4. *Let $d : S \times S \rightarrow [0, \infty]$ be a metric (where $S \subset \Omega$). The following is a sufficient condition for d to be continuous: For all $\mu \in \Omega$, all $\nu_1 \in S \cap \Omega_\mu$ and all $\varepsilon > 0$, there exists $\delta > 0$ such that, for all $\nu_2 \in S \cap \Omega_\mu$,*

$$\|\nu_1 - \nu_2\|_\infty^\mu < \delta \Rightarrow d(\nu_1, \nu_2) < \varepsilon.$$

Proof. This result follows by the triangle inequality and symmetry: $|d(\nu_1, \nu_2) - d(\nu_1, \nu_3)| \leq d(\nu_2, \nu_3)$.

□

The following proposition proves that D_{MULT} is continuous, in the setting considered in [Montes et al. \(2020a\)](#).

Proposition 9. *Suppose that \mathcal{T} has finite cardinality. Let $\mathcal{F} = 2^{\mathcal{T}}$ and $\Omega^* = \{\nu \in \Omega \mid \nu(\{t\}) > 0 \forall t \in \mathcal{T}\}$. Then D_{MULT} is continuous when restricted to the domain $\Omega^* \times \Omega^*$.*

Proposition 9 implies that the interval of measures $\mathcal{I}(L, U) \cap \Omega^*$ is closed with respect to the topology induced by the supremum norm, as long as \mathcal{T} has finite cardinality.

Proof. Lemma C.4.4 describes a sufficient condition for continuity. We will prove that this sufficient condition holds under the assumption that \mathcal{T} has finite cardinality and under the restriction to σ -finite measures ν with $\nu(\{t\}) > 0$.

Without loss of generality, we assume that the dominating measure μ is the counting measure on \mathcal{T} .

Then $\Omega^* \subset \Omega_\mu$. Let $\nu_1 \in \Omega^*$ and $\varepsilon > 0$. Let f be the μ -density of ν_1 . Choose some δ which satisfies

$$0 < \delta < \min_{t \in \mathcal{T}} f(t)(1 - e^{-\varepsilon}).$$

(Such a δ exists because $f(t) > 0$ for all $t \in \mathcal{T}$ and because \mathcal{T} has finite cardinality.) Fix some $\nu_2 \in \Omega^*$

with $\|\nu_1 - \nu_2\|_\infty^\mu < \delta$. Then, for any $S \neq \emptyset$,

$$\frac{\nu_2(S)}{\nu_1(S)} \leq \frac{\nu_1(S) + \delta}{\nu_1(S)} \leq \max_{t \in \mathcal{T}} \frac{f(t) + \delta}{f(t)} \leq \max_{t \in \mathcal{T}} \frac{f(t)}{f(t) - \delta} < e^\varepsilon.$$

(The last inequality follows because $\delta < f(t) - f(t)e^{-\varepsilon}$ for all t .) Similarly,

$$\frac{\nu_1(S)}{\nu_2(S)} \leq \frac{\nu_1(S)}{\nu_1(S) - \delta} \leq \max_{t \in \mathcal{T}} \frac{f(t)}{f(t) - \delta} < e^\varepsilon.$$

Hence $D_{\text{MULT}}(\nu_1, \nu_2) < \varepsilon$. □

If we remove either of the two restriction in Proposition 9 (to finite \mathcal{T} and to $\mu, \nu \in \Omega^*$), then D_{MULT} is no longer continuous. We will demonstrate that both of these restrictions are necessary with two counterexamples: In the first case, suppose $\mathcal{T} = \{1, 2, \dots\}$ with $\mathcal{F} = 2^\mathcal{T}$ and let μ be the counting measure. Define $\nu_1(\{t\}) = t^{-2}$. Let $\varepsilon = \ln 2$ and fix $0 < \delta < 1$. Define

$$\nu_2(\{t\}) = \begin{cases} 1 - \delta/4 & \text{if } t = 1, \\ t^{-2} + \delta/4 & \text{if } t = t_0, \\ t^{-2} & \text{otherwise,} \end{cases}$$

where $t_0 = \left\lceil \sqrt{4\delta^{-1}} + 1 \right\rceil$. Then $\|v_1 - v_2\|_\infty^\mu < \delta$ but

$$D_{\text{MULT}}(v_1, v_2) \geq \ln \frac{t_0^{-2} + \delta/4}{t_0^{-2}} > \ln 2.$$

In the second case, let \mathcal{T} be finite, $\mathcal{F} = 2^{\mathcal{T}}$ and μ be the counting measure. For some $t_0 \in \mathcal{T}$, define $v_1(\{t_0\}) = \delta/2$ and $v_2(\{t_0\}) = 0$, and suppose that $v_1(\{t\}) = v_2(\{t\})$ for all $t \neq t_0$. Then $\|v_1 - v_2\|_\infty^\mu < \delta$ but $D_{\text{MULT}}(v_1, v_2) = \infty$.

These two counterexamples can easily be modified so that v_1 and v_2 are probability measures. Hence D_{MULT} is also not continuous on the space of probability measures, except when \mathcal{T} is finite and we restrict to probabilities with support \mathcal{T} .

C.5 SUPPLEMENTARY RESULTS

C.5.1 AN EQUIVALENT DEFINITION OF THE MULTIPLICATIVE DISTANCE D_{MULT}

Lemma C.5.1. *Suppose that $\mu \in \Omega$. Let f be a μ -density of ν . Then ν is σ -finite if and only if f is finite μ -almost everywhere.*

Proof. “ \Rightarrow ” by the contrapositive: (This direction does not require that μ is σ -finite.) Suppose that $\mu(f = \infty) > 0$. Let $\{E_n : n \in \mathbb{N}\} \subset \mathcal{F}$ be any countable partition of \mathcal{T} . We will show that necessarily there exists some E_n with $\nu(E_n) = \infty$. Since

$$0 < \mu(f = \infty) = \sum_{n=1}^{\infty} \mu(E_n \cap \{f = \infty\}),$$

there exists some E_n with $\mu(E_n \cap \{f = \infty\}) > 0$. Then

$$\nu(E_n) \geq \nu(E_n \cap \{f = \infty\}) = \int_{E_n \cap \{f = \infty\}} f d\mu = \infty.$$

“ \Leftarrow ”: (This direction requires that μ is σ -finite – the case that $\mu = \nu$ with $f = 1$ serves as a counterexample.) Let $A = \{f = \infty\} \in \mathcal{F}$. Then $\nu(A) = \mu(A) = 0$ because ν is absolutely continuous with respect to μ by assumption. Define $E_n = \{t \in \mathcal{T} \setminus A : n - 1 \leq f(t) < n\}$ for all $n \in \mathbb{N}$. Let $\{S_m : m \in \mathbb{N}\}$ be a partition of \mathcal{T} such that $\mu(S_m) < \infty$ for all $m \in \mathbb{N}$. Then $\nu(E_n \cap S_m) < n\mu(S_m) < \infty$. Hence $\{A, E_n \cap S_m : n, m \in \mathbb{N}\}$ is a countable partition of \mathcal{T} with each component having finite ν -measure. \square

The following proposition gives an alternative definition of the multiplicative distance D_{MULT} .

Proposition 10. *Let $\mu, \nu \in \Omega$. Then*

$$D_{\text{MULT}}(\mu, \nu) = \begin{cases} 0 & \text{if } \mu = \nu = 0, \\ \text{ess sup}_{t \in \mathcal{T}^\circ} \left| \ln \frac{f(t)}{g(t)} \right| & \text{else if } \mu, \nu \text{ are mutually absolutely continuous,} \\ \infty & \text{otherwise,} \end{cases}$$

where f and g are any densities of μ and ν respectively, with respect to any common dominating measure

$\tau \in \Omega$; $\mathcal{T}^\circ = \{t \in \mathcal{T} \mid 0 < f(t), g(t) < \infty\}$; and the essential supremum is with respect to τ .

To prove the above proposition, we need two lemmata.

Lemma C.5.2. *Suppose that $\mu, \nu \in \Omega$ are non-zero and mutually absolutely continuous and $\tau \in \Omega$ is a dominating measure. Let f and g be τ -densities of μ and ν respectively.*

Then

$$\sup_{S \in \mathcal{F}} \ln \frac{\mu(S)}{\nu(S)} = \text{ess sup}_{t \in \mathcal{T}^\circ} \ln \frac{f(t)}{g(t)},$$

where the essential supremum is over τ .

Proof. We first need that (A) $f, g < \infty$ holds τ -almost everywhere and that (B) $f, g > 0$ holds μ - and ν -almost everywhere. (A) is a direct result of Lemma C.5.1. To prove (B), observe that

$$\mu(f = 0) = \int_{\{f=0\}} f d\tau = 0,$$

and then $\nu(f = 0) = 0$ follows by absolute continuity. That $g > 0$ holds μ - and ν -almost everywhere has a similar proof.

For $a \in \mathbb{R}$, define $E_a = \{t \in \mathcal{T}^0 \mid f(t) > \exp(a)g(t)\}$. We need to prove the following result holds for all $a \in \mathbb{R}$: There exists $S \in \mathcal{F}$ such that $\mu(S) > \exp(a)\nu(S)$ if and only if $\tau(E_a) > 0$. Denote this result by (*).

Suppose that there exists $S \in \mathcal{F}$ such that $\mu(S) > \exp(a)\nu(S)$. By (A) and (B), this implies

$$\mu(S \cap \mathcal{T}^0) > \exp(a)\nu(S \cap \mathcal{T}^0),$$

and hence

$$\int_{S \cap \mathcal{T}^0} (f - e^a g) d\tau > 0.$$

Thus, there exists some $E \subset S \cap \mathcal{T}^0$ such that $\tau(E) > 0$ and $f(t) - \exp(a)g(t) > 0$ for all $t \in E$. Hence $\tau(E_a) \geq \tau(E) > 0$.

In the other direction, suppose $\tau(E_a) > 0$. Then

$$\int_{E_a} (f - e^a g) d\tau > 0,$$

which implies $\mu(E_a) > \exp(a)\nu(E_a)$. This proves (*).

Finally, we have

$$\begin{aligned}
\sup_{S \in \mathcal{F}} \ln \frac{\mu(S)}{\nu(S)} &= \sup\{a \in \mathbb{R} \mid \text{there exists } S \in \mathcal{F} \text{ s.t. } \mu(S) > \exp(a)\nu(S)\} \\
&= \sup\{a \in \mathbb{R} \mid \tau(E_a) > 0\} \\
&= \text{ess sup}_{t \in T^o} \ln \frac{f(t)}{g(t)},
\end{aligned}$$

where the first line follows by continuity of $\exp(\cdot)$; the second by $(*)$; and the third by the definition of the essential supremum. \square

Lemma C.5.3. *Let $\mu, \nu \in \Omega$ be non-zero and not mutually absolutely continuous. Then*

$$\sup_{S \in \mathcal{F}} \ln \frac{\mu(S)}{\nu(S)} = \infty \quad \text{or} \quad \sup_{S \in \mathcal{F}} \ln \frac{\nu(S)}{\mu(S)} = \infty.$$

Proof. Without loss of generality, there exists some $E \in \mathcal{F}$ such that $\mu(E) > 0$ but $\nu(E) = 0$. Then

$$\sup_{S \in \mathcal{F}} \ln \frac{\mu(S)}{\nu(S)} \geq \ln \frac{\mu(E)}{\nu(E)} = \infty.$$

\square

Proof. of Proposition 10: Suppose that $\mu = \nu = 0$. Then, for all $S \in \mathcal{F}$, we have that $\ln \frac{\mu(S)}{\nu(S)} = \ln \frac{0}{0} = 0$

where we define $0/0 = 1$ as in the definition of the multiplicative distance (Definition 5.2.1). Hence

$$D_{\text{MULT}}(\mu, \nu) = 0.$$

Now suppose that μ and ν are both non-zero and not mutually absolutely continuous. By Lemma C.5.3,

$$D_{\text{MULT}}(\mu, \nu) = \infty.$$

Finally, suppose that μ and ν are both non-zero and mutually absolutely continuous. Fix a dominating

measure τ and τ -densities f and g . By symmetry, it suffices to show that

$$\operatorname{ess\,sup}_{t \in \mathcal{T}^o} \ln \frac{f(t)}{g(t)} = \sup_{S \in \mathcal{F}} \ln \frac{\mu(S)}{\nu(S)}.$$

This equation is given by Lemma C.5.2. □

C.5.2 THE MULTIPLICATIVE DISTANCE D_{MULT} IS COMPOSABLE

Proposition 11. *Let $(\mathcal{T}_1, \mathcal{F}_1)$ and $(\mathcal{T}_2, \mathcal{F}_2)$ be measurable spaces and let $\Omega^{(i)}$ be the collection of σ -finite measures on $(\mathcal{T}_i, \mathcal{F}_i)$. Then, for all $\mu_1, \nu_1 \in \Omega^{(1)}$ and all $\mu_2, \nu_2 \in \Omega^{(2)}$,*

$$D_{\text{MULT}}(\mu_1 \times \mu_2, \nu_1 \times \nu_2) \leq D_{\text{MULT}}(\mu_1, \nu_1) + D_{\text{MULT}}(\mu_2, \nu_2), \quad (\text{C.10})$$

where $\mu_1 \times \mu_2$ and $\nu_1 \times \nu_2$ are product measures.

Proof. We will use the definition of the multiplicative distance D_{MULT} given in Proposition 10.

Suppose that μ_1 and ν_1 are zero. Then $\mu_1 \times \mu_2$ and $\nu_1 \times \nu_2$ are also zero. Hence (C.10) simplifies to $0 \leq D_{\text{MULT}}(\mu_2, \nu_2)$, which always holds because D_{MULT} is a metric.

Suppose that μ_1 and ν_2 are non-zero but not absolutely continuous. Then the RHS of (C.10) is infinite, and so (C.10) holds vacuously.

Suppose that μ_1 and ν_1 are non-zero and mutually absolutely continuous. Then there exists a common dominating measure $\tau \in \Omega_1$. Let f and g denote τ -densities of μ_1 and ν_1 respectively. Then, Lemma C.5.1 and Proposition 10 together imply that

$$f(t) \leq g(t) \exp[D_{\text{MULT}}(\mu_1, \nu_1)], \quad (\text{C.11})$$

for τ -a.e. $t \in \mathcal{T}_1$. Thus, for $S \in \mathcal{F}_1 \otimes \mathcal{F}_2$,

$$\begin{aligned}
(\mu_1 \times \mu_2)(S) &= \int_{\mathcal{T}_1} \mu_2(S') d\mu_1(t) \\
&= \int_{\mathcal{T}_1} \mu_2(S') f(t) d\tau(t) \\
&\leq \int_{\mathcal{T}_1} \mu_2(S') g(t) \exp[D_{\text{MULT}}(\mu_1, \nu_1)] d\tau(t) \\
&\leq \int_{\mathcal{T}_1} \exp[D_{\text{MULT}}(\mu_2, \nu_2)] \nu_2(S') g(t) \exp[D_{\text{MULT}}(\mu_1, \nu_1)] d\tau(t) \\
&= \exp[D_{\text{MULT}}(\mu_1, \nu_1) + D_{\text{MULT}}(\mu_2, \nu_2)] \int_{\mathcal{T}_1} \nu_2(S') g(t) d\tau(t) \\
&= \exp[D_{\text{MULT}}(\mu_1, \nu_1) + D_{\text{MULT}}(\mu_2, \nu_2)] (\nu_1 \times \nu_2)(S),
\end{aligned}$$

where the third line follows by (C.11) and the fourth by the definition of D_{MULT} (Definition 5.2.1). Hence

$$\ln \frac{(\mu_1 \times \mu_2)(S)}{(\nu_1 \times \nu_2)(S)} \leq D_{\text{MULT}}(\mu_1, \nu_1) + D_{\text{MULT}}(\mu_2, \nu_2),$$

for all $S \in \mathcal{F}_1 \otimes \mathcal{F}_2$. The proof of the bound

$$\ln \frac{(\nu_1 \times \nu_2)(S)}{(\mu_1 \times \mu_2)(S)} \leq D_{\text{MULT}}(\mu_1, \nu_1) + D_{\text{MULT}}(\mu_2, \nu_2),$$

follows analogously. □

C.5.3 AN EQUIVALENT DEFINITION OF THE DENSITY RATIO METRIC d_{DR}

The following proposition gives an alternative definition of the density ratio metric d_{DR} .

Proposition 12. *For any non-zero $\mu, \nu \in \Omega$,*

$$d_{\text{DR}}(\mu, \nu) = \sup_{S \in \mathcal{F}} \ln \frac{\mu(S)}{\nu(S)} + \sup_{S \in \mathcal{F}} \ln \frac{\nu(S)}{\mu(S)}, \quad (\text{C.12})$$

where $0/0 = \infty/\infty = 1$. Also, $d_{\text{DR}}(\mu, \nu) = \infty$ if exactly one of μ or ν is zero; and $d_{\text{DR}}(\mu, \nu) = 0$ if μ and ν are both zero.

Compare the two definitions of the multiplicative distance D_{MULT} – Definition 5.2.1 and Proposition 10 – with the two definitions of the density ratio metric d_{DR} – Definition 5.8.3 and Proposition 12. Each of these two distances have definitions in terms of densities and in terms of measures.

Proof. of Proposition 12: Firstly, we must verify that the RHS of (C.12) is well defined – i.e. that the RHS cannot take on the form of $\infty - \infty$. We need to prove that (*) both $\sup_{S \in \mathcal{F}} \ln \frac{\mu(S)}{\nu(S)}$ and $\sup_{S \in \mathcal{F}} \ln \frac{\nu(S)}{\mu(S)}$ are bounded away from negative infinity.

Let E_1, E_2, \dots be a partition of \mathcal{T} such that $\mu(E_n) < \infty$. Since $\nu(\mathcal{T}) > 0$ there must exist some E_n with $\nu(E_n) > 0$. Then

$$\sup_{S \in \mathcal{F}} \ln \frac{\nu(S)}{\mu(S)} \geq \ln \frac{\nu(E_n)}{\mu(E_n)} > -\infty.$$

The proof of $\sup_{S \in \mathcal{F}} \ln \frac{\mu(S)}{\nu(S)} > -\infty$ is analogous.

Suppose that $\mu = \nu = 0$. Then

$$\ln \frac{\mu(S)}{\nu(S)} = \ln \frac{\nu(S)}{\mu(S)} = 0,$$

for all $S \in \mathcal{F}$. Hence (C.12) holds.

Suppose that μ and ν are non-zero and not mutually absolutely continuous. Then by Lemma C.5.3 and (*),

$$\sup_{S \in \mathcal{F}} \ln \frac{\mu(S)}{\nu(S)} + \sup_{S \in \mathcal{F}} \ln \frac{\nu(S)}{\mu(S)} = \infty.$$

Hence (C.12) holds once again.

Finally, suppose that μ and ν are non-zero and mutually absolutely continuous. Let τ be a dominating

measure of μ and ν and suppose that f and g are τ -densities of μ and ν respectively. By Lemma C.5.2,

$$\sup_{S \in \mathcal{F}} \ln \frac{\mu(S)}{\nu(S)} + \sup_{S \in \mathcal{F}} \ln \frac{\nu(S)}{\mu(S)} = \operatorname{ess\,sup}_{t \in \mathcal{T}^\circ} \ln \frac{f(t)}{g(t)} + \operatorname{ess\,sup}_{t' \in \mathcal{T}^\circ} \ln \frac{g(t')}{f(t')} = d_{\text{DR}}(\mu, \nu),$$

where the essential supremum is over τ . □

Corollary C.5.4. *For finite $\mu, \nu \in \Omega$,*

$$d_{\text{DR}}(\mu, \nu) = \begin{cases} 0 & \text{if } \mu = \nu = 0, \\ \sup_{S, S' \in \mathcal{F}^*} \ln \frac{\mu(S)\nu(S')}{\mu(S')\nu(S)} & \text{else if } \mu, \nu \text{ are mutually absolutely continuous,} \\ \infty & \text{otherwise,} \end{cases}$$

where $\mathcal{F}^* = \{S \in \mathcal{F} : \mu(S) > 0\}$.

Proof. Proving the result when $\mu = \nu = 0$ or when μ, ν not mutually absolutely continuous is straightforward. We may thus assume that μ and ν are non-zero and mutually absolutely continuous. Then $\mathcal{F}^* = \{S \in \mathcal{F} : \nu(S) > 0\}$ and moreover

$$\ln \frac{\mu(S)}{\nu(S)} = 0,$$

for all $S \notin \mathcal{F}^*$. Hence

$$d_{\text{DR}}(\mu, \nu) = \sup_{S \in \mathcal{F}^*} \ln \frac{\mu(S)}{\nu(S)} + \sup_{S \in \mathcal{F}^*} \ln \frac{\nu(S)}{\mu(S)}.$$

The result then follows because $0 < \mu(S), \nu(S) < \infty$ for all $S \in \mathcal{F}^*$ implies we may combine the log-terms without introducing an undefined operation. That is,

$$\sup_{S \in \mathcal{F}^*} \ln \frac{\mu(S)}{\nu(S)} + \sup_{S \in \mathcal{F}^*} \ln \frac{\nu(S)}{\mu(S)} = \sup_{S, S' \in \mathcal{F}^*} \ln \frac{\mu(S)\nu(S')}{\mu(S')\nu(S)},$$

because $0 < \mu(S), \nu(S) < \infty$ for all $S \in \mathcal{F}^*$. □

Corollary C.5.5. *For finite $\mu, \nu \in \Omega$,*

$$d_{\text{DR}}(\mu, \nu) = -\ln[1 - d_{\text{COR}}(\mu, \nu)],$$

where d_{COR} is the constant odds metric (defined in Remark 5.8.7).

Proof. The cases where μ and ν are zero or not mutually absolutely continuous are straightforward to verify. When μ and ν are non-zero and mutually absolutely continuous,

$$\frac{1}{-d_{\text{COR}}(\mu, \nu) + 1} = \exp(d_{\text{DR}}(\mu, \nu)),$$

by Corollary C.5.4. □

C.5.4 THE MULTIPLICATIVE DISTANCE D_{MULT} AND THE DENSITY RATIO METRIC d_{DR} ARE STRONGLY EQUIVALENT FOR PROBABILITIES

The following proposition proves that the multiplicative distance D_{MULT} and the density ratio metric d_{DR} are strongly equivalent (Definition C.3.2) – but not equal – on the set of probability measures. This proposition also shows that D_{MULT} and d_{DR} are not strongly equivalent on the set Ω of σ -finite measures.

See also Theorem 1 of [Wasserman and Kadane \(1992\)](#), which describes the relationships between intervals of measures (or density bounded classes) and density ratio neighbourhoods.

Proposition 13. *Let $P, Q \in \Omega_1$ be probability measures on $(\mathcal{T}, \mathcal{F})$. Then*

$$D_{\text{MULT}}(P, Q) \leq d_{\text{DR}}(P, Q) \leq 2D_{\text{MULT}}(P, Q). \quad (\text{C.13})$$

Moreover, for each of the two inequalities in (C.13):

- *there exist $P, Q \in \Omega_1$ such that the inequality is strict; and*

- there exist probabilities $P, Q \in \Omega_1$ (which can even be mutually absolutely continuous and not equal) such that the inequality is an equality.

Finally, when the probabilities P and Q are replaced by measures $\mu, \nu \in \Omega$, the first inequality of (C.13) does not hold (even up to a non-zero multiplicative constant), but the second does.

Proof. Let P and Q be probability measures on $(\mathcal{T}, \mathcal{F})$. Proving (C.13) when $P = Q$ is trivial by the metric properties of D_{MULT} and d_{DR} . Hence, we may assume that $P \neq Q$. Then there exists $S \in \mathcal{F}$ such that $P(S) > Q(S)$ and consequently,

$$\sup_{S \in \mathcal{F}} \ln \frac{P(S)}{Q(S)} > 0. \quad (\text{C.14})$$

Symmetrically,

$$\sup_{S \in \mathcal{F}} \ln \frac{Q(S)}{P(S)} > 0. \quad (\text{C.15})$$

Thus,

$$\begin{aligned} D_{\text{MULT}}(P, Q) &= \max \left(\sup_{S \in \mathcal{F}} \ln \frac{P(S)}{Q(S)}, \sup_{S \in \mathcal{F}} \ln \frac{Q(S)}{P(S)} \right) \\ &\leq \sup_{S \in \mathcal{F}} \ln \frac{P(S)}{Q(S)} + \sup_{S \in \mathcal{F}} \ln \frac{Q(S)}{P(S)} \\ &= d_{\text{DR}}(P, Q), \end{aligned} \quad (\text{C.16})$$

where the second line follows by (C.14) and (C.15), and the third by Proposition 12.

Also,

$$d_{\text{DR}}(P, Q) \leq \sup_{S \in \mathcal{F}} \left| \ln \frac{P(S)}{Q(S)} \right| + \sup_{S \in \mathcal{F}} \left| \ln \frac{Q(S)}{P(S)} \right| = 2D_{\text{MULT}}(P, Q),$$

by Proposition 12.

Note that the first inequality, $D_{\text{MULT}}(P, Q) \leq d_{\text{DR}}(P, Q)$, does not hold when P, Q are replaced by

(non-probability) measures. For example, for μ and ν defined in (C.7), $d_{\text{DR}}(\mu, \nu) = 0$ while $D_{\text{MULT}}(\mu, \nu) = \ln 2$. (It is easy to modify this example to construct μ and ν such that $0 < d_{\text{DR}}(\mu, \nu) < D_{\text{MULT}}(\mu, \nu)$. Simply change ν on some $E \neq \mathcal{T}$ so that $0 < \mu(E) < \nu(E) < 2\mu(E)$ while still maintaining $\nu(E') = 2\mu(E') > 0$ on some other E' .) However, the second inequality, $d_{\text{DR}}(\mu, \nu) \leq D_{\text{MULT}}(\mu, \nu)$, does hold for any $\mu, \nu \in \Omega$, by the same proof as given above.

The inequalities in (C.13) are tight, even when P and Q are mutually absolutely continuous and $P \neq Q$:

To see that the first inequality is tight, define the probability P on $[-1, 1]$ by the density

$$f(x) = \begin{cases} \frac{2}{3} & \text{if } x \in [-1, 0], \\ \frac{2}{3}(1-x) & \text{if } x \in (0, 1], \end{cases}$$

and Q by

$$g(x) = \begin{cases} \frac{2}{3}(1+x) & \text{if } x \in [-1, 0], \\ \frac{2}{3} & \text{if } x \in (0, 1], \end{cases}.$$

Then $D_{\text{MULT}}(P, Q) = d_{\text{DR}}(P, Q) = \infty$. However, if $0 < D_{\text{MULT}}(P, Q) < \infty$, then the inequality in (C.16) will be strict and thus $D_{\text{MULT}}(P, Q) < d_{\text{DR}}(P, Q)$.

To see that the second inequality in (C.13) is tight, define P, Q on $\{-1, 1\}$ by $P(-1) = Q(1) = 1/3$ and $P(1) = Q(-1) = 2/3$. Then $d_{\text{DR}}(P, Q) = 2 \ln 2 = 2D_{\text{MULT}}(P, Q)$. However, it is also possible that $d_{\text{DR}}(P, Q) < 2D_{\text{MULT}}(P, Q)$: Define P, Q on $\{1, 2, \dots, 9\}$ by

- $P(1) = 9/10$;
- $P(x) = 1/80$ for $x = 2, 3, \dots, 9$;
- $Q(1) = 1/5$; and

- $Q(x) = 1/10$; for $x = 1, 2, \dots, 10$.

Then $d_{\text{DR}}(P, Q) = \ln 36 < 2 \ln 8 = 2D_{\text{MULT}}(P, Q)$. □

C.5.5 AN EQUIVALENCE BETWEEN INTERVALS OF MEASURES AND DENSITY BOUNDED CLASSES

The following proposition provides an equivalence between intervals of measures and density bounded classes. Recall that $\mu \ll \nu$ denotes that μ is absolutely continuous with respect to ν .

Proposition 14. *Given $L, U, \nu \in \Omega$ with $L \leq U$ and $U \ll \nu$, there exists some ν -densities $l \leq u$ such that*

- (a) *Every $\mu \in \mathcal{I}(L, U)$ has a ν -density f which is in the density bounded class $\mathcal{I}(l, u)$; and*
- (b) *The measure $\mu(S) = \int_S f d\nu$ given by any density $f \in \mathcal{I}(l, u)$ is in the interval of measure $\mathcal{I}(L, U)$.*

Moreover, if $\nu = U$ then u is the constant function: $u(t) = 1$ for all $t \in \mathcal{T}$; and l is a ν -density of L , which exists because L is absolutely continuous with respect to U .

Additionally, if $L = a\tau$ and $U = b\tau$ for some $\tau \in \Omega$ with $\tau \ll \nu$ and constants $0 < a \leq 1 \leq b < \infty$, then $l = ag$ and $u = bg$, where g is a ν -density of τ .

In the other direction, given some $\nu \in \Omega$ and ν -densities $l \leq u$ (which are finite ν -almost everywhere), define $L, U \in \Omega$ by

$$L(S) = \int_S l d\nu, \tag{C.17}$$

$$U(S) = \int_S u d\nu, \tag{C.18}$$

for all $S \in \mathcal{F}$. Then $L \leq U$ and the properties (a) and (b) above hold.

Note that the condition that l and u are finite ν -almost everywhere is necessary and sufficient for L, U to be in Ω (see Lemma C.5.1).

The first half of this proposition (before “In the other direction...”) shows that an interval of measure $\mathcal{I}(L, U)$ can be considered as a density bounded class $\mathcal{I}(l, u)$, and the second half shows the converse – that a density bounded class $\mathcal{I}(l, u)$ can be considered as an interval of measure $\mathcal{I}(L, U)$.

Proof. Let $L, U \in \Omega$ and suppose $L \leq U$. We first consider the case that $\nu = U$. The fact that $L \leq U$ implies that $L \ll U$: if $U(S) = 0$ then $L(S) = 0$, for all $S \in \mathcal{F}$. Thus, by the Radon-Nikodym theorem, L has a density l with respect to U . Define the density u by $u(t) = 1$ for all $t \in \mathcal{T}$. It is straightforward to verify that u is a density of U with respect to U (i.e. u is a U -density of U).

Suppose, for contradiction, that there exists $E \in \mathcal{F}$ with $U(E) > 0$ and $l(t) > u(t)$ for all $t \in E$. Then

$$0 < U(E) = \int_E u dU < \int_E l dU = L(E),$$

contradicting the assumption $L \leq U$. Hence there exists \tilde{l} such that $\tilde{l}(t) \leq u(t)$ for all $t \in \mathcal{T}$ and $\tilde{l}(t) = l(t)$, for U -almost all $t \in \mathcal{T}$. This implies \tilde{l} is also a U -density of L because

$$\int_S \tilde{l} dU = \int_S l dU = L(S),$$

for all $S \in \mathcal{F}$. From herein, replace l by its modification \tilde{l} . This proves $l \leq u$.

Now we prove (a). Let $\mu \in \mathcal{I}(L, U)$. Since $\mu \leq U$, we know that μ is absolutely continuous with respect to U and hence has a U -density f by the Radon-Nikodym theorem. We know that $f(t) \leq u(t)$ for U -almost all $t \in \mathcal{T}$. Otherwise there exists $E \in \mathcal{F}$ with $U(E) > 0$ and $f(t) > u(t)$ for all $t \in E$, which would imply

$$0 < U(E) = \int_E u dU < \int_E f dU = \mu(E),$$

contradicting the assumption $\mu \leq U$. Thus, $f \leq u$, U -almost everywhere. This implies there exists \tilde{f} which

differs from f on a U -null set such that $\tilde{f}(t) \leq u(t)$ for all $t \in \mathcal{T}$. Since f and \tilde{f} differ only on a U -null set, \tilde{f} is also a U -density of μ . By exactly the same reasoning as above, we can show that $\tilde{f} \geq l$, U -almost everywhere and hence there exists a modification \check{f} of \tilde{f} such that

$$l(t) \leq \check{f}(t) \leq u(t),$$

and $\check{f} \neq \tilde{f}$ only on a U -null set. Thus, \check{f} is a U -density of μ such that $\check{f} \in \mathcal{I}(l, u)$. This proves (a).

Next we prove (b). Fix some density $f \in \mathcal{I}(l, u)$ and define $\mu \in \Omega$ by

$$\mu(S) = \int_S f dU.$$

Then

$$L(S) = \int_S l dU \leq \int_S f dU = \mu(S),$$

and

$$\mu(S) = \int_S f dU \leq \int_S u dU = U(S).$$

This proves $\mu \in \mathcal{I}(L, U)$.

Now we consider the general case of the first half of the proposition. Let $L, U, \nu \in \Omega$ with $L \leq U$ and $U \ll \nu$. Then $L \ll \nu$ as well, and let l and u be ν -densities of L and U respectively. By the same reasoning as above, we can replace l and u with their modifications \tilde{l}, \tilde{u} . Hence we may assume that $l(t) \leq u(t)$ for all $t \in \mathcal{T}$. Now take $\mu \in \mathcal{I}(L, U)$. Since $\mu \ll U \ll \nu$, the Radon-Nikodym theorem states that μ has a ν -density f . Analogous to above, we can modify f on a ν -null set to produce a ν -density \check{f} of μ such that $\check{f} \in \mathcal{I}(l, u)$. This proves (a). The proof of (b) is analogous to the case when $\nu = U$ with the integrals $\int \cdot dU$ replaced by $\int \cdot d\nu$.

Finally, we consider when $L = a\tau$ and $U = b\tau$ for some $\tau \in \Omega$ with $\tau \ll \nu$ and constants $0 < a \leq 1 \leq b < \infty$. Define $l = ag$ and $u = bg$, where g is some ν -density of τ . Then $l \leq u$ and the proof of properties (a) and (b) are as before.

We now prove the second half of the proposition (that which follows “In the other direction...”). Let $\nu \in \Omega$ and suppose l and u are ν -densities. Define $L, U \in \Omega$ according to equations (C.17) and (C.18). Then $L \leq U$ since

$$L(S) = \int_S l d\nu \leq \int_S u d\nu = U(S).$$

We will show that property (a) is satisfied. Let $\mu \in \mathcal{I}(L, U)$. By the same reasoning as in the proof of the first half of the proposition, modify the ν -density f of μ on a ν -null set to produce \check{f} satisfying $l(t) \leq \check{f}(t) \leq u(t)$. This implies \check{f} is a ν -density of μ and that $\check{f} \in \mathcal{I}(l, u)$.

Finally, property (b) also follows by the same reasoning as in the proof of it in the first half of the proposition. □

C.5.6 CHARACTERISING MECHANISMS WITH ZERO PRIVACY LOSS

The following proposition formalises the relationship between complete privacy and releasing pure noise, as described on page 175:

Proposition 15. *Let M be a data-release mechanism. Denote the set of connected components of \mathcal{X} by*

$$\text{Comp}(\mathcal{X}) = \{[x] : x \in \mathcal{X}\},$$

where $[x] = \{x' \in \mathcal{X} \mid d(x, x') < \infty\}$ (see Definition 5.3.2). The following statements are equivalent:

I M satisfies ε -DP with $\varepsilon = 0$.

II M is a function of x only through its connected component $[x]$. That is, there exists a function $M' : \text{Comp}(\mathcal{X}) \times [0, 1] \rightarrow \mathcal{T}$ such that $M = M' \circ c$, where

$$\begin{aligned} c : \mathcal{X} \times [0, 1] &\rightarrow \text{Comp}(\mathcal{X}) \times [0, 1], \\ (x, u) &\mapsto ([x], u). \end{aligned}$$

III The probability P_x induced by M is a function of x only through $[x]$. That is, $P_x = P_{x'}$ whenever $[x] = [x']$, or equivalently, the map $x \mapsto P_x$ factors as $\varphi \circ c$, where

$$\begin{aligned} c : \mathcal{X} &\rightarrow \text{Comp}(\mathcal{X}), \\ x &\mapsto [x], \end{aligned}$$

and φ is some function which maps $[x]$ to a probability on $(\mathcal{T}, \mathcal{F})$.

Proof. The equivalence between II and III follows by the definition of P_x as the probability induced by M – see equation (5.1).

We now prove that I and III are equivalent. The mechanism M satisfies ε -DP with $\varepsilon = 0$ if and only if $D_{\text{MULT}}(P_x, P_{x'}) = 0$ for all x, x' with $d(x, x') < \infty$. Because metrics are positive definite (see Definition C.3.1.1), this is equivalent to $P_x = P_{x'}$ for all x, x' with $d(x, x') < \infty$. But $P_x = P_{x'}$ holds for $x' \in [x]$ if and only if P_x is a function of x only through $[x]$. \square

The following proposition formalises the result in Remark 5.4.2, which states that publishing $[x]$ alongside $T = M(x, U)$ does not increase the privacy loss but ensures that $\text{supp}(x \mid t, \theta)$ is connected. (Recall that $[x]$ is the connected component $[x] = \{x' \in \mathcal{X} \mid d(x, x') < \infty\}$.)

Proposition 16. *Let $M : \mathcal{X} \times [0, 1] \rightarrow \mathcal{T}$ be an ε -DP mechanism. Then the mechanism M' defined by*

$$\begin{aligned} M' : \mathcal{X} \times [0, 1] &\rightarrow 2^{\mathcal{X}} \times \mathcal{T}, \\ (x, u) &= ([x], M(x, u)), \end{aligned}$$

is also ε -DP, and, moreover, the support $\text{supp}(x \mid t, \theta)$ for M' is d -connected for all the possible outputs $t \in \mathcal{T} \times 2^{\mathcal{X}}$ of M' and all $\theta \in \Theta$.

Proof. Proposition 15 implies that the data-release mechanism

$$M_0 : \mathcal{X} \times [0, 1] \rightarrow \text{Comp}(\mathcal{X}),$$

$$(x, u) \mapsto [x],$$

is ε -DP with $\varepsilon = 0$. Observe that $P_x(M' \in \cdot)$ is the product measure of $P_x(M_0 \in \cdot)$ and $P_x(M \in \cdot)$.

Thus M' is also ε -DP by Proposition 11. This proves the first half of the proposition.

To see the second half of the proposition, fix some $\theta \in \Theta$ and some $t \in \mathcal{T}$. Let $E \subset \mathcal{X}$. If there does not exist some $x \in \mathcal{X}$ with $E = [x]$ then $\text{supp}(x \mid (t, E), \theta) = \emptyset$ which is trivially d -connected. On the other hand, if $E = [x]$ for some $x \in \mathcal{X}$ then $\text{supp}(x \mid (t, E), \theta) \subset [x]$. Since $[x]$ is d -connected by definition, $\text{supp}(x \mid (t, E), \theta)$ must also be. \square

C.5.7 ALTERNATIVE CHARACTERISATIONS OF THE PRIVATISED DATA PROBABILITY $P(T \in \cdot \mid \theta)$

The following proposition provides a number of characterisations of the privatised data probability $P(T \in \cdot \mid \theta)$, defined in equation (5.6). Recall that λ is the Lebesgue measure.

Proposition 17. *Let M be a data-release mechanism. Then the privatised data probability is given by:*

$$\begin{aligned} P(T \in S \mid \theta) &= \int_{\mathcal{X}} P_x(S) dP_{\theta}(x), \\ &= \int_{\mathcal{X} \times [0, 1]} 1_{M(x, u) \in S} d(P_{\theta} \times \lambda)(x, u), \end{aligned}$$

for every $S \in \mathcal{F}$, where $P_\theta \times \lambda$ is the product measure of P_θ and λ , and $1_{M(x,u) \in S}$ is the indicator function:

$$1_{M(x,u) \in S} = \begin{cases} 1 & \text{if } M(x, u) \in S, \\ 0 & \text{otherwise.} \end{cases}$$

Moreover, $P(T \in \cdot \mid \theta)$ is a well defined probability on $(\mathcal{T}, \mathcal{F})$, for every $\theta \in \Theta$.

Proof. Recall that M is $(\mathcal{G} \otimes \mathcal{B}[0, 1], \mathcal{F})$ -measurable. Hence $1_{M(x,u) \in S} : \mathcal{X} \times [0, 1] \rightarrow \mathbb{R}$ is also measurable for every $S \in \mathcal{F}$. Thus,

$$\int_{\mathcal{X} \times [0, 1]} 1_{M(x,u) \in S} d(P_\theta \times \lambda)(x, u)$$

is a well-defined probability on $(\mathcal{T}, \mathcal{F})$. Then

$$\begin{aligned} \int_{\mathcal{X} \times [0, 1]} 1_{M(x,u) \in S} d(P_\theta \times \lambda)(x, u) &= \int_{\mathcal{X}} \left(\int_{[0, 1]} 1_{M(x,u) \in S} du \right) dP_\theta(x) \\ &= \int_{\mathcal{X}} \lambda(\{u \in [0, 1] : M(x, u) \in S\}) dP_\theta(x) \\ &= \int_{\mathcal{X}} P_x(S) dP_\theta(x), \end{aligned}$$

where the first line follows by Fubini's theorem, and the second by the definition of the Lebesgue measure λ and the third by the definition of $P_x(S)$ in equation (5.1). (Note that $x \mapsto P_x(S)$ is indeed $(\mathcal{G}, \mathcal{F})$ -measurable – see (Durrett, 2019, Lemma 1.7.3).) \square

C.5.8 ALTERNATIVE CHARACTERISATIONS OF THE PROBABILITY OF A DATA-PROVISION PROCEDURE M_G

The following lemma will be useful in proving that the probability of a data-provision procedure M_G is well defined.

Lemma C.5.6. *Let (X, Σ_X) , (Y, Σ_Y) and (Z, Σ_Z) be measurable spaces. Suppose that $f : X \rightarrow Y$ is (Σ_X, Σ_Y) -measurable and $g : Y \times Z \rightarrow W$ is $(\Sigma_Y \otimes \Sigma_Z, \Sigma_W)$ -measurable.*

Then the map

$$h : X \times Z \rightarrow W,$$

$$(x, z) \mapsto g(f(x), z),$$

is $(\Sigma_X \otimes \Sigma_Z, \Sigma_W)$ -measurable.

Proof. Define

$$\varphi : X \times Z \rightarrow Y \times Z,$$

$$(x, z) \mapsto (f(x), z),$$

and

$$\mathcal{E} = \{E_2 \in \Sigma_Y \otimes \Sigma_Z \mid \exists E_1 \in \Sigma_X \otimes \Sigma_Z \text{ s.t. } \varphi^{-1}(E_2) = E_1\}.$$

It is easy to verify that \mathcal{E} is a σ -algebra because $\varphi^{-1}(\emptyset) = \emptyset$; $\varphi^{-1}(E_2^c) = [\varphi^{-1}(E_2)]^c$; and $\varphi^{-1}(\cup_i E_2^{(i)}) = \cup_i \varphi^{-1}(E_2^{(i)})$. (Here E_2^c is the complement of E_2 in $Y \times Z$.) Define the rectangles

$$\mathcal{R} = \{E \times F \mid E \in \Sigma_Y, F \in \Sigma_Z\}.$$

Because f is measurable, $\mathcal{R} \subset \mathcal{E}$ and hence $\Sigma_Y \otimes \Sigma_Z = \sigma(\mathcal{R}) = \mathcal{E}$.

Now let $E_3 \in \Sigma_W$. Because g is measurable, there exists $E_2 \in \Sigma_Y \otimes \Sigma_Z$ such that $g^{-1}(E_3) = E_2$. Because $\Sigma_Y \otimes \Sigma_Z \subset \mathcal{E}$, there exists $E_1 \in \Sigma_X \otimes \Sigma_Z$ such that $\varphi^{-1}(E_2) = E_1$. Thus, $h^{-1}(E_3) = \varphi^{-1}(g^{-1}[E_3]) = E_1$. □

The following proposition provides a number of characterisations of the probability $P(T \in \cdot \mid \theta)$ of the output of a data-provision procedure M_G , defined in equation (5.15).

Proposition 18. *Let M_G be a data-provision procedure. The probability on M_G 's output T is given by*

$$P(T \in S \mid \theta) = \lambda(\{u_1, u_2 \in [0, 1] : M_G(\theta, u_1, u_2) \in S\}) \quad (\text{C.19})$$

$$= \int_{[0,1]^2} 1_{M_G(\theta, u_1, u_2) \in S} d(u_1, u_2) \quad (\text{C.20})$$

$$= \int_{\mathcal{X}} P_x(S) dP_\theta(x), \quad (\text{C.21})$$

for $S \in \mathcal{F}$, where $1_{M_G(\theta, u_1, u_2) \in S}$ is the indicator function:

$$1_{M_G(\theta, u_1, u_2) \in S} = \begin{cases} 1 & \text{if } M_G(\theta, u_1, u_2) \in S, \\ 0 & \text{otherwise.} \end{cases}$$

Moreover, for each $\theta \in \Theta$, $P(T \in \cdot \mid \theta)$ is a well-defined probability on $(\mathcal{T}, \mathcal{F})$.

Proof. We first show that $P(T \in \cdot \mid \theta)$ is a well-defined probability on $(\mathcal{T}, \mathcal{F})$. By the definition of $P(T \in \cdot \mid \theta)$ as a pushforward probability in equation (5.15), it suffices to show that the map

$$M_G(\theta, \cdot, \cdot) : [0, 1]^2 \rightarrow \mathcal{T},$$

$$(u_1, u_2) \mapsto M_G(\theta, u_1, u_2),$$

is $(\mathcal{B}[0, 1]^2, \mathcal{F})$ -measurable for every $\theta \in \Theta$. (Here $\mathcal{B}[0, 1]^2$ is the Borel σ -algebra on $[0, 1] \times [0, 1]$.) Recall that we assume that M is $(\mathcal{G} \otimes \mathcal{B}[0, 1], \mathcal{F})$ -measurable and that $G(\theta, \cdot)$ is $(\mathcal{B}[0, 1], \mathcal{G})$ -measurable for all $\theta \in \Theta$. Thus, measurability of M_G follows from Lemma C.5.6 because $\mathcal{B}[0, 1]^2 = \mathcal{B}[0, 1] \otimes \mathcal{B}[0, 1]$.

Now we prove equations (C.19)-(C.21). Equation (C.19) is simply the definition of $P(T \in S \mid \theta)$,

as given in equation (5.15). Equation (C.20) follows from the definition of the Lebesgue measure λ on $[0, 1]^2$. Equation (C.21) follows by the calculation:

$$\begin{aligned}
\int_{[0,1]^2} \mathbf{1}_{M_G(\theta, u_1, u_2) \in S} d(u_1, u_2) &= \int_{[0,1]} \left(\int_{[0,1]} \mathbf{1}_{M(G(\theta, u_1), u_2) \in S} du_2 \right) du_1 \\
&= \int_{[0,1]} \left(\int_{[0,1]} \mathbf{1}_{M(x, u_2) \in S} du_2 \right) \mathbf{1}_{G(\theta, u_1) = x} du_1 \\
&= \int_{[0,1]} P_x(S) \mathbf{1}_{G(\theta, u_1) = x} du_1 \\
&= \int_{\mathcal{X}} P_x(S) dP_\theta(x),
\end{aligned}$$

where the first line is Fubini's theorem; the second line is a substitution of $G(\theta, u_1)$ with x ; the third line follows by the definition of P_x in equation (5.1):

$$\begin{aligned}
P_x(S) &= \lambda(\{u_2 \in [0, 1] : M(x, u_2) \in S\}) \\
&= \int_{[0,1]} \mathbf{1}_{M(x, u_2) \in S} du_2,
\end{aligned}$$

and the fourth line follows by the definition of P_θ in equation (5.14):

$$\begin{aligned}
P_\theta(X \in E) &= \lambda(\{u_1 \in [0, 1] : G(\theta, u_1) \in E\}) \\
&= \int_{[0,1]} \mathbf{1}_{G(\theta, u_1) \in E} du_1.
\end{aligned}$$

(Note that $P_x(S)$ is indeed $(\mathcal{G}, \mathcal{F})$ -measurable – see (Durrett, 2019, Lemma 1.7.3).)

□

C.5.9 PUFFERFISH BOUNDS THE PRIOR-TO-POSTERIOR ODDS RATIO

The following proposition formalises the result described by equation (5.17).

Proposition 19. *Fix a data-generating process G . Let M be a data-release mechanism. Then M satisfies*

ε -PufferFish(\mathbb{D}, \mathbb{S}) if and only if

$$e^{-\varepsilon} \leq \frac{P_{\theta^*}(X \in E \mid T = t)}{P_{\theta^*}(X \in E' \mid T = t)} \bigg/ \frac{P_{\theta^*}(X \in E)}{P_{\theta^*}(X \in E')} \leq e^{\varepsilon}, \quad (\text{C.22})$$

for all $\theta^* \in \mathbb{D}$, all $(E, E') \in \mathbb{S}$, and all $t \in \mathcal{T}_*$, where \mathcal{T}_*^c is a null set under $P(T \in \cdot \mid \theta^*, X \in E)$ and under $P(T \in \cdot \mid \theta^*, X \in E')$.

Proof. Suppose throughout that $\varepsilon < \infty$ (otherwise the proposition is vacuous).

“ \Rightarrow ”: Suppose that M satisfies ε -PufferFish(\mathbb{D}, \mathbb{S}). Fix some $\theta^* \in \mathbb{D}$ and some $(E, E') \in \mathbb{S}$ such that $P_{\theta^*}(X \in \cdot \mid X \in E)$ and $P_{\theta^*}(X \in \cdot \mid X \in E')$ are both well-defined. By 5.8.1.IV, there is a common dominating measure ν of $P(T \in \cdot \mid \theta^*, X \in E)$ and $P(T \in \cdot \mid \theta^*, X \in E')$. Additionally, $P(T \in \cdot \mid \theta^*, X \in E)$ and $P(T \in \cdot \mid \theta^*, X \in E')$ have ν -densities satisfying

$$e^{-\varepsilon} \leq \frac{p(t \mid \theta^*, X \in E)}{p(t \mid \theta^*, X \in E')} \leq e^{\varepsilon},$$

for all $t \in \mathcal{T}$. Now Bayes rule states that

$$P_{\theta^*}(X \in E \mid T = t) \propto P_{\theta^*}(X \in E) p(t \mid \theta^*, X \in E), \quad (\text{C.23})$$

for $P(T \in \cdot \mid \theta^*, X \in E)$ -almost all t . Let $\mathcal{T}_*^{(1)}$ be the null set where (C.23) doesn't hold. Similarly, let $\mathcal{T}_*^{(2)}$ be the $P(T \in \cdot \mid \theta^*, X \in E')$ -null set where

$$P_{\theta^*}(X \in E' \mid T = t) \propto P_{\theta^*}(X \in E') p(t \mid \theta^*, X \in E'),$$

does not hold. On $\mathcal{T}_* = (\mathcal{T}_1^* \cup \mathcal{T}_2^*)^c$, the posterior odds is equal to product of the prior odds and the likelihood ratio:

$$\frac{P_{\theta^*}(X \in E \mid T = t)}{P_{\theta^*}(X \in E' \mid T = t)} = \frac{P_{\theta^*}(X \in E)}{P_{\theta^*}(X \in E')} \frac{p(t \mid \theta^*, X \in E)}{p(t \mid \theta^*, X \in E')}.$$

Hence (C.22) holds for all $t \in \mathcal{T}_*$.

“ \Leftarrow ”: Fix some $\theta^* \in \mathbb{D}$ and some $(E, E') \in \mathbb{S}$ such that $P_{\theta^*}(X \in \cdot \mid X \in E)$ and $P_{\theta^*}(X \in \cdot \mid X \in E')$ are both well-defined. Suppose that (C.22) holds. We will show that 5.8.1.II must also hold. Firstly, we note that $P(T \in \cdot \mid \theta^*, X \in E)$ is absolutely continuous with respect to $P(T \in \cdot \mid \theta^*)$. Also,

$$\frac{P_{\theta^*}(X \in E \mid T = t)}{P_{\theta^*}(X \in E)}$$

is a $P(T \in \cdot \mid \theta^*)$ -density for $P(T \in \cdot \mid \theta^*, X \in E)$ because

$$P(T \in S \mid \theta^*, X \in E) = \int_S \frac{P_{\theta^*}(X \in E \mid T = t)}{P_{\theta^*}(X \in E)} dP(t \mid \theta^*)$$

by Bayes rule. All the above also applies when E is replaced by E' . Hence, for any $S \in \mathcal{F}$,

$$\begin{aligned} P(T \in S \mid \theta^*, X \in E) &= P(T \in S \cap \mathcal{T}_* \mid \theta^*, X \in E) \\ &= \int_{S \cap \mathcal{T}_*} \frac{P_{\theta^*}(X \in E \mid T = t)}{P_{\theta^*}(X \in E)} dP(t \mid \theta^*) \\ &\leq e^\varepsilon \int_{S \cap \mathcal{T}_*} \frac{P_{\theta^*}(X \in E' \mid T = t)}{P_{\theta^*}(X \in E')} dP(t \mid \theta^*) \\ &= e^\varepsilon P(T \in S \cap \mathcal{T}_* \mid \theta^*, X \in E') \\ &= e^\varepsilon P(T \in S \mid \theta^*, X \in E'). \end{aligned}$$

The second half of 5.8.1.II follows analogously. □

C.6 PROOFS OMITTED FROM THE MAIN TEXT

Recall that $\mu \ll \nu$ denotes that μ is absolutely continuous with respect to ν .

Proof. of Lemma 5.3.4: Fix two σ -finite measures $\mu, \nu \in \Omega$ and a constants $\varepsilon > 0$.

“ \Rightarrow ”: Suppose that $\nu \in \mathcal{I}(e^{-\varepsilon}\mu, e^\varepsilon\mu)$. Then, for all $S \in \mathcal{F}$,

$$\nu(S) \geq e^{-\varepsilon}\mu(S),$$

and hence

$$\ln \frac{\mu(S)}{\nu(S)} \leq \varepsilon.$$

Similarly,

$$-\ln \frac{\mu(S)}{\nu(S)} = \ln \frac{\nu(S)}{\mu(S)} \leq \varepsilon,$$

since $\nu(S) \leq e^\varepsilon\mu(S)$. Putting these two results together,

$$D_{\text{MULT}}(\mu, \nu) = \sup_{S \in \mathcal{F}} \left| \ln \frac{\mu(S)}{\nu(S)} \right| \leq \varepsilon.$$

“ \Leftarrow ”: Suppose that $D_{\text{MULT}}(\mu, \nu) \leq \varepsilon$. Then, for all $S \in \mathcal{F}$,

$$\frac{\mu(S)}{\nu(S)} \leq e^\varepsilon.$$

This proves $e^{-\varepsilon}\mu \leq \nu$. We also have that

$$-\ln \frac{\mu(S)}{\nu(S)} \leq \varepsilon,$$

for all $S \in \mathcal{F}$, which implies $\nu \leq e^\varepsilon\mu$. Thus, $\nu \in \mathcal{I}(e^{-\varepsilon}\mu, e^\varepsilon\mu)$.

Now we prove the second half of the lemma. Fix two constants $0 < a \leq 1 \leq b < \infty$. Suppose $\nu \in \mathcal{I}(a\mu, b\mu)$. Let $\varepsilon = \max(-\ln a, \ln b)$. Since $\mathcal{I}(a\mu, b\mu) \subset \mathcal{I}(e^{-\varepsilon}\mu, e^\varepsilon\mu)$, we have $D_{\text{MULT}} \leq \varepsilon$ by the result of the first half of the lemma. Now let $\varepsilon = \min(-\ln a, \ln b)$ and suppose $D_{\text{MULT}}(\mu, \nu) \leq \varepsilon$. Then $\nu \in \mathcal{I}(e^{-\varepsilon}\mu, e^\varepsilon\mu) \subset \mathcal{I}(a\mu, b\mu)$ as required. \square

Proof. of Theorem 5.3.5: First we prove that I implies II. Suppose that M satisfies pure ε -DP. Fix some

$S \in \mathcal{F}$ and some $x, x' \in \mathcal{X}$ with $d(x, x') = 1$. By assumption, $D_{\text{MULT}}(P_x, P_{x'}) \leq \varepsilon$. This implies $P_{x'}(S) \leq e^\varepsilon P_x(S)$ by Lemma 5.3.4, which is exactly II.

Next we will prove that II implies I. To do this, we need the following lemma (*): II implies that $D_{\text{MULT}}(P_x, P_{x'}) \leq \varepsilon$ for all $x, x' \in \mathcal{X}$ with $d(x, x') = 1$.

To prove this lemma, suppose II holds and fix some $x, x' \in \mathcal{X}$ with $d(x, x') = 1$. By assumption $P_{x'}(S) \leq e^\varepsilon P_x(S)$ for all $S \in \mathcal{F}$. Yet, by symmetry of d (i.e. because $d(x', x) = d(x, x') = 1$), II also implies that $P_x(S) \leq e^\varepsilon P_{x'}(S)$ for all $S \in \mathcal{F}$. Then Lemma 5.3.4 provides the desired result: $D_{\text{MULT}}(P_x, P_{x'}) \leq \varepsilon$.

Now we return to proving that II implies I. Fix $x, x' \in \mathcal{X}$. If $d(x, x') = \infty$ then the condition $D_{\text{MULT}}(P_x, P_{x'}) \leq \varepsilon d(x, x')$ is vacuous. Similarly, if $d(x, x') = 0$ then, by the properties of d as a metric, $x = x'$. Thus $P_x = P_{x'}$ and $D_{\text{MULT}}(P_x, P_{x'}) = 0$.

On the other hand, if $d(x, x') = n < \infty$, then by Assumption 5.3.3, there exists $x = x_0, x_1, \dots, x_n = x' \in \mathcal{X}$ with $d(x_i, x_{i+1}) = 1$ for all $0 \leq i \leq n-1$. Then,

$$D_{\text{MULT}}(P_x, P_{x'}) \leq \sum_{i=0}^{n-1} D_{\text{MULT}}(P_{x_i}, P_{x_{i+1}}) \leq \sum_{i=0}^{n-1} \varepsilon = \varepsilon d(x, x'),$$

where the first line follows by the triangle inequality (Lemma C.3.3) of D_{MULT} , and the second by (*).

This proves that II implies I.

Now we prove that I is equivalent to III. Fix some $x, x' \in \mathcal{X}$ with $\delta = d(x, x')$. Then \mathcal{M} is ε -DP if and only if $D_{\text{MULT}}(P_x, P_{x'}) \leq \delta \varepsilon$ for all such $x, x' \in \mathcal{X}$. Yet Lemma 5.3.4 states that this is equivalent to $P_{x'} \in \mathcal{I}(e^{-\delta \varepsilon} P_x, e^{\delta \varepsilon} P_x)$, which is III.

We move to proving that III implies IV. This follows by Proposition 14. Fix $x, x' \in \mathcal{X}$ and $\nu \in \Omega$. Suppose that x' and x are d -connected, so that $d(x, x') = \delta < \infty$. Assume that III holds for these x, x' . Suppose that P_x has a ν -density p_x . By the Radon-Nikodym theorem, this implies $P_x \ll \nu$. Then we can

apply Proposition 14(a) with $\tau = P_x$, $a = \exp(-\delta\varepsilon)$, $b = \exp(\delta\varepsilon)$ and $g = p_x$. Then $P_{x'}$ has a ν -density which is in the density bounded class $\mathcal{I}(\exp(-\delta\varepsilon)p_x, \exp(\delta\varepsilon)p_x)$. This is precisely IV.

Finally, we prove that IV implies III. Set $\nu = P_x$ so that P_x has a constant ν -density $p_x = 1$. Then IV implies that $p_{x'} \in \mathcal{I}(l, u)$, where $l = \exp(-\delta\varepsilon)$ and $u = \exp(\delta\varepsilon)$ are also constant ν -densities. Apply the second half of Proposition 14. This states that $P_{x'} \in \mathcal{I}(L, U)$ where L and U are defined as

$$L(S) = \int_S e^{-\delta\varepsilon} dP_x = e^{-\delta\varepsilon} P_x(S),$$

and

$$U(S) = \int_S e^{\delta\varepsilon} dP_x = e^{\delta\varepsilon} P_x(S).$$

Yet this is exactly III. □

We now turn to proving Theorem 5.4.1. We first need to establish some lemmata.

Lemma C.6.1. *Consider the same set-up as in Theorem 5.4.1. If $x \in \text{supp}_0(x \mid t_0)^\complement \cap \text{supp}(P_\theta)$, then $P_x(\mathcal{T}_0) = 0$.*

Proof. For $t \in \mathcal{T}_0$, we have

$$\begin{aligned} \text{supp}_0(x \mid t_0)^\complement \cap \text{supp}(P_\theta) &\subset (\text{supp}(P_\theta) \cap \text{supp}_0(x \mid t_0))^\complement \cap \text{supp}(P_\theta) \\ &\subset (\text{supp}(P_\theta) \cap \text{supp}_0(x \mid t))^\complement \cap \text{supp}(P_\theta) \\ &\subset \text{supp}_0(x \mid t)^\complement \\ &= \{x \in \mathcal{X} \mid t \in \text{supp}(P_x)^\complement\}. \end{aligned}$$

Therefore, for $x \in \text{supp}_0(x \mid t_0)^\complement \cap \text{supp}(P_\theta)$, we have that $\mathcal{T}_0 \in \text{supp}(P_x)^\complement$. The result then follows by (C.2). □

Lemma C.6.2. *Consider the same set-up as in Theorem 5.4.1. Then*

$$\int_{\text{supp}(x|t_0, \theta)^c} P_x(\mathcal{T}_0) dP_\theta(x) = 0.$$

Proof. Compute

$$\begin{aligned} \int_{\text{supp}(x|t_0, \theta)^c} P_x(\mathcal{T}_0) dP_\theta(x) &= \int_{\text{supp}(P_\theta)^c} P_x(\mathcal{T}_0) dP_\theta(x) + \int_{\text{supp}_0(x|t_0)^c \cap \text{supp}(P_\theta)} P_x(\mathcal{T}_0) dP_\theta(x) \\ &= \int_{\text{supp}(P_\theta)^c} P_x(\mathcal{T}_0) dP_\theta(x) \\ &\leq P_\theta(\text{supp}(P_\theta)^c) \\ &= 0, \end{aligned}$$

where the second line follows by Lemma C.6.1 and the fourth by (C.2). \square

Lemma C.6.3. *Consider the same set-up as in Theorem 5.4.1. Then, a density $p(t \mid \theta)$ for $P(T \in \cdot \mid \theta)$ exists in $\mathcal{T}_0 = \{t \in \mathcal{T} \mid \text{supp}(x \mid t, \theta) \subset \text{supp}(x \mid t_0, \theta)\}$ in the sense that*

$$P(T \in S \cap \mathcal{T}_0 \mid \theta) = \int_{S \cap \mathcal{T}_0} p(t \mid \theta) d\nu(t),$$

for all $S \in \mathcal{F}$, where ν is the dominating measure for the density $p(t \mid \theta)$. Moreover, this density satisfies

$$p(t \mid \theta) \in p_{x_*}(t) \exp(\pm \varepsilon d_*), \tag{C.24}$$

for every $t \in \mathcal{T}_0$ and every $x_* \in \text{supp}(x \mid t_0, \theta)$, where $d_* = \sup_{x \in \text{supp}(x|t_0, \theta)} d(x, x_*)$.

Proof. Fix some $x_* \in \text{supp}(x \mid t_0, \theta)$ and some $\nu \in \Omega$ with $P_{x_*} \ll \nu$. Let p_{x_*} be a ν -density of P_{x_*} . By Theorem 5.3.5.IV and the assumption that $\text{supp}(x \mid t_0, \theta)$ is d -connected, the probability P_x also has a

ν -density p_x , for all $x \in \text{supp}(x \mid t_0, \theta)$. For $t \in \mathcal{T}_0$, define

$$p(t \mid \theta) = \int_{\text{supp}(x \mid t_0, \theta)} p_x(t) dP_\theta(x).$$

We want to prove that $p(t \mid \theta)$ is a ν -density of $P(T \in \cdot \mid \theta)$ in \mathcal{T}_0 – namely, that

$$P(T \in S \cap \mathcal{T}_0 \mid \theta) = \int_{S \cap \mathcal{T}_0} p(t \mid \theta) d\nu(t),$$

for all $S \in \mathcal{F}$.

We have

$$\begin{aligned} \int_{S \cap \mathcal{T}_0} p(t \mid \theta) d\nu(t) &= \int_{S \cap \mathcal{T}_0} \left(\int_{\text{supp}(x \mid t_0, \theta)} p_x(t) dP_\theta(x) \right) d\nu(t) \\ &= \int_{\text{supp}(x \mid t_0, \theta)} \left(\int_{S \cap \mathcal{T}_0} p_x(t) d\nu(t) \right) dP_\theta(x) \\ &= \int_{\text{supp}(x \mid t_0, \theta)} P_x(S \cap \mathcal{T}_0) dP_\theta(x) \\ &= P(T \in S \cap \mathcal{T}_0 \mid \theta), \end{aligned}$$

where the second line follows by Fubini's theorem and the fourth by Lemma C.6.2.

Now we move to proving the upper bound of (C.24):

$$\begin{aligned} p(t \mid \theta) &= \int_{\text{supp}(x \mid t_0, \theta)} p_x(t) dP_\theta(x) \\ &\leq \int_{\text{supp}(x \mid t_0, \theta)} e^{\varepsilon d(x, x_*)} p_{x_*}(t) dP_\theta(x) \\ &\leq e^{\varepsilon d_*} p_{x_*}(t), \end{aligned}$$

where the second line follows from Theorem 5.3.5.IV. The lower bound follows similarly. \square

Proof. of Theorem 5.4.1: By Theorem 5.3.5, we can fix a measure $\nu \in \Omega$ which dominates all P_x for

$x \in \text{supp}(x \mid t_0, \theta)$. Let p_x denote a ν -density of P_x for $x \in \text{supp}(x \mid t_0, \theta)$. Take the essential supremum with respect to ν over the collection of densities $\{\exp(-\varepsilon d_*) p_{x_*}(t) : x_* \in \text{supp}(x \mid t_0, \theta)\}$ to produce

$$l_{\theta, \varepsilon}(t) = \text{ess sup}_{x_* \in \text{supp}(x \mid t_0, \theta)} \exp(-\varepsilon d_*) p_{x_*}(t). \quad (\text{C.25})$$

This function $l_{\theta, \varepsilon} : \mathcal{T} \rightarrow [0, \infty]$ exists and is measurable as ν is σ -finite. Thus, $l_{\theta, \varepsilon}$ is a ν -density for some measure $L_{\theta, \varepsilon}$ on $(\mathcal{T}, \mathcal{F})$. By Lemma C.6.3, we can construct a ν -density $p(t \mid \theta)$ for $P(T \in \cdot \cap \mathcal{T}_0 \mid \theta)$ which satisfies

$$p(t \mid \theta) \geq l_{\theta, \varepsilon}(t),$$

for ν -almost all $t \in \mathcal{T}_0$. This proves that $P(T \in S \mid \theta) \geq L_{\theta, \varepsilon}(S)$ for any measurable $S \subset \mathcal{T}_0$.

Since $P(T \in \cdot \cap \mathcal{T}_0 \mid \theta)$ is zero outside of \mathcal{T}_0 , technically we must modify $l_{\theta, \varepsilon}$ to also be zero outside of \mathcal{T}_0 . It then follows that $L_{\theta, \varepsilon}(S) \leq P(T \in S \cap \mathcal{T}_0 \mid \theta)$ for all $S \in \mathcal{F}$. This proves the lower bound of (5.7).

The argument for the upper measure $U_{\theta, \varepsilon}$ is almost analogous. We can construct $u_{\theta, \varepsilon}$ on \mathcal{T} as the essential infimum

$$u_{\theta, \varepsilon}(t) = \text{ess inf}_{x_* \in \text{supp}(x \mid t_0, \theta)} \exp(\varepsilon d_*) p_{x_*}(t).$$

(Note that it is possible that $u_{\theta, \varepsilon}$ is not finite almost everywhere – even though all of the $p_{x_*}(t)$ are – so that the measure $U_{\theta, \varepsilon}$ is not σ -finite by Lemma C.5.1.) Then Lemma C.6.3 implies that

$$p(t \mid \theta) \leq u_{\theta, \varepsilon}(t),$$

for ν -almost all $t \in \mathcal{T}_0$. This proves that $P(T \in S \mid \theta) \leq U_{\theta, \varepsilon}(S)$ for any measurable $S \subset \mathcal{T}_0$. Thus, $P(T \in S \cap \mathcal{T}_0 \mid \theta) \leq U_{\theta, \varepsilon}(S)$ for any $S \in \mathcal{F}$. This proves the upper bound of (5.7). \square

Proof. of Theorem 5.5.1: Fix some $x_* \in S_0 \cup S_1$ and some $\nu \in \Omega$ with $P_{x_*} \ll \nu$. By Theorem 5.3.5.IV

and the assumption that $S_0 \cup S_1$ is d -connected, the probability P_x has a ν -density p_x , for all $x \in S_0 \cup S_1$.

For $t \in \mathcal{T}$, define

$$p(t \mid \theta_i) = \int_{S_i} p_x(t) dP_{\theta_i}(x).$$

We can show that $p(t \mid \theta_i)$ is a ν -density of $P(T \in \cdot \mid \theta_0)$: For any $E \in \mathcal{F}$,

$$\begin{aligned} \int_E p(t \mid \theta_i) d\nu(t) &= \int_E \int_{S_i} p_x(t) dP_{\theta_i}(x) d\nu(t) \\ &= \int_{S_i} \int_E p_x(t) d\nu(t) dP_{\theta_i}(x) \\ &= \int_{S_i} P_x(E) dP_{\theta_i}(x) \\ &= P(T \in E \mid \theta_i), \end{aligned}$$

where the second line follows by Fubini's theorem and the fourth by (C.2).

Let R be the rejection region of a test with size $P(T \in R \mid \theta_0) \leq \alpha$. Then

$$\begin{aligned} P(T \in R \mid \theta_1) &= \int_R p(t \mid \theta_1) d\nu(t) \\ &\leq \exp(d_{**}\varepsilon) \int_R p(t \mid \theta_0) d\nu(t) \\ &\leq \alpha \exp(d_{**}\varepsilon), \end{aligned}$$

where the second line follows by the computation:

$$\begin{aligned} p(t \mid \theta_1) &= \int_{S_1} p_x(t) dP_{\theta_1}(x) \\ &= \int_{S_0} \left(\int_{S_1} p_x(t) dP_{\theta_1}(x) \right) dP_{\theta_0}(x') \\ &\in \int_{S_0} \left(\int_{S_1} \exp(\pm \varepsilon d_{**}) p_{x'}(t) dP_{\theta_1}(x) \right) dP_{\theta_0}(x') \end{aligned}$$

$$\begin{aligned}
&= \exp(\pm \varepsilon d_{**}) \int_{S_0} \left(\int_{S_1} dP_{\theta_1}(x) \right) p_{x'}(t) dP_{\theta_0}(x') \\
&= \exp(\pm \varepsilon d_{**}) p(t \mid \theta_0).
\end{aligned}$$

In the above computation, the third line follows from Theorem 5.3.5.IV and the other lines simply pull constant factors into – or out of – some integral over dP_{θ_i} . \square

We begin proving Theorem 5.6.2 by establishing the following lemma (which is analogous to Lemma C.6.3).

Lemma C.6.4. *Consider the same set-up as in Theorem 5.6.2. Then, a density $p(t)$ for the prior predictive probability $P(T \in \cdot)$ exists in \mathcal{T}_0 in the sense that*

$$P(T \in S \cap \mathcal{T}_0) = \int_{S \cap \mathcal{T}_0} p(t) d\nu(t), \quad (\text{C.26})$$

for all $S \in \mathcal{F}$, where ν is the dominating measure for the density $p(t)$. Moreover, this density satisfies

$$p(t) \in p_{x_*}(t) \exp(\pm \varepsilon d_*), \quad (\text{C.27})$$

for every $t \in \mathcal{T}_0$ and every $x_* \in \text{supp}(x \mid t_0)$, where $d_* = \sup_{x \in \text{supp}(x \mid t_0)} d(x, x_*)$.

Proof. We proceed as for the proof of Lemma C.6.3. Fix some $x_* \in \text{supp}(x \mid t_0)$ and some $\nu \in \Omega$ with $P_{x_*} \ll \nu$. Let p_{x_*} be a ν -density of P_{x_*} . By Theorem 5.3.5.IV and the assumption that $\text{supp}(x \mid t_0)$ is d -connected, the probability P_x also has a ν -density p_x , for all $x \in \text{supp}(x \mid t_0)$. For $t \in \mathcal{T}_0$, define

$$p(t) = \int_{\Theta} \left(\int_{\text{supp}(x \mid t_0)} p_x(t) dP_{\theta}(x) \right) d\pi(\theta).$$

We can compute

$$\int_{S \cap \mathcal{T}_0} p(t) d\nu(t) = \int_{S \cap \mathcal{T}_0} \int_{\Theta} \int_{\text{supp}(x \mid t_0)} p_x(t) dP_{\theta}(x) d\pi(\theta) d\nu(t)$$

$$\begin{aligned}
&= \int_{\Theta} \int_{\text{supp}(x|t_0)} \int_{S \cap \mathcal{T}_0} p_x(t) d\nu(t) dP_{\theta}(x) d\pi(\theta) \\
&= \int_{\Theta} \int_{\text{supp}(x|t_0)} P_x(S \cap \mathcal{T}_0) dP_{\theta}(x) d\pi(\theta) \\
&= \int_{\Theta} P(T \in S \cap \mathcal{T}_0 \mid \theta) d\pi(\theta) \\
&= P(T \in S \cap \mathcal{T}_0),
\end{aligned}$$

where the second line follows by Fubini's theorem and the fourth by Lemma C.6.2. This proves (C.26).

Now we move to proving the upper bound of (C.27):

$$\begin{aligned}
p(t) &= \int_{\Theta} \left(\int_{\text{supp}(x|t_0)} p_x(t) dP_{\theta}(x) \right) d\pi(\theta) \\
&\leq \int_{\Theta} \left(\int_{\text{supp}(x|t_0)} e^{\varepsilon d(x, x_*)} p_{x_*}(t) dP_{\theta}(x) \right) d\pi(\theta) \\
&\leq e^{\varepsilon d_*} p_{x_*}(t),
\end{aligned}$$

where the second line follows from Theorem 5.3.5.IV. The lower bound of (C.27) follows similarly. \square

Proof. of Theorem 5.6.2: This is exactly analogous to the proof of Theorem 5.4.1, where references to Lemma C.6.3 are replaced by references to Lemma C.6.4; $p(t \mid \theta)$ is replaced with $p(t)$; and $\text{supp}(x \mid t_0, \theta)$ is replaced with $\text{supp}(x \mid t_0)$. \square

Proof. of Theorem 5.6.3: By Theorem 5.3.5, we can fix a measure $\nu \in \Omega$ which dominates all P_x for $x \in \text{supp}(x \mid t_0)$. Let p_x denote a ν -density of P_x for $x \in \text{supp}(x \mid t_0)$. Define

$$p(t \mid \theta) = \int_{\text{supp}(x|t_0)} p_x(t) dP_{\theta}(x).$$

Exactly as in the proof of Lemma C.6.3, we can show that $p(t \mid \theta)$ is a ν -density of $P(T \in \cdot \mid \theta)$ in \mathcal{T}_0 .

Since $t_0 \in \mathcal{T}_0$, we can use this density to compute the posterior via Bayes rule:

$$\pi(\theta \mid t_0) = \frac{p(t_0 \mid \theta)\pi(\theta)}{\int_{\Theta} p(t_0 \mid \theta')d\pi(\theta')}.$$

We have that

$$\begin{aligned} \frac{p(t_0 \mid \theta)\pi(\theta)}{\int_{\Theta} p(t_0 \mid \theta')d\pi(\theta')} &\leq \frac{p(t_0 \mid \theta)\pi(\theta)}{\int_{\Theta} \exp(-\varepsilon d_{**})p(t_0 \mid \theta')d\pi(\theta')} \\ &= \exp(\varepsilon d_{**})\pi(\theta), \end{aligned} \tag{C.28}$$

where the first line follows by the calculations

$$\begin{aligned} p(t_0 \mid \theta') &= \int_{\text{supp}(x|t_0)} p_x(t) dP_{\theta'}(x) \\ &= \int_{\text{supp}(x|t_0)} \left(\int_{\text{supp}(x|t_0)} p_x(t) dP_{\theta'}(x) \right) dP_{\theta'}(x') \\ &\leq \int_{\text{supp}(x|t_0)} \left(\int_{\text{supp}(x|t_0)} \exp(\varepsilon d_{**}) p_{x'}(t) dP_{\theta'}(x) \right) dP_{\theta'}(x') \\ &= \exp(\varepsilon d_{**}) \int_{\text{supp}(x|t_0)} p_{x'}(t) dP_{\theta'}(x') \\ &= p(t_0 \mid \theta). \end{aligned}$$

In the above computation, the third line follows from Theorem 5.3.5.IV and the other lines simply pull constant factors into – or out of – some integral over $dP_{\theta'}(x')$ or $dP_{\theta'}(x)$.

Using (C.28), we obtain the upper bound of (5.12):

$$\begin{aligned} \pi(\theta \in S \mid t_0) &= \int_S \pi(\theta \mid t_0) d\mu(\theta) \\ &\leq \int_S \exp(\varepsilon d_{**}) \pi(\theta) d\mu(\theta) \end{aligned}$$

$$= \exp(\varepsilon d_{**}) \pi(\theta \in S),$$

where μ is the dominating measure of the prior density $\pi(\theta)$. The proof of the lower bound of (5.12) is analogous. □

D

Appendices to Chapter 6

D.I PROOFS

Proof of Lemma 6.2.7. Let $T \in \mathcal{M}(\mathcal{X}, \mathcal{D}', d_{\mathcal{D}_0}, D_{\text{Pr}}, \varepsilon_{\mathcal{D}'})$. Take some $\mathcal{D} \in \mathcal{D}$ and some $\mathfrak{d}, \mathfrak{d}' \in \mathcal{D}$.

Then

$$\begin{aligned} D_{\text{Pr}}(\mathbf{P}_{\mathfrak{d}}, \mathbf{P}_{\mathfrak{d}'}) &\leq \inf\{\varepsilon_{\mathcal{D}'} d_{\mathcal{D}_0}(\mathfrak{d}, \mathfrak{d}') : \mathcal{D}' \in \mathcal{D}' \text{ s.t. } \mathfrak{d}, \mathfrak{d}' \in \mathcal{D}'\} \\ &\leq \inf\{\varepsilon_{\mathcal{D}'} d_{\mathcal{D}_0}(\mathfrak{d}, \mathfrak{d}') : \mathcal{D}' \in \mathcal{D}' \text{ s.t. } \mathcal{D} \subset \mathcal{D}'\} \\ &= \varepsilon_{\mathcal{D}} d_{\mathcal{D}_0}(\mathfrak{d}, \mathfrak{d}'), \end{aligned}$$

where

$$\varepsilon_{\mathcal{D}} = \inf\{\varepsilon_{\mathcal{D}'} : \mathcal{D}' \in \mathcal{D}' \text{ s.t. } \mathcal{D} \subset \mathcal{D}'\}.$$

□

Proof of Lemma 6.2.11. Let $\mathcal{D}' \in \overline{\mathcal{D}}$. Then there exists some $\mathcal{D} \in \mathcal{D}$ and $\mathfrak{d} \in \mathcal{D}$ such that $\mathcal{D}' = \mathcal{D} \cap [\mathfrak{d}]$.

Since every $\mathfrak{d}', \mathfrak{d}'' \in [\mathfrak{d}]$ are connected, it follows that every $\mathfrak{d}', \mathfrak{d}'' \in \mathcal{D}'$ are also connected. This proves that $\overline{\mathcal{D}}$ is complete.

Suppose that $T \in \mathcal{M}(\mathcal{X}, \overline{\mathcal{D}}, d_{\mathcal{D}_0}, D_{\text{Pr}}, \varepsilon_{\mathcal{D}'})$. Take some $\mathcal{D} \in \mathcal{D}$ and some $\mathfrak{d}, \mathfrak{d}' \in \mathcal{D}$. We wish to show that

$$D_{\text{Pr}}(\mathbf{P}_{\mathfrak{d}}, \mathbf{P}_{\mathfrak{d}'}) \leq \varepsilon_{\mathcal{D}} d_{\mathcal{D}_0}(\mathfrak{d}, \mathfrak{d}'). \quad (\text{D.1})$$

We may assume without loss of generality that $d_{\mathcal{D}_0}(\mathfrak{d}, \mathfrak{d}') < \infty$. Define $\mathcal{D}' = \mathcal{D} \cap [\mathfrak{d}]$. Since $\mathcal{D}' \in \overline{\mathcal{D}}$ and $\mathfrak{d}, \mathfrak{d}' \in \mathcal{D}'$, we know that

$$D_{\text{Pr}}(\mathbf{P}_{\mathfrak{d}}, \mathbf{P}_{\mathfrak{d}'}) \leq \varepsilon_{\mathcal{D}'} d_{\mathcal{D}_0}(\mathfrak{d}, \mathfrak{d}').$$

(D.1) then follows by observing that $\varepsilon_{\mathcal{D}'} \leq \varepsilon_{\mathcal{D}}$.

Suppose that $T \in \mathcal{M}(\mathcal{D}_0, \mathcal{D}, d_{\mathcal{D}_0}, D_{\text{Pr}}, \varepsilon_{\mathcal{D}})$. Take some $\mathcal{D}' \in \overline{\mathcal{D}}$ and some $\mathfrak{d}, \mathfrak{d}' \in \mathcal{D}'$. Then

$$\begin{aligned} D_{\text{Pr}}(\mathbf{P}_{\mathfrak{d}}, \mathbf{P}_{\mathfrak{d}'}) &\leq \inf\{\varepsilon_{\mathcal{D}} d_{\mathcal{D}_0}(\mathfrak{d}, \mathfrak{d}') : \mathcal{D} \in \mathcal{D} \text{ s.t. } \mathfrak{d}, \mathfrak{d}' \in \mathcal{D}\} \\ &\leq \inf\{\varepsilon_{\mathcal{D}} d_{\mathcal{D}_0}(\mathfrak{d}, \mathfrak{d}') : \mathcal{D} \in \mathcal{D} \text{ s.t. } \mathcal{D}' \subset \mathcal{D}\} \\ &= \varepsilon_{\mathcal{D}'} d_{\mathcal{D}_0}(\mathfrak{d}, \mathfrak{d}'). \end{aligned}$$

□

Proof of Theorem 6.4.3. Let $\mathcal{D} \in \mathcal{D}$ and $\mathfrak{d}, \mathfrak{d}' \in \mathcal{D}$. The density of $\mathbf{P}_{\mathfrak{d}}(T \in \cdot)$ is

$$f_x(t) = (2\Delta_q([x]_{\mathcal{D}}))^{-k} \exp\left(-\frac{\|t - q(\mathfrak{d})\|_1}{\Delta_q([x]_{\mathcal{D}})}\right).$$

Thus,

$$\begin{aligned}
D_{\text{MULT}}(\mathbf{P}_{\mathfrak{d}}, \mathbf{P}_{\mathfrak{d}'}) &= \sup_{t \in \mathbb{R}} \left| \ln \frac{f_x(t)}{f_{x'}(t)} \right| \\
&= \sup_{t \in \mathbb{R}} \left| \frac{\|t - q(\mathfrak{d}')\|_1 - \|t - q(\mathfrak{d})\|_1}{\Delta_q([x]_{\mathcal{D}})} \right| \\
&\leq \varepsilon d_{\mathcal{D}_0}(\mathfrak{d}, \mathfrak{d}'),
\end{aligned}$$

where the first line follows by Proposition 38 of [Baile and Gong \(2024\)](#), the second because $[\mathfrak{d}]_{\mathcal{D}} = [\mathfrak{d}']_{\mathcal{D}}$ and the third by the reverse triangle inequality. \square

References

- Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., and Zhang, L. (2016). Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, pages 308–318.
- Abowd, J., Ashmead, R., Cumings-Menon, R., Garfinkel, S., Heineck, M., Heiss, C., Johns, R., Kifer, D., Leclerc, P., Machanavajjhala, A., Moran, B., Sexton, W., Spence, M., and Zhuravlev, P. (2022a). The 2020 Census disclosure avoidance system TopDown Algorithm. *Harvard Data Science Review*, (Special Issue 2).
- Abowd, J. M. (2018). The US Census Bureau adopts differential privacy. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2867–2867. ACM.
- Abowd, J. M. (2021). Declaration of John M. Abowd. Exhibit A, Document 10-1, Case 1:21-cv-01361-ABJ of the United States District Court for the District of Columbia.
- Abowd, J. M., Adams, T., Ashmead, R., Darais, D., Dey, S., Garfinkel, S. L., Goldschlag, N., Kifer, D., Leclerc, P., Lew, E., Moore, S., Rodríguez, R. A., Tadros, R. N., and Vilhuber, L. (2023). The 2010 Census confidentiality protections failed, here’s how and why. Working Paper 31995.
- Abowd, J. M., Ashmead, R., Cumings-Menon, R., Garfinkel, S., Heineck, M., Heiss, C., Johns, R., Kifer, D., Leclerc, P., Machanavajjhala, A., Moran, B., Sexton, W., Spence, M., and Zhuravlev, P. (2022b). The 2020 Census disclosure avoidance system TopDown algorithm. *Harvard Data Science Review*, (Special Issue 2).
- Abowd, J. M. and Hawes, M. B. (2023). Confidentiality protection in the 2020 US Census of Population and Housing. *Annual Review of Statistics and Its Application*, 10:119–144.
- Abowd, J. M. and Schmutte, I. M. (2016). Economic analysis and statistical disclosure limitation. *Brookings Papers on Economic Activity*, 2015(1):221–293.

- Abowd, J. M. and Schmutte, I. M. (2019). An economic analysis of privacy protection and statistical accuracy as social choices. *American Economic Review*, 109(1):171–202.
- Abowd, J. M., Schneider, M. J., and Vilhuber, L. (2013). Differential privacy applications to Bayesian and linear mixed model estimation. *Journal of Privacy and Confidentiality*, 5(1).
- Acquisti, A., Brandimarte, L., and Loewenstein, G. (2015). Privacy and human behavior in the age of information. *Science*, 347(6221):509–514.
- Acquisti, A., Brandimarte, L., and Loewenstein, G. (2020). Secrets and likes: The drive for privacy and the difficulty of achieving it in the digital age. *Journal of Consumer Psychology*, 30(4):736–758.
- Acquisti, A., Taylor, C., and Wagman, L. (2016). The Economics of Privacy. *Journal of Economic Literature*, 54(2):442–492.
- Adjerid, I., Samat, S., and Acquisti, A. (2016). A Query-Theory Perspective of Privacy Decision Making. *The Journal of Legal Studies*, 45(S2):S97–S121.
- Advisory Committee on Data for Evidence Building (2022). Year 2 report. Technical report.
- Agre, P. and Rotenberg, M., editors (1997). *Technology and Privacy: The New Landscape*. MIT Press, Cambridge, Mass., 1st edition.
- Allen, A. L. (1988). *Uneasy Access: Privacy for Women in a Free Society*. Rowman & Littlefield, Totowa, NJ.
- Allen, A. L. (2007). Privacy. In Staples, W. G., editor, *Encyclopedia of Privacy*, pages 393–403. Greenwood Press, Westport, Conn.
- Allen, A. L. (2010). Privacy Torts: Unreliable Remedies for LGBT Plaintiffs. *California Law Review*, 98(6):1711–1764.
- Allen, A. L. and Rotenberg, M. (2016). *Privacy Law and Society*. American Casebook Series. West Academic Publishing, St. Paul, MN, third edition edition.
- Altman, I. (1977). Privacy regulation: Culturally universal or culturally specific? *Journal of Social Issues*, 33(3):66–84.
- Altman, I. and Taylor, D. A. (1973). *Social Penetration: The Development of Interpersonal Relationships*. Holt, Rinehart & Winston, Oxford, England.

- Andrés, M. E., Bordenabe, N. E., Chatzikokolakis, K., and Palamidessi, C. (2013). Geo-indistinguishability: Differential privacy for location-based systems. In *Proceedings of the 2013 ACM SIGSAC Conference on Computer & Communications Security, CCS '13*, pages 901–914, New York, NY, USA. Association for Computing Machinery.
- Anthony, D., Campos-Castillo, C., and Horne, C. (2017). Toward a Sociology of Privacy. *Annual Review of Sociology*, 43(1):249–269.
- Apple Inc. (2017). Apple differential privacy technical overview. https://www.apple.com/privacy/docs/Differential_Privacy_Overview.pdf (accessed: Feb 18, 2023).
- Apple’s Differential Privacy Team (2017). Learning with privacy at scale. *Apple Machine Learning Journal*, 1(8).
- Ashmead, R., Kifer, D., Leclerc, P., Machanavajjhala, A., and Sexton, W. (2019). Effective privacy after adjusting for invariants with applications to the 2020 Census. Technical report.
- Asi, H., Duchi, J. C., and Javidbakht, O. (2022). Element level differential privacy: The right granularity of privacy. In *AAAI Workshop on Privacy-Preserving Artificial Intelligence*. Association for the Advancement of Artificial Intelligence.
- Asi, H., Ullman, J., and Zakyntinou, L. (2023). From robustness to privacy and back. In *Proceedings of the 40th International Conference on Machine Learning*, pages 1121–1146. PMLR.
- Augustin, T., Coolen, F. P. A., de Cooman, G., and Troffaes, M. C. M., editors (2014). *Introduction to Imprecise Probabilities*. Wiley Series in Probability and Statistics. John Wiley & Sons, Ltd, Chichester, UK.
- Australian Bureau of Statistics (2021a). Five Safes framework. <https://www.abs.gov.au/about/data-services/data-confidentiality-guide/five-safes-framework>.
- Australian Bureau of Statistics (2021b). Treating microdata. <https://www.abs.gov.au/about/data-services/data-confidentiality-guide/treating-microdata>.
- Avella-Medina, M. (2020). The role of robust statistics in private data analysis. *CHANCE*, 33(4):37–42.
- Avella-Medina, M. (2021). Privacy-preserving parametric inference: A case for robust statistics. *Journal of the American Statistical Association*, 116(534):969–983.
- Awan, J. A. and Slavković, A. (2018). Differentially private uniformly most powerful tests for binomial data. In *Advances in Neural Information Processing Systems 31*, pages 4208–4218. Curran Associates,

Inc.

Awan, J. A. and Slavković, A. (2020). Differentially private inference for binomial data. *Journal of Privacy and Confidentiality*, 10(1).

Bailie, J. (2020). Big data, differential privacy and national statistical organisations. *Statistical Journal of the IAOS*, 36(4):1067–1074.

Bailie, J. and Chien, C.-H. (2019). ABS perturbation methodology through the lens of differential privacy. In *Work Session on Statistical Data Confidentiality, UN Economic Commission for Europe*, page 13.

Bailie, J. and Drechsler, J. (2024). Whose data is it anyway? Towards a formal treatment of differential privacy for surveys. In *Data Privacy Protection and the Conduct of Applied Research: Methods, Approaches and Their Consequences*, page 33. National Bureau of Economic Research.

Bailie, J. and Gong, R. (2023a). Differential privacy: General inferential limits via intervals of measures. In *Proceedings of the Thirteenth International Symposium on Imprecise Probability: Theories and Applications*, volume 215, pages 11–24. PMLR. <https://proceedings.mlr.press/v215/bailie23a.html>.

Bailie, J. and Gong, R. (2023b). The Five Safes as a privacy context. In *The 5th Annual Symposium on Applications of Contextual Integrity*, Toronto, Canada.

Bailie, J. and Gong, R. (2024). General inferential limits under differential and Pufferfish privacy. *International Journal of Approximate Reasoning*, 172. <https://www.sciencedirect.com/science/article/pii/S0888613X24001294>.

Bailie, J., Gong, R., and Meng, X.-L. (2025a). On the uniqueness of differential privacy. *In preparation*.

Bailie, J., Gong, R., and Meng, X.-L. (2025b). A refreshment stirred, not shaken (I): Building blocks of differential privacy. *In preparation*.

Bailie, J., Gong, R., and Meng, X.-L. (2025c). A refreshment stirred, not shaken (II): Invariant-preserving deployments of differential privacy for the US Decennial Census. *Under submission*.

Bailie, J., Gong, R., and Meng, X.-L. (2025d). A refreshment stirred, not shaken (III): Can swapping be differentially private? *In preparation*.

Bailie, J., Gong, R., and Meng, X.-L. (2025e). A statistical appreciation and assessment of differential privacy. *In preparation*.

- Balle, B., Barthe, G., and Gaboardi, M. (2018). Privacy amplification by subsampling: Tight analyses via couplings and divergences. *Advances in neural information processing systems*, 31.
- Balle, B., Barthe, G., and Gaboardi, M. (2020). Privacy profiles and amplification by subsampling. *Journal of Privacy and Confidentiality*, 10(1).
- Balle, B., Barthe, G., Gaboardi, M., Hsu, J., and Sato, T. (2019). Hypothesis testing interpretations and Rényi differential privacy.
- Barak, B., Chaudhuri, K., Dwork, C., Kale, S., McSherry, F., and Talwar, K. (2007). Privacy, accuracy, and consistency too: A holistic solution to contingency table release. In *Proceedings of the Twenty-Sixth ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, pages 273–282.
- Barber, R. F. and Duchi, J. C. (2014). Privacy and statistical risk: Formalisms and minimax bounds. <http://arxiv.org/abs/1412.4451>.
- Barth, A., Datta, A., Mitchell, J. C., and Nissenbaum, H. (2006). Privacy and contextual integrity. In *Proceedings of the 2006 IEEE Symposium on Security and Privacy*, pages 184–198, Berkeley, United States.
- Barthe, G. and Olmedo, F. (2013). Beyond differential privacy: Composition theorems and relational logic for f -divergences between probabilistic programs. In Fomin, F. V., Freivalds, R., Kwiatkowska, M., and Peleg, D., editors, *Automata, Languages, and Programming*, Lecture Notes in Computer Science, pages 49–60, Berlin, Heidelberg. Springer.
- Bassily, R., Groce, A., Katz, J., and Smith, A. (2013). Coupled-worlds privacy: Exploiting adversarial uncertainty in statistical data privacy. In *Proceedings of the 2013 IEEE 54th Annual Symposium on Foundations of Computer Science, FOCS '13*, pages 439–448, USA. IEEE Computer Society.
- Bassily, R., Smith, A., and Thakurta, A. (2014). Private empirical risk minimization: Efficient algorithms and tight error bounds. In *Proceedings of the 2014 IEEE 55th Annual Symposium on Foundations of Computer Science, FOCS '14*, pages 464–473, USA. IEEE Computer Society.
- Baumeister, R. F. and Tierney, J. (2011). *Willpower: Rediscovering the Greatest Human Strength*. Penguin.
- Beimel, A., Kasiviswanathan, S. P., and Nissim, K. (2010). Bounds on the sample complexity for private learning and private data release. In Micciancio, D., editor, *Proceedings of the 7th Theory of Cryptography Conference, TCC 2010, Zurich, Switzerland*, Lecture Notes in Computer Science, pages 437–454, Berlin, Heidelberg. Springer.
- Beimel, A., Nissim, K., and Omri, E. (2008). Distributed private data analysis: Simultaneously solving

- how and what. In Wagner, D., editor, *Advances in Cryptology – CRYPTO 2008*, volume 5157 of *Lecture Notes in Computer Science*, pages 451–468, Berlin, Heidelberg. Springer Berlin Heidelberg. http://link.springer.com/10.1007/978-3-540-85174-5_25.
- Benschop, T. and Welch, M. (2024). *Statistical Disclosure Control: A Practice Guide*. <https://sdcppractice.readthedocs.io/en/latest>.
- Berger, J. O. (1990). Robust Bayesian analysis: sensitivity to the prior. *Journal of Statistical Planning and Inference*, 25(3):303–328.
- Berger, J. O. and Wolpert, R. L. (1988). *The Likelihood Principle*, volume 6 of *Lecture Notes–Monograph Series*. Institute of Mathematical Statistics, Hayward, California, 2nd edition.
- Bernstein, G. and Sheldon, D. R. (2018). Differentially private Bayesian inference for exponential families. *Advances in Neural Information Processing Systems*, 31.
- Bernstein, G. and Sheldon, D. R. (2019). Differentially private Bayesian linear regression. *Advances in Neural Information Processing Systems*, 32.
- Bhaskar, R., Bhowmick, A., Goyal, V., Laxman, S., and Thakurta, A. (2011). Noiseless database privacy. In Lee, D. H. and Wang, X., editors, *Advances in Cryptology – ASIACRYPT 2011*, Lecture Notes in Computer Science, pages 215–232, Berlin, Heidelberg. Springer.
- Biddle, N., Gray, M., and Sollis, K. (2021). The use of QR codes to identify COVID-19 contacts and the role of data trust and data privacy. Technical report, Centre for Social Research and Methods, Australian National University.
- Billingsley, P. (2012). *Probability and Measure*. Wiley Series in Probability and Statistics. Wiley, Hoboken, N.J, anniversary ed (4th) edition.
- Blackwell, D. (1953). Equivalent comparisons of experiments. *The Annals of Mathematical Statistics*, 24(2):265–272.
- Bloustein, E. J. (1964). Privacy as an aspect of human dignity: An answer to Dean Prosser. *New York University Law Review*, 39:962–1007.
- Bok, S. (1982). *Secrets: On the Ethics of Concealment and Revelation*. Pantheon, New York.
- Borgman, C. L. (2019). The lives and after lives of data. *Harvard Data Science Review*, 1(1).
- Boswell, C. (2020). What is politics?

- Bowen, C. M., Bryant, V., Burman, L., Czajka, J., Khitatrakun, S., MacDonald, G., McClelland, R., Muciolio, L., Pickens, M., Ueyama, K., et al. (2022). Synthetic individual income tax data: Methodology, utility, and privacy implications. In *Privacy in Statistical Databases: International Conference, PSD 2022, Paris, France, September 21–23, 2022, Proceedings*, pages 191–204. Springer.
- Bowker, G. C. (2005). *Memory Practices in the Sciences*. Inside Technology. MIT Press, Cambridge, MA, USA, 1st edition.
- Bowles, C. (2018). *Future Ethics*. NowNext Press.
- boyd, d. and Sarathy, J. (2022). Differential perspectives: Epistemic disconnects surrounding the U.S. Census Bureau’s use of differential privacy. *Harvard Data Science Review*, (Special Issue 2).
- Brenton, M. (1964). *The Privacy Invaders*. Coward-McCann, New York.
- Brick, J. M. and Williams, D. (2013). Explaining rising nonresponse rates in cross-sectional surveys. *The ANNALS of the American Academy of Political and Social Science*, 645(1):36–59.
- Bruce, R. A. and McDonough, J. R. (1969). Stress testing in screening for cardiovascular disease. *Bulletin of the New York Academy of Medicine*, 45(12):1288–1305. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1750554/>.
- Bu, Z., Dong, J., Long, Q., and Weijie, S. (2020). Deep learning with Gaussian differential privacy. *Harvard Data Science Review*, 2(3).
- Bun, M., Drechsler, J., Gaboardi, M., McMillan, A., and Sarathy, J. (2022). Controlling privacy loss in sampling schemes: An analysis of stratified and cluster sampling. In *Foundations of Responsible Computing (FORC 2022)*, page 24.
- Bun, M., Nissim, K., Stemmer, U., and Vadhan, S. (2015). Differentially private release and learning of threshold functions. In *Proceedings of the 56th Annual IEEE Symposium on Foundations of Computer Science*, FOCS ’15, pages 634–649, Washington, DC, USA. IEEE Computer Society.
- Bun, M. and Steinke, T. (2016). Concentrated differential privacy: Simplifications, extensions, and lower bounds. In Hirt, M. and Smith, A., editors, *Theory of Cryptography*, Lecture Notes in Computer Science, pages 635–658, Berlin, Heidelberg. Springer.
- Cai, T. T., Wang, Y., and Zhang, L. (2021). The cost of privacy: Optimal rates of convergence for parameter estimation with differential privacy. *The Annals of Statistics*, 49(5):2825–2850.
- Canonne, C., Kamath, G., and Steinke, T. (2022). The discrete Gaussian for differential privacy. *Journal*

of *Privacy and Confidentiality*, 12(1).

Cantwell, P. (2021). How we complete the Census when households or group quarters don't respond. <https://www.census.gov/newsroom/blogs/random-samplings/2021/04/imputation-when-households-or-group-quarters-dont-respond.html>.

Carothers, N. L. (2000). *Real Analysis*. Cambridge University Press, Cambridge, UK.

Carroll, R. J. and Hall, P. (1988). Optimal rates of convergence for deconvolving a density. *Journal of the American Statistical Association*, 83(404):1184–1186.

Casacuberta, S., Shoemate, M., Vadhan, S., and Wagaman, C. (2022). Widespread underestimation of sensitivity in differentially private libraries and how to fix it. In *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security*, CCS'22, pages 471–484, New York, NY, USA. Association for Computing Machinery.

Chang, J. T. and Pollard, D. (1997). Conditioning as disintegration. *Statistica Neerlandica*, 51(3):287–317.

Charest, A.-S. and Hou, Y. (2016). On the meaning and limits of empirical differential privacy. *Journal of Privacy and Confidentiality*, 7(3):53–66.

Chatzikokolakis, K., Andrés, M. E., Bordenabe, N. E., and Palamidessi, C. (2013). Broadening the Scope of Differential Privacy Using Metrics. In De Cristofaro, E. and Wright, M., editors, *Privacy Enhancing Technologies*, Lecture Notes in Computer Science, pages 82–102, Berlin, Heidelberg. Springer.

Cheu, A. (2022). Differential privacy in the shuffle model: A survey of separations. <http://arxiv.org/abs/2107.11839>.

Cheu, A., Smith, A., Ullman, J., Zeber, D., and Zhilyaev, M. (2019). Distributed differential privacy via shuffling. In Ishai, Y. and Rijmen, V., editors, *Advances in Cryptology – EUROCRYPT 2019*, Lecture Notes in Computer Science, pages 375–403, Cham. Springer International Publishing.

Chhor, J. and Sentenac, F. (2023). Robust estimation of discrete distributions under local differential privacy. In *International Conference on Algorithmic Learning Theory*, pages 411–446. PMLR.

Chien, C.-H. and Sadeghi, P. (2024). On the connection between the ABS perturbation methodology and differential privacy. *Journal of Privacy and Confidentiality*, 14(2).

Chipperfield, J., Gow, D., and Loong, B. (2016). The Australian Bureau of Statistics and releasing frequency tables via a remote server. *Statistical Journal of the LAOS*, 32(1):53–64.

- Cho, Y. H. and Awan, J. (2024). Formal privacy guarantees with invariant statistics. <http://arxiv.org/abs/2410.17468>.
- Christ, M., Radway, S., and Bellovin, S. M. (2022). Differential privacy and swapping: Examining deidentification’s impact on minority representation and privacy preservation in the U.S. Census. In *2022 IEEE Symposium on Security and Privacy*, pages 457–472.
- Clifton, C., Dajani, A. N., Hanson, E. J., Clark, S., Merrill, K., Merrill, S., and Rodriguez, R. (2023). Preliminary report on differentially private post-stratification. Working paper ced-wp-2023-004, US Census Bureau.
- Cohen, A. (2022). Attacks on deidentification’s defenses. In *31st USENIX Security Symposium (USENIX Security 22)*, pages 1469–1486. <https://www.usenix.org/conference/usenixsecurity22/presentation/cohen>.
- Couper, M. P., Singer, E., Conrad, F. G., and Groves, R. M. (2008). Risk of Disclosure, Perceptions of Risk, and Concerns about Privacy and Confidentiality as Factors in Survey Participation. *Journal of official statistics*, 24(2):255–275.
- Couper, M. P., Singer, E., Conrad, F. G., and Groves, R. M. (2010). Experimental Studies of Disclosure Risk, Disclosure Harm, Topic Sensitivity, and Survey Participation. *Journal of official statistics*, 26(2):287–300.
- Cover, T. M. and Thomas, J. A. (2005). *Elements of Information Theory*. Wiley, 1 edition. <https://onlinelibrary.wiley.com/doi/book/10.1002/047174882X>.
- Cuff, P. and Yu, L. (2016). Differential privacy as a mutual information constraint. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, pages 43–54.
- Culnan, M. J. and Armstrong, P. K. (1999). Information Privacy Concerns, Procedural Fairness, and Impersonal Trust: An Empirical Investigation. *Organization Science*, 10(1):104–115.
- Culnane, C., Rubinstein, B. I. P., and Teague, V. (2019). Stop the Open Data Bus, We Want to Get Off. *arXiv:1908.05004 [cs]*.
- Cummings, R., Desfontaines, D., Evans, D., Geambasu, R., Huang, Y., Jagielski, M., Kairouz, P., Kamath, G., Oh, S., Ohrimenko, O., Papernot, N., Rogers, R., Shen, M., Song, S., Su, W., Terzis, A., Thakurta, A., Vassilvitskii, S., Wang, Y.-X., Xiong, L., Yekhanin, S., Yu, D., Zhang, H., and Zhang, W. (2024). Advancing differential privacy: Where we are now and future directions for real-world deployment. *Harvard Data Science Review*, 6(1).

- Dalenius, T. and Reiss, S. P. (1978). Data-swapping: A technique for disclosure control (extended abstract). In *Proceedings of the ASA Section on Survey Research Methods*, volume 6, pages 191–194, Washington, DC. American Statistical Association. <https://www.sciencedirect.com/science/article/pii/0378375882900581>.
- Dalenius, T. and Reiss, S. P. (1982). Data-swapping: A technique for disclosure control. *Journal of Statistical Planning and Inference*, 6(1):73–85.
- Das, S., Drechsler, J., Merrill, K., and Merrill, S. (2022). Imputation under differential privacy. <http://arxiv.org/abs/2206.15063>.
- Data Quality Hub (2020). Monitoring and reducing respondent burden. Technical report, UK Government Statistical Service.
- de Campos, L. M., Huete, J. F., and Moral, S. (1994). Probability intervals: A tool for uncertain reasoning. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 2(2):167–196.
- de Vries, M., Golmajer, M., Tent, R., Giessing, S., and de Wolf, P.-P. (2023). An overview of used methods to protect the European Census 2021 tables. In *UNECE Conference of European Statisticians Expert Meeting on Statistical Data Confidentiality*, page 16, Wiesbaden. <https://unece.org/statistics/documents/2023/08/working-documents/overview-used-methods-protect-european-census-2021>.
- Debenedetti, E., Severi, G., Carlini, N., Choquette-Choo, C. A., Jagielski, M., Nasr, M., Wallace, E., and Tramèr, F. (2024). Privacy side channels in machine learning systems. <http://arxiv.org/abs/2309.05610>.
- DeCew, J. (2018). Privacy. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, spring 2018 edition. <https://plato.stanford.edu/archives/spr2018/entries/privacy/>.
- DeCew, J. W. (1997). *In Pursuit of Privacy: Law, Ethics, and the Rise of Technology*. Cornell University Press.
- Demmers, Joris (2018). *Consumers and Their Data*. PhD thesis, University of Amsterdam, Amsterdam.
- DePersio, M., Lemons, M., Ramanayake, K. A., Tsay, J., and Zayatz, L. (2012). n-cycle swapping for the American Community Survey. In Domingo-Ferrer, J. and Tinnirello, I., editors, *PSD 2012: Privacy in Statistical Databases*, volume 7556 of *Lecture Notes in Computer Science*, pages 143–164, Berlin, Heidelberg. Springer. <https://link-springer-com.ezp-prod1.hul.harvard.edu/chapter/10>.

1007/978-3-642-33627-0_12.

- DeRobertis, L. and Hartigan, J. A. (1981). Bayesian inference using intervals of measures. *The Annals of Statistics*, 9(2):235–244.
- Desai, T., Ritchie, F., and Welpton, R. (2016). Five Safes: Designing data access for research. Working paper 1601, University of the West of England, Bristol.
- Desfontaines, D. (2023). A list of real-world uses of differential privacy. Blog post on *Ted Is Writing Things*. <https://desfontain.es/privacy/real-world-differential-privacy.html>.
- Desfontaines, D., Mohammadi, E., Krahmer, E., and Basin, D. (2020). Differential privacy with partial knowledge. <http://arxiv.org/abs/1905.00650>.
- Desfontaines, D. and Pejó, B. (2022). SoK: Differential privacies.
- Desfontaines, D. and Pejó, B. (2020). SoK: Differential privacies. In *Proceedings on Privacy Enhancing Technologies*, volume 2020, pages 288–313.
- Destercke, S., Montes, I., and Miranda, E. (2022). Processing distortion models: A comparative study. *International Journal of Approximate Reasoning*, 145:91–120.
- Dharangutte, P., Gao, J., Gong, R., and Yu, F.-Y. (2023). Integer subspace differential privacy. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI-23)*.
- Dinev, T. and Hart, P. (2006). An Extended Privacy Calculus Model for E-Commerce Transactions. *Information Systems Research*, 17(1):61–80.
- Dinev, T., McConnell, A. R., and Smith, H. J. (2015). Research Commentary—Informing Privacy Research Through Information Systems, Psychology, and Behavioral Economics: Thinking Outside the “APCO” Box. *Information Systems Research*, 26(4):639–655.
- Ding, B., Kulkarni, J., and Yekhanin, S. (2017). Collecting telemetry data privately. In *Advances in Neural Information Processing Systems*, pages 3571–3580.
- Ding, J. and Ding, B. (2022). Interval privacy: A framework for privacy-preserving data collection. *IEEE Transactions on Signal Processing*, 70:2443–2459.
- Ding, N. (2024). Approximation of Pufferfish privacy for Gaussian priors. <http://arxiv.org/abs/2401.12391>.

- Dinur, I. and Nissim, K. (2003). Revealing information while preserving privacy. In *Proceedings of the Twenty-Second ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems - PODS '03*, pages 202–210, San Diego, California. ACM Press.
- Dong, J., Roth, A., and Su, W. J. (2022). Gaussian differential privacy. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 84(1):3–37.
- Drechsler, J. (2023). Differential privacy for government agencies—Are we there yet? *Journal of the American Statistical Association*, 118(541):761–773. <https://doi.org/10.1080/01621459.2022.2161385>.
- Drechsler, J. and Bailie, J. (2024). The complexities of differential privacy for survey data. <http://arxiv.org/abs/2408.07006>.
- Drechsler, J. and Reiter, J. P. (2010). Sampling with synthesis: A new approach for releasing public use census microdata. *Journal of the American Statistical Association*, 105(492):1347–1357.
- Duchi, J. C., Jordan, M. I., and Wainwright, M. J. (2018). Minimax optimal procedures for locally private estimation. *Journal of the American Statistical Association*, 113(521):182–201.
- Duhigg, C. (2012). How companies learn your secrets. *The New York Times*.
- Duncan, G. T. and Lambert, D. (1986). Disclosure-limited data dissemination. *Journal of the American Statistical Association*, 81(393):10–18.
- Durrett, R. (2019). *Probability: Theory and Examples*. Cambridge University Press, 5 edition.
- Dwork, C. (2006). Differential privacy. In *International colloquium on automata, languages, and programming*, pages 1–12. Springer.
- Dwork, C., Kenthapadi, K., McSherry, F., Mironov, I., and Naor, M. (2006a). Our data, ourselves: Privacy via distributed noise generation. In Vaudenay, S., editor, *Advances in Cryptology - EUROCRYPT 2006*, Lecture Notes in Computer Science, pages 486–503, Berlin, Heidelberg. Springer.
- Dwork, C., Kohli, N., and Mulligan, D. (2019). Differential privacy in practice: Expose your epsilons! *Journal of Privacy and Confidentiality*, 9(2).
- Dwork, C. and Lei, J. (2009). Differential privacy and robust statistics. In *Proceedings of the forty-first annual ACM symposium on Theory of computing*, pages 371–380. ACM.
- Dwork, C., McSherry, F., Nissim, K., and Smith, A. (2006b). Calibrating noise to sensitivity in private

- data analysis. In *Theory of cryptography conference*, pages 265–284. Springer.
- Dwork, C., McSherry, F., Nissim, K., and Smith, A. (2016). Calibrating noise to sensitivity in private data analysis. *Journal of Privacy and Confidentiality*, 7(3):17–51.
- Dwork, C. and Naor, M. (2010). On the difficulties of disclosure prevention in statistical databases or the case for differential privacy. *Journal of Privacy and Confidentiality*, 2(1).
- Dwork, C., Naor, M., Pitassi, T., and Rothblum, G. N. (2010a). Differential privacy under continual observation. In *Proceedings of the Forty-Second ACM Symposium on Theory of Computing*, STOC ’10, pages 715–724, New York, NY, USA. Association for Computing Machinery. <https://dl.acm.org/doi/10.1145/1806689.1806787>.
- Dwork, C. and Roth, A. (2014). The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3–4):211–407.
- Dwork, C. and Rothblum, G. N. (2016). Concentrated differential privacy. *arXiv:1603.01887*.
- Dwork, C., Rothblum, G. N., and Vadhan, S. (2010b). Boosting and differential privacy. In *2010 IEEE 51st Annual Symposium on Foundations of Computer Science*, pages 51–60, Las Vegas, NV, USA. IEEE.
- Dwork, C., Smith, A., Steinke, T., and Ullman, J. (2017). Exposed! A survey of attacks on private data. *Annual Review of Statistics and Its Application*, 4(1):61–84.
- Dwork, C., Talwar, K., Thakurta, A., and Zhang, L. (2014). Analyze Gauss: Optimal bounds for privacy-preserving principal component analysis. In *Proceedings of the Forty-Sixth Annual ACM Symposium on Theory of Computing*, STOC ’14, pages 11–20, New York, NY, USA. Association for Computing Machinery.
- Ebadi, H., Sands, D., and Schneider, G. (2015). Differential Privacy: Now it’s Getting Personal. *ACM SIGPLAN Notices*, 50(1):69–81.
- Erlingsson, Ú., Feldman, V., Mironov, I., Raghunathan, A., Talwar, K., and Thakurta, A. (2019). Amplification by shuffling: From local to central differential privacy via anonymity. In *Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA ’19, pages 2468–2479, USA. Society for Industrial and Applied Mathematics.
- Erlingsson, Ú., Pihur, V., and Korolova, A. (2014). RAPPOR: Randomized aggregatable privacy-preserving ordinal response. In *Proceedings of the 2014 ACM SIGSAC conference on computer and communications security*, pages 1054–1067.

- Evans, M. and Moshonov, H. (2006). Checking for prior-data conflict. *Bayesian Analysis*, 1(4):893–914.
- Feldman, V., McMillan, A., and Talwar, K. (2022). Hiding among the clones: A simple and nearly optimal analysis of privacy amplification by shuffling. In *2021 IEEE 62nd Annual Symposium on Foundations of Computer Science (FOCS)*, pages 954–964.
- Feldman, V. and Zrnic, T. (2022). Individual privacy accounting via a Rényi filter. <http://arxiv.org/abs/2008.11193>.
- Fienberg, S. and McIntyre, J. (2004). Data swapping: Variations on a theme by Dalenius and Reiss. In *Privacy in Statistical Databases*.
- Finck, M. and Pallas, F. (2020). They who must not be identified—distinguishing personal from non-personal data under the GDPR. *International Data Privacy Law*, 10(1):11–36. <https://doi.org/10.1093/idpl/ipz026>.
- Flood, M., Katz, J., Ong, S., and Smith, A. (2013). Cryptography and the Economics of Supervisory Information: Balancing Transparency and Confidentiality. Working Paper #0011 #0011, Office of Financial Research, US Department of Treasury.
- Foote, A. D., Machanavajjhala, A., and McKinney, K. (2019). Releasing earnings distributions using differential privacy: Disclosure avoidance system for post-secondary employment outcomes (pseo). *Journal of Privacy and Confidentiality*, 9(2).
- Foster, I., Ghani, R., Jarmin, R. S., Kreuter, F., and Lane, J., editors (2021). *Big Data and Social Science: Data Science Methods and Tools for Research and Practice*. Chapman and Hall/CRC Statistics in the Social and Behavioural Sciences. CRC Press, Boca Raton, FL, second edition edition.
- Francis, L. and Francis, J. G. (2017). *Privacy: What Everyone Needs to Know*. Oxford University Press.
- Francis, P. (2022). A note on the misinterpretation of the US Census re-identification attack. <http://arxiv.org/abs/2202.04872>.
- Fraser, B. and Wooton, J. (2005). A proposed method for confidentialising tabular output to protect against differencing. In *Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality*, Geneva, Switzerland.
- Ganta, S. R., Kasiviswanathan, S. P., and Smith, A. (2008). Composition attacks and auxiliary information in data privacy. In *Proceeding of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD 08*, page 265, Las Vegas, Nevada, USA. ACM Press.

- Gao, J., Gong, R., and Yu, F.-Y. (2022). Subspace differential privacy. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(4):3986–3995.
- Garfinkel, S. (2001). *Database Nation: The Death of Privacy in the 21st Century*. O'Reilly, Cambridge, Mass.
- Garfinkel, S. (2023). Comment to Muralidhar and Domingo-Ferrer (2023) – legacy statistical disclosure limitation techniques were not an option for the 2020 US Census of Population And Housing. *Journal of Official Statistics*, 39(3):399–410. <https://sciendo.com/it/article/10.2478/jos-2023-0018>.
- Garfinkel, S. L. (2019). Formal privacy: Making an impact at large organizations - Deploying differential privacy for the 2020 Census of Population and Housing.
- Gavison, R. (1980). Privacy and the Limits of Law. *The Yale Law Journal*, 89(3):421–471.
- Ghosh, A. and Roth, A. (2015). Selling privacy at auction. *Games and Economic Behavior*, 91:334–346.
- Gitelman, L., editor (2013). *"Raw Data" Is an Oxymoron*. Infrastructures Series. The MIT Press, Cambridge, Massachusetts ; London, England, 1st edition. <https://direct.mit.edu/books/edited-volume/3992/Raw-Data-Is-an-Oxymoron>.
- Glessing, S. and Schulte Nordholt, E. (2017). Recommendations for best practices to protect the Census 2021 hypercubes. Technical report, Centre of Excellence on Statistical Disclosure Control, Eurostat. https://cros-legacy.ec.europa.eu/content/recommendations-protection-hypercubes_en.
- Goldwasser, S. and Micali, S. (1984). Probabilistic encryption. *Journal of Computer and System Sciences*, 28(2):270–299.
- Gong, R. (2022a). Exact inference with approximate computation for differentially private data via perturbations. *Journal of Privacy and Confidentiality*, 12(2).
- Gong, R. (2022b). Transparent privacy is principled privacy. *Harvard Data Science Review*, (Special Issue 2).
- Gong, R., Groshen, E. L., and Vadhan, S. (2022). Differential privacy for the 2020 US Census: Can we make data both private and useful? *Harvard Data Science Review*, (Special Issue 2).
- Gong, R. and Meng, X.-L. (2020). Congenial differential privacy under mandated disclosure. In *Proceedings of the 2020 ACM-IMS on Foundations of Data Science Conference*, FODS '20, pages 59–70, New

- York, NY, USA. Association for Computing Machinery.
- Gross, H. (1971). Privacy and Autonomy. In Pennock, J. and Chapman, J., editors, *NOMOS XIII: Privacy*, pages 169–181. Atherton Press, New York.
- Hacking, I. (1965). *Logic of Statistical Inference*. Cambridge University Press, Cambridge.
- Hall, R., Rinaldo, A., and Wasserman, L. (2013). Differential privacy for functions and functional data. *Journal of machine learning research*, 14(Feb):703–727.
- Haney, S., Machanavajjhala, A., Abowd, J. M., Graham, M., Kutzbach, M., and Vilhuber, L. (2017). Utility cost of formal privacy for releasing national employer-employee statistics. In *Proceedings of the 2017 ACM International Conference on Management of Data - SIGMOD '17*, pages 1339–1354, Chicago, Illinois, USA. ACM Press.
- Hannig, J., Iyer, H., Lai, R. C., and Lee, T. C. (2016). Generalized fiducial inference: A review and new results. *Journal of the American Statistical Association*, 111(515):1346–1361.
- Hawala, S. (2008). Producing partially synthetic data to avoid disclosure. In *JSM Section on Government Statistics*, pages 1345–1350. American Statistical Association.
- Hawes, M. (2021a). The 2020 Census Disclosure Avoidance System. https://planning.maryland.gov/MSDC/Documents/affiliate_meeting/2021/Census2021_MHawes.pdf.
- Hawes, M. (2021b). Understanding the 2020 Census Disclosure Avoidance System. Presentation. <https://www2.census.gov/about/training-workshops/2021/2021-08-10-das-presentation.pdf>.
- Hawes, M. and Rodríguez, R. (2021). Determining the Privacy-loss Budget: Research into Alternatives to Differential Privacy. Presentation to the Census Scientific Advisory Committee. <https://www2.census.gov/about/partners/cac/sac/meetings/2021-05/presentation-research-on-alternatives-to-differential-privacy.pdf>.
- Hawes, M., Rodriguez, R., and Goldschlag, N. (2021). The Census Bureau’s simulated reconstruction-abetted re-identification attack on the 2010 Census. [transcript of presentation at NWX-US Dept Of Commerce].
- Hay, M., Li, C., Miklau, G., and Jensen, D. (2009). Accurate estimation of the degree distribution of private networks. In *2009 Ninth IEEE International Conference on Data Mining*, pages 169–178.
- Hay, M., Rastogi, V., Miklau, G., and Suciu, D. (2010). Boosting the accuracy of differentially private

- histograms through consistency. *Proceedings of the VLDB Endowment*, 3(1).
- He, X., Machanavajjhala, A., and Ding, B. (2014). Blowfish privacy: Tuning privacy-utility trade-offs using policies. In *Proceedings of the 2014 ACM SIGMOD international conference on Management of data*, pages 1447–1458.
- Heffetz, O. (2022). What will it take to get to acceptable privacy-accuracy combinations? *Harvard Data Science Review*, (Special Issue 2).
- Heitjan, D. F. and Rubin, D. B. (1990). Inference from coarse data via multiple imputation with application to age heaping. *Journal of the American Statistical Association*, 85(410):304–314.
- Henkin, L. (1974). Privacy and Autonomy. *Columbia Law Review*, 74(8):1410–1433.
- Heywood, A. (2013). *Politics*. Palgrave Foundations. Macmillan Education - Palgrave, Oxford, 4. ed edition.
- Hod, S. and Canetti, R. (2025). Differentially private release of Israel’s National Registry of Live Births. In *Proceedings of the 2025 IEEE Symposium on Security and Privacy (SP)*. IEEE Computer Society.
- Holzberg, J., Katz, J., and Davis, M. (2021). Measuring Respondents’ Perceptions of Burden in the American Community Survey. Technical Report 2021-04, Center for Behavioral Science Methods, Research and Methodology Directorate, US Census Bureau.
- Homer, N., Szelingier, S., Redman, M., Duggan, D., Tembe, W., Muehling, J., Pearson, J. V., Stephan, D. A., Nelson, S. F., and Craig, D. W. (2008). Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. *PLoS genetics*, 4(8):e1000167.
- Hopkins, S. B., Kamath, G., Majid, M., and Narayanan, S. (2023). Robustness implies privacy in statistical estimation. In *Proceedings of the 55th Annual ACM Symposium on Theory of Computing*, pages 497–506, Orlando FL USA. ACM.
- Horowitz, J. L. and Manski, C. F. (1995). Identification and robustness with contaminated and corrupted data. *Econometrica*, 63(2):281–302.
- Horvitz, D. G. and Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American statistical Association*, 47(260):663–685.
- Hotz, V. J., Bollinger, C. R., Komarova, T., Manski, C. F., Moffitt, R. A., Nekipelov, D., Sojourner, A., and Spencer, B. D. (2022). Balancing data privacy and usability in the federal statistical system. *Proceedings*

of the *National Academy of Sciences*, 119(31):e2104906119.

Hotz, V. J. and Salvo, J. (2020). Assessing the use of differential privacy for the 2020 Census: Summary of what we learned from the CNSTAT workshop. Technical report, National Academies Committee on National Statistics. https://www.amstat.org/asa/files/pdfs/POL-CNSTAT_CensusDP_WorkshopLessonsLearnedSummary.pdf [Accessed: 04-08-2020].

Hotz, V. J. and Salvo, J. (2022). A chronicle of the application of differential privacy to the 2020 Census. *Harvard Data Science Review*, (Special Issue 2).

Hsu, J., Gaboardi, M., Haeberlen, A., Khanna, S., Narayan, A., Pierce, B. C., and Roth, A. (2014). Differential Privacy: An economic method for choosing epsilon. In *Proceedings of the 2014 IEEE 27th Computer Security Foundations Symposium*, pages 398–410, Vienna. IEEE.

Hu, Y., Sanyal, A., and Schölkopf, B. (2024). Provable privacy with non-private pre-processing. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *ICML'24*, pages 19402–19437. JMLR.org.

Inness, J. (1992). *Privacy, Intimacy, and Isolation*. Oxford University Press.

Ito, S. and Hoshino, N. (2014). Data swapping as a more efficient tool to create anonymized Census microdata in Japan. In *Privacy in Statistical Databases*. <https://www.semanticscholar.org/paper/Data-Swapping-as-a-More-Efficient-Tool-to-Create-in-Ito-Hoshino/8786e94ab1b714c6e7536b6dd2fbb19a0c0bdaad>.

Jacob, P. E., Gong, R., Edlefsen, P. T., and Dempster, A. P. (2021). A Gibbs sampler for a class of random convex polytopes. *Journal of the American Statistical Association*, 116(535):1181–1192. <https://doi.org/10.1080/01621459.2021.1881523>.

Jarmin, R. S., Abowd, J. M., Ashmead, R., Cumings-Menon, R., Goldschlag, N., Hawes, M. B., Keller, S. A., Kifer, D., Leclerc, P., Reiter, J. P., Rodríguez, R. A., Schmutte, I., Velkoff, V. A., and Zhuravlev, P. (2023). An in-depth examination of requirements for disclosure risk assessment. *Proceedings of the National Academy of Sciences*, 120(43):e2220558120.

Jorgensen, Z., Yu, T., and Cormode, G. (2015). Conservative or liberal? Personalized differential privacy. In *2015 IEEE 31st International Conference on Data Engineering*, pages 1023–1034. <https://ieeexplore.ieee.org/document/7113353>.

Ju, N., Awan, J. A., Gong, R., and Rao, V. A. (2022). Data augmentation MCMC for Bayesian inference from privatized data. *Thirty-sixth Annual Conference on Neural Information Processing Systems*.

- Kahneman, D. (2011). *Thinking, Fast and Slow*. Farrar, Straus and Giroux.
- Kahneman, D. and Tversky, A. (1979). Prospect Theory: An Analysis of Decision under Risk. *Econometrica*, 47(2):263–291.
- Kairouz, P., Oh, S., and Viswanath, P. (2017). The composition theorem for differential privacy. *IEEE Transactions on Information Theory*, 63(6):4037–4049.
- Kalven, H. (1966). Privacy in Tort Law—Were Warren and Brandeis Wrong? *Law and Contemporary Problems*, 31(2):326–341.
- Kamath, G., Mouzakis, A., Regehr, M., Singhal, V., Steinke, T., and Ullman, J. (2023). A bias-variance-privacy trilemma for statistical estimation. *arXiv preprint arXiv:2301.13334*.
- Kasiviswanathan, S. P., Lee, H. K., Nissim, K., Raskhodnikova, S., and Smith, A. (2011). What can we learn privately? *SIAM Journal on Computing*, 40(3):793–826.
- Kasiviswanathan, S. P. and Smith, A. (2014). On the ‘semantics’ of differential privacy: A Bayesian formulation. *Journal of Privacy and Confidentiality*, 6(1).
- Kasper, D. V. (2007). Privacy as a Social Good. *Social Thought & Research*, 28:165–189.
- Keller, S. A. and Abowd, J. M. (2023). Database reconstruction does compromise confidentiality. *Proceedings of the National Academy of Sciences*, 120(12):e2300976120. <https://www.pnas.org/doi/10.1073/pnas.2300976120>.
- Kenny, C. T., Kuriwaki, S., McCartan, C., Rosenman, E. T., Simko, T., and Imai, K. (2021). The use of differential privacy for Census data and its impact on redistricting: The case of the 2020 US Census. *Science Advances*, 7(41):eabk3283.
- Kenny, C. T., McCartan, C., Kuriwaki, S., Simko, T., and Imai, K. (2024). Evaluating bias and noise induced by the U.S. Census Bureau’s privacy protection methods. *Science Advances*, 10(18):eadl2524.
- Kenthapadi, K. and Tran, T. T. L. (2018). PriPeARL: A framework for privacy-preserving analytics and reporting at LinkedIn. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, CIKM ’18, pages 2183–2191, New York, NY, USA. Association for Computing Machinery.
- Keulen, S. and Kroeze, R. (2018). Privacy from a Historical Perspective. In van der Sloot, B. and de Groot, A., editors, *The Handbook of Privacy Studies: An Interdisciplinary Introduction*. Amsterdam University Press.

- Keyes, O. and Flaxman, A. D. (2022). How Census data put trans children at risk. *Scientific American*. <https://www.scientificamerican.com/article/how-census-data-put-trans-children-at-risk/>.
- Kifer, D. (2019). Design principles of the TopDown algorithm. Presentation at JASON, La Jolla, CA.
- Kifer, D., Abowd, J. M., Ashmead, R., Cumings-Menon, R., Leclerc, P., Machanavajjhala, A., Sexton, W., and Zhuravlev, P. (2022). Bayesian and frequentist semantics for common variations of differential privacy: Applications to the 2020 Census.
- Kifer, D. and Machanavajjhala, A. (2011). No free lunch in data privacy. In *Proceedings of the 2011 International Conference on Management of Data - SIGMOD '11*, pages 193–204, Athens, Greece. ACM Press.
- Kifer, D. and Machanavajjhala, A. (2014). Pufferfish: A framework for mathematical privacy definitions. *ACM Transactions on Database Systems*, 39(1):1–36.
- Kim, N. (2015). The effect of data swapping on analyses of American Community Survey data. *Journal of Privacy and Confidentiality*, 7(1). <https://journalprivacyconfidentiality.org/index.php/jpc/article/view/644>.
- Kim, S. and Misra, A. (2007). SNP genotyping: Technologies and biomedical applications. *Annual Review of Biomedical Engineering*, 9(1):289–320.
- Kirchner, A. (2015). Validating Sensitive Questions: A Comparison of Survey and Register Data. *Journal of Official Statistics*, 31(1):31–59.
- Kitchin, R. (2015). The opportunities, challenges and risks of big data for official statistics. *Statistical Journal of the IAOS*, 31(3):471–481.
- Komarova, T. and Nekipelov, D. (2020). Identification and formal privacy guarantees. *arXiv preprint arXiv:2006.14732*.
- Kotsogiannis, I., Doudalis, S., Haney, S., Machanavajjhala, A., and Mehrotra, S. (2020). One-sided differential privacy. In *2020 IEEE 36th International Conference on Data Engineering (ICDE)*, pages 493–504.
- Kreuter, F. (2019). The social survey statistician’s perspective. Presentation at the Simons Institute, Berkeley.
- Laird, A. (2001). Ringing the changes on Gyges: Philosophy and the formation of fiction in Plato’s *Re-*

- public. The Journal of Hellenic Studies*, 121:12–29.
- Landsheer, J. A., Van Der Heijden, P., and Van Gils, G. (1999). Trust and Understanding, Two Psychological Aspects of Randomized Response. *Quality and Quantity*, 33(1):1–12.
- Lateral Economics (2019). Value of the Australian Census. Technical report, Australian Bureau of Statistics.
- Laufer, R. and Wolfe, M. (1977). Privacy as a Concept and a Social Issue: A Multidimensional Developmental Theory. *Journal of Social Issues*, 33(3):22–42.
- Lauger, A., Wisniewski, B., and McKenna, L. (2014). Disclosure avoidance techniques at the U.S. Census Bureau: Current practices and research. Research Report Series - Disclosure Avoidance #2014-02, Center for Disclosure Avoidance Research, US Census Bureau, Washington DC.
- Lavine, M. (1991a). An approach to robust Bayesian analysis for multidimensional parameter spaces. *Journal of the American Statistical Association*, 86(414):400–403.
- Lavine, M. (1991b). Sensitivity in Bayesian statistics: The prior and the likelihood. *Journal of the American Statistical Association*, 86(414):396–399.
- Leftwich, A., editor (2004). *What Is Politics? The Activity and Its Study*. Polity, Cambridge.
- Lemons, M., Dajani, A., You, J., and Jordan, J. (2015). Measuring the degree of difference in perturbed data. In *Proceedings of the 2015 Joint Statistical Meetings*, Seattle, WA. American Statistical Association.
- Leonelli, S. (2019). Data governance is key to interpretation: Reconceptualizing data in data science. *Harvard Data Science Review*, 1(1).
- Leonelli, S. and Tempini, N., editors (2020). *Data Journeys in the Sciences*. Springer International Publishing, Cham.
- Lever, A. (2015). Privacy, Democracy and Freedom of Expression. In Roessler, B. and Mokrosinska, D., editors, *Social Dimensions of Privacy: Interdisciplinary Perspectives*. Cambridge University Press.
- Levi, I. (1980). *The Enterprise of Knowledge: An Essay on Knowledge, Credal Probability, and Chance*. MIT Press.
- Li, C., Li, D. Y., Miklau, G., and Suci, D. (2017). A theory of pricing private data. *Communications of the ACM*, 60(12):79–86.

- Li, Q., Zhou, C., Qin, B., and Xu, Z. (2022). Local differential privacy for belief functions. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(9):10025–10033.
- Ligett, K. and Roth, A. (2012). Take It or Leave It: Running a Survey When Privacy Comes at a Cost. In Goldberg, P. W., editor, *Internet and Network Economics*, Lecture Notes in Computer Science, pages 378–391, Berlin, Heidelberg, Springer.
- Lillington, K. (2021). Britain looks to weaken rules on data privacy. *The Irish Times*.
- Lin, S., Bun, M., Gaboardi, M., Kolaczyk, E. D., and Smith, A. (2023). Differentially private confidence intervals for proportions under stratified random sampling. *arXiv preprint arXiv:2301.08324*.
- Lindgreen, E. R. (2018). Privacy from an Economic Perspective. In van der Sloot, B. and de Groot, A., editors, *The Handbook of Privacy Studies: An Interdisciplinary Introduction*, pages 181–208. Amsterdam University Press.
- Liu, L., Sun, K., Zhou, C., and Feng, Y. (2023). Two views of constrained differential privacy: Belief revision and update. In *Proceedings of the AAAI Conference on Artificial Intelligence*.
- Machanavajjhala, A. (2022). Candidate differential privacy algorithms for 2020 Decennial Census Group II products. Workshop on the Analysis of Census Noisy Measurement Files and Differential Privacy. <http://dimacs.rutgers.edu/events/details?eID=2038>. <http://dimacs.rutgers.edu/events/details?eID=2038> [Accessed May 2023].
- Machanavajjhala, A., Kifer, D., Abowd, J. M., Gehrke, J., and Vilhuber, L. (2008). Privacy: Theory meets practice on the map. In *Proceedings of the 2008 IEEE 24th International Conference on Data Engineering*, pages 277–286. IEEE Computer Society.
- Machanavajjhala, A., Kifer, D., Gehrke, J., and Venkitasubramaniam, M. (2007). ℓ -Diversity: Privacy beyond k -anonymity. *ACM Transactions on Knowledge Discovery from Data*, 1(1). <http://www.scopus.com/inward/record.url?scp=34248181923&partnerID=8YFLogxK>.
- MacKinnon, C. A. (1989). *Toward a Feminist Theory of the State*. Harvard University Press, Cambridge, Mass.
- Margulis, S. T. (2003). Privacy as a social issue and behavioral concept. *Journal of Social Issues*, 59(2):243–261.
- Marks, R. and Rios-Vargas, M. (2021). Improvements to the 2020 Census race and hispanic origin question designs, data processing, and coding procedures. <https://www.census.gov/newsroom/blogs/random-samplings/2021/08/improvements-to-2020-census-race-hispanic-ori>

[gin-question-designs.html](#).

- Marley, J. K. and Leaver, V. L. (2011). A method for confidentialising user-defined tables: Statistical properties and a risk-utility analysis. In *Proceedings of the 58th World Statistical Congress*, pages 1072–1081, Dublin, Ireland. International Statistical Institute.
- Mayer, T. S. (2002). Privacy and Confidentiality Research and the U.S. Census Bureau: Recommendations based on a review of the literature. Technical Report #2002-01, Statistical Research Division US Bureau of the Census.
- McCartan, C., Simko, T., and Imai, K. (2023). Researchers need better access to US Census data. *Science*, 380(6648):902–903. <https://www.science.org/doi/10.1126/science.adf7004>.
- McClure, D. and Reiter, J. P. (2012). Differential privacy and statistical disclosure risk measures: An investigation with binary synthetic data. *Trans. Data Priv.*, 5(3):535–552.
- McDougall, B. S. and Hansson, A. (2002). *Chinese Concepts of Privacy*. Brill, Lieden.
- McKenna, L. (2018). Disclosure avoidance techniques used for the 1970 through 2010 Decennial Censuses of Population and Housing. Working paper, The Research and Methodology Directorate - US Census Bureau.
- McKenna, L. and Haubach, M. (2019). Legacy techniques and current research in disclosure avoidance at the U.S. Census Bureau. Working paper, Research and Methodology Directorate, United States Census Bureau.
- McSherry, F. and Mahajan, R. (2010). Differentially-private network trace analysis. In *Proceedings of the ACM SIGCOMM 2010 Conference, SIGCOMM '10*, pages 123–134, New York, NY, USA. Association for Computing Machinery.
- McSherry, F. and Talwar, K. (2007). Mechanism design via differential privacy. In *48th Annual IEEE Symposium on Foundations of Computer Science (FOCS'07)*, pages 94–103.
- Mead, M. (1928). *Coming of Age in Samoa: A Psychological Study of Primitive Youth for Western Civilization*. Blue Ribbon Books, New York.
- Meng, X.-L. (2021). Enhancing (publications on) data quality: Deeper data minding and fuller data confession. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 184(4):1161–1175.
- Messing, S., DeGregorio, C., Hillenbrand, B., King, G., Mahanti, S., Mukerjee, Z., Nayak, C., Persily, N., State, B., and Wilkins, A. (2020a). Facebook Privacy-Protected Full URLs Data Set.

<https://doi.org/10.7910/DVN/TDOAPG>.

- Messing, S., DeGregorio, C., Hillenbrand, B., King, G., Mahanti, S., Mukerjee, Z., Nayak, C., Persily, N., State, B., and Wilkins, A. (2020b). *Urls-v3.pdf*. In *Facebook Privacy-Protected Full URLs Data Set*. Harvard Dataverse.
- Miller, D. (2020a). Covid-19 and the cult of privacy. <https://blogs.ucl.ac.uk/assa/2020/04/30/covid-19-and-the-cult-of-privacy/>.
- Miller, D. (2020b). There is a fine line between care and surveillance. <https://blogs.ucl.ac.uk/assa/2020/03/31/there-is-a-fine-line-between-care-and-surveillance/>.
- Miller, D. (2021). Smartphones and contact-tracing: Balancing care and surveillance. *The Conversation*.
- Miller, D., Rabho, L. A., Awondo, P., de Vries, M., Duque, M., Garvey, P., Haapio-Kirk, L., Hawkins, C., Otaegui, A., Walton, S., and Wang, X. (2021). *The Global Smartphone: Beyond a Youth Technology*. Ageing with Smartphones. UCL Press, London.
- Miranda, E., Pelessoni, R., and Vicig, P. (2024). Evaluating uncertainty with Vertical Barrier Models. *International Journal of Approximate Reasoning*, 167:109132.
- Mironov, I. (2017). Rényi differential privacy. *2017 IEEE 30th Computer Security Foundations Symposium (CSF)*, pages 263–275.
- Mironov, I., Pandey, O., Reingold, O., and Vadhan, S. (2009). Computational differential privacy. In Halevi, S., editor, *Advances in Cryptology - CRYPTO 2009*, pages 126–142, Berlin, Heidelberg. Springer.
- Mitra, R. and Reiter, J. P. (2006). Adjusting survey weights when altering identifying design variables via synthetic data. In *Privacy in Statistical Databases: CENEX-SDC Project International Conference, PSD 2006, Rome, Italy, December 13-15, 2006. Proceedings*, pages 177–188. Springer.
- Montes, I. (2023). Neighbourhood models induced by the Euclidean distance and the Kullback-Leibler divergence. In *Proceedings of the Thirteenth International Symposium on Imprecise Probability: Theories and Applications*, pages 367–378. PMLR. <https://proceedings.mlr.press/v215/montes23a.html>.
- Montes, I., Miranda, E., and Destercke, S. (2020a). Unifying neighbourhood and distortion models: Part I – new results on old models. *International Journal of General Systems*, 49(6):602–635. <https://doi.org/10.1080/03081079.2020.1778682>.

- Montes, I., Miranda, E., and Destercke, S. (2020b). Unifying neighbourhood and distortion models: Part II – new models and synthesis. *International Journal of General Systems*, 49(6):636–674. <https://doi.org/10.1080/03081079.2020.1778683>.
- Moore, A. (2003). Privacy: Its Meaning and Value. *American Philosophical Quarterly*, 40(3):215–227.
- Moore, A. (2008). Defining Privacy. *Journal of Social Philosophy*, 39(3):411–428.
- Moore, A. D. (2015). *Privacy, Security and Accountability: Ethics, Law and Policy*. Rowman & Littlefield.
- Moore, B. (1984). *Privacy: Studies in Social and Cultural History*. M.E. Sharpe.
- Moore, B. (1985). Privacy. *Society*, 22(4):17–27.
- Muralidhar, K. and Domingo-Ferrer, J. (2023). Database reconstruction is not so easy and is different from reidentification. *Journal of Official Statistics*, 39(3):381–398. <https://sciendo.com/en/article/10.2478/jos-2023-0017>.
- Nanayakkara, P., Smart, M. A., Cummings, R., Kaptchuk, G., and Redmiles, E. M. (2023). What are the chances? Explaining the epsilon parameter in differential privacy. In *Proceedings of the 32nd USENIX Conference on Security Symposium, SEC ’23*, pages 1613–1630, USA. USENIX Association.
- Narayanan, A. and Shmatikov, V. (2008). Robust de-anonymization of large sparse datasets. In *2008 IEEE Symposium on Security and Privacy (Sp 2008)*, pages 111–125.
- National Academies of Sciences, Engineering, and Medicine (2017a). Current Challenges and Opportunities in Federal Statistics. In Harris-Kojetin, B. A. and Groves, R. M., editors, *Innovations in Federal Statistics: Combining Data Sources While Protecting Privacy*. National Academies Press (US).
- National Academies of Sciences, Engineering, and Medicine (2017b). Protecting privacy and confidentiality while providing access to data for research use. In Harris-Kojetin, B. A. and Groves, R. M., editors, *Innovations in Federal Statistics: Combining Data Sources While Protecting Privacy*. National Academies Press (US).
- National Academies of Sciences, Engineering, and Medicine (2021). Legislation and regulations that govern federal statistics. In *Principles and Practices for a Federal Statistical Agency*. The National Academies Press (US), Washington DC, seventh edition. <https://www.ncbi.nlm.nih.gov/books/NBK573405/>.
- National Academies of Sciences, Engineering, and Medicine (2024). *A Roadmap for Disclosure Avoidance in the Survey of Income and Program Participation*. The National Academies Press, Washington, D.C.

- National Research Council (1995). *Modernizing the U.S. Census*. National Academies Press.
- Near, J. P. and Abuah, C. (2025). *Programming Differential Privacy*.
- Neunhoeffler, M., Lattner, J., and Drechsler, J. (2024). On the formal privacy guarantees of synthetic data (generated without formal privacy guarantees). In *National Bureau of Economic Research (2024): Data Privacy Protection and the Conduct of Applied Research: Methods, Approaches and Their Consequences*, Washington D.C.
- Nissenbaum, H. (2019). Contextual integrity up and down the data food chain. *Theoretical Inquiries in Law*, 20(1):221–256.
- Nissenbaum, H. F. (2010). *Privacy in Context: Technology, Policy, and the Integrity of Social Life*. Stanford Law Books, Stanford, Calif.
- Nissim, K., Smorodinsky, R., and Tennenholtz, M. (2012). Approximately optimal mechanism design via differential privacy. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference, ITCS '12*, pages 203–213, New York, NY, USA. Association for Computing Machinery.
- Nissim, K. and Wood, A. (2018). Is privacy privacy? *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 376:20170358.
- Norberg, P. A., Horne, D. R., and Horne, D. A. (2007). The privacy paradox: Personal information disclosure intentions versus behaviors. *Journal of Consumer Affairs*, 41(1):100–126.
- Oberski, D. L. and Kreuter, F. (2020). Differential privacy and social science: An urgent puzzle. *Harvard Data Science Review*, 2(1).
- Office for National Statistics (2023). Protecting personal data in Census 2021 results. ONS Website, Methodology. <https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/methodologies/protectingpersonaldataincensus2021results>.
- Office of the Australian Information Commissioner and Lonergan Research (2020). Australian Community Attitudes to Privacy Survey 2020. Technical report, Office of the Australian Information Commissioner.
- O’Keefe, C. M. and Charest, A.-S. (2019). Bootstrap differential privacy. *Transactions on Data Privacy*, 12:1–28.
- Oulasvirta, A., Pihlajamaa, A., Perkiö, J., Ray, D., Vähäkangas, T., Hasu, T., Vainio, N., and Myllymäki,

- P. (2012). Long-term effects of ubiquitous surveillance in the home. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*, UbiComp '12, pages 41–50, New York, NY, USA. Association for Computing Machinery.
- Packard, V. (1964). *The Naked Society*. D. McKay Company, New York.
- Papacharissi, Z. (2010). Privacy as a luxury commodity. *First Monday*.
- Parent, W. A. (1983). Privacy, Morality, and the Law. *Philosophy & Public Affairs*, 12(4):269–288.
- Pearl, J. (1995). Causal diagrams for empirical research. *Biometrika*, 82(4):669–688.
- Pelessoni, R., Vicig, P., and Corsato, C. (2021). Inference with nearly-linear uncertainty models. *Fuzzy Sets and Systems*, 412:1–26. <https://www.sciencedirect.com/science/article/pii/S0165011420301287>.
- Phillips, P. J. (2012). Oral glucose tolerance testing. *Australian Family Physician*, 41(6):391–393. <https://www.racgp.org.au/afp/2012/june/oral-glucose-tolerance-testing>.
- Population Reference Bureau and US Census Bureau’s 2020 Census Data Products and Dissemination Team (2023). Disclosure avoidance and the 2020 Census: How the TopDown algorithm works. 2020 Census Briefs C2020BR-04. <https://www2.census.gov/library/publications/decennial/2020/census-briefs/c2020br-04.pdf> [Accessed: 04-25-2023].
- Posner, R. A. (1981). The Economics of Privacy. *The American Economic Review*, 71(2):405–409.
- Prosser, W. L. (1960). Privacy. *California Law Review*, 48(3):383–423.
- Protivash, P., Durrell, J., Ding, Z., Zhang, D., and Kifer, D. (2022). Reconstruction attacks on aggressive relaxations of differential privacy.
- Raab, C. (2018). Political Science and Privacy. In van der Sloot, B. and de Groot, A., editors, *The Handbook of Privacy Studies: An Interdisciplinary Introduction*, pages 257–262. Amsterdam University Press.
- Rachels, J. (1975). Why Privacy is Important. *Philosophy & Public Affairs*, 4(4):323–333.
- Radway, S. and Christ, M. (2023). The impact of de-identification on single-year-of-age counts in the U.S. Census. <http://arxiv.org/abs/2308.12876>.
- Ramirez, R. and Borman, C. (2021). How we complete the Census when demographic and housing characteristics are missing. <https://www.census.gov/newsroom/blogs/random-samplings/2021/>

[08/census-when-demographic-and-housing-characteristics-are-missing.html](#).

- Raskhodnikova, S. and Smith, A. (2016). Differentially private analysis of graphs. In Kao, M.-Y., editor, *Encyclopedia of Algorithms*, pages 543–547. Springer, New York, NY.
- Reamer, A. (2019). Brief 7: Comprehensive accounting of census-guided federal spending (FY2017). Technical report, George Washington Institute of Public Policy.
- Redberg, R. and Wang, Y.-X. (2021). Privately publishable per-instance privacy. In *Advances in Neural Information Processing Systems*, volume 34, pages 17335–17346. Curran Associates, Inc.
- Reiter, J. P. (2005). Estimating risks of identification disclosure in microdata. *Journal of the American Statistical Association*, 100(472):1103–1112.
- Reiter, J. P. (2019). Differential privacy and federal data releases. *Annual review of statistics and its application*, 6:85–101.
- Rinott, Y., O’Keefe, C. M., Shlomo, N., and Skinner, C. (2018). Confidentiality and differential privacy in the dissemination of frequency tables. *Statistical Science*, 33(3):358–385.
- Robertson Ishii, T. and Atkins, P. (2023). Essential vs. accidental properties. In Zalta, E. N. and Nodelman, U., editors, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, spring 2023 edition.
- Romano, A., Sotis, C., Dominioni, G., and Guidi, S. (2020). The Scale of COVID-19 Graphs Affects Understanding, Attitudes, and Policy Preferences. SSRN Scholarly Paper ID 3588511, Social Science Research Network, Rochester, NY.
- Rubel, A. (2011). The particularized judgment account of privacy. *Res Publica*, 17(3):275–290.
- Rubel, A. P. (2006). *A Philosophical Account of Privacy*. PhD thesis, University of Wisconsin-Madison.
- Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63(3):581–592.
- Rubin, D. B. (2005). Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, 100(469):322–331.
- Ruggles, S., Fitch, C., Magnuson, D., and Schroeder, J. (2019). Differential privacy and Census data: Implications for social and economic research. *AEA Papers and Proceedings*, 109:403–408.
- Ruggles, S., Fitch, C. A., Goeken, R., Hacker, J. D., Nelson, M. A., Roberts, E., Schouweiler, M., and

- Sobek, M. (2021). IPUMS ancestry full count data: Version 3.0 [dataset]. Minneapolis, MN: IPUMS. <https://doi.org/10.18128/D014.V3.0>.
- Russell, J. N. (2022). National center for health statistics virtual data enclave (vde) project overview. <https://www.cdc.gov/nchs/data/bsc/bsc-pres-j-neil-russell-5-26-2022.pdf>.
- Sánchez, D., Jebreel, N., Domingo-Ferrer, J., Muralidhar, K., and Blanco-Justicia, A. (2023). An examination of the alleged privacy threats of confidence-ranked reconstruction of Census microdata. <http://arxiv.org/abs/2311.03171>.
- Schmutte, I. M. (2016). Differentially private publication of data on wages and job mobility. *Statistical Journal of the IAOS*, 32(1):81–92.
- Schneider, M. J., Bailie, J., and Iacobucci, D. (2025). Why data anonymization hasn’t taken off. *Submitted to Customer Needs and Solutions*.
- Seeman, J., Reimherr, M., and Slavkovic, A. (2022). Formal privacy for partially private data. <http://arxiv.org/abs/2204.01102>.
- Seeman, J., Sexton, W., Pujol, D., and Machanavajjhala, A. (2023). Privately answering queries on skewed data via per record differential privacy. <http://arxiv.org/abs/2310.12827>.
- Seeman, J., Si, Y., and Reiter, J. P. (2024). Differentially private population quantity estimates via survey weight regularization. https://conference.nber.org/conf_papers/f194295.pdf.
- Seeman, J. and Susser, D. (2023). Between privacy and utility: On differential privacy in theory and practice. *ACM Journal on Responsible Computing*.
- Seidenfeld, T. and Wasserman, L. (1993). Dilation for sets of probabilities. *The Annals of Statistics*, 21(3):1139–1154.
- Shafer, G. (1976). *A mathematical theory of evidence*. Princeton University Press, Princeton, NJ.
- Shapiro, F. R. and Pearse, M. (2012). The most-cited law review articles of all time. *Michigan Law Review*, 110(8):1483–1520.
- Sharma, T. and Bashir, M. (2020). Use of apps in the COVID-19 response and the loss of privacy protection. *Nature Medicine*, 26(8):1165–1167.
- Shlomo, N., Tudor, C., and Groom, P. (2010). Data swapping for protecting Census tables. In Domingo-Ferrer, J. and Magkos, E., editors, *Privacy in Statistical Databases*, Lecture Notes in Computer Science,

pages 41–51, Berlin, Heidelberg. Springer.

Simon, H. A. (1971). Designing Organizations for an Information-Rich World. In Greenberger, M., editor, *Computers, Communication, and the Public Interest*, pages 37–52. Johns Hopkins University Press, Baltimore, MD.

Singer, E. and Couper, M. P. (2010). Communicating Disclosure Risk in Informed Consent Statements. *Journal of Empirical Research on Human Research Ethics*, 5(3):1–8.

Slavković, A. and Seeman, J. (2023). Statistical data privacy: A song of privacy and utility. *Annual Review of Statistics and Its Application*, 10(1):189–218.

Slavković, A. B. and Lee, J. (2010). Synthetic two-way contingency tables that preserve conditional frequencies. *Statistical Methodology*, 7(3):225–239.

Smart, M. A., Sood, D., and Vaccaro, K. (2022). Understanding risks of privacy theater with differential privacy. In *Proceedings of the ACM on Human-Computer Interaction*, volume 6, pages 342:1–342:24.

Smith, A. (2011). Privacy-preserving statistical estimation with optimal convergence rates. In *Proceedings of the Forty-Third Annual ACM Symposium on Theory of Computing*, STOC ’11, pages 813–822, New York, NY, USA. Association for Computing Machinery.

Solove, D. J. (2002). Conceptualizing Privacy. *California Law Review*, 90(4):1087–1155.

Solove, D. J. (2008). *Understanding Privacy*. Harvard University Press, Cambridge, MA.

Solove, D. J. and Schwartz, P. M. (2021). *Information Privacy Law*. Aspen Casebook Series. Wolters Kluwer, New York, seventh edition edition.

Soria-Comas, J., Domingo-Ferrer, J., Sánchez, D., and Megías, D. (2017). Individual differential privacy: A utility-preserving formulation of differential privacy guarantees. *IEEE Transactions on Information Forensics and Security*, 12(6):1418–1429.

Spicer, K. (2020). Statistical disclosure control (SDC) for 2021 UK Census. Technical Report EAP125, UK Statistics Authority. <https://uksa.statisticsauthority.gov.uk/wp-content/uploads/2020/07/EAP125-Statistical-Disclosure-Control-SDC-for-2021-UK-Census.docx>.

Stanford, J. (2020). Why some Americans don’t trust the census. *The Conversation*.

Statistics New Zealand (2022). How we keep integrated data safe. <https://www.stats.govt.nz/integrated-data/how-we-keep-integrated-data-safe/>.

- Steel, P. and Zayatz, L. (2003). The effects of the disclosure limitation procedure on Census 2000 tabular data products (abridged). Technical Report Census 2000 Evaluation C.1, Statistical Research Division US Census Bureau.
- Steinke, T. (2022). Composition of differential privacy & privacy amplification by subsampling. <http://arxiv.org/abs/2210.00597>.
- Stigler, G. J. (1980). An Introduction to Privacy in Economics and Politics. *The Journal of Legal Studies*, 9(4):623–644.
- Stokes, P. (2017). The ‘Five Safes’ – Data Privacy at ONS. <https://blog.ons.gov.uk/2017/01/27/the-five-safes-data-privacy-at-ons/>.
- Stuart, A., Bandara, A. K., and Levine, M. (2019). The psychology of privacy in the digital age. *Social and Personality Psychology Compass*, 13(11):e12507.
- Su, W. J. (2024). A statistical viewpoint on differential privacy: Hypothesis testing, representation and Blackwell’s theorem. <http://arxiv.org/abs/2409.09558>.
- Sweeney, L. (2000). Simple demographics often identify people uniquely. Data Privacy Working Paper 3, Carnegie Mellon University, Pittsburgh.
- Sweeney, L. (2002). *k*-Anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(5):557–570. <https://www.worldscientific.com/doi/abs/10.1142/S0218488502001648>.
- Takagi, S., Kato, F., Cao, Y., and Yoshikawa, M. (2022). Asymmetric differential privacy. In *2022 IEEE International Conference on Big Data (Big Data)*, pages 1576–1581. <https://ieeexplore.ieee.org/document/10020709>.
- Talwar, K., Guha Thakurta, A., and Zhang, L. (2015). Nearly optimal private LASSO. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.
- Thompson, G., Broadfoot, S., and Elazar, D. (2013). Methodology for the automatic confidentialisation of statistical outputs from remote servers at the Australian Bureau of Statistics. In *Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality*, page 38, Ottawa, Canada. <https://www.unece.org/stats/documents/2013.10.confidentiality.html>.
- Thomson, J. J. (1975). The Right to Privacy. *Philosophy & Public Affairs*, 4(4):295–314.
- Trope, Y. and Liberman, N. (2010). Construal-Level Theory of Psychological Distance. *Psychological*

review, 117(2):440–463.

Tschantz, M. C., Sen, S., and Datta, A. (2020). SoK: Differential privacy as a causal property. In *2020 IEEE Symposium on Security and Privacy (SP)*, pages 354–371. IEEE.

Tumult Labs (2022). SafeTab: DP algorithms for 2020 Census Detailed DHC Race & Ethnicity. Technical report.

Uber Security (2017). Uber releases open source project for differential privacy. <https://medium.com/uber-security-privacy/differential-privacy-open-source-7892c82c42b6>.

UK Data Service (2023). What is the Five Safes framework? <https://ukdataservice.ac.uk/help/secure-lab/what-is-the-five-safes-framework/>.

UK Statistics Authority (2021). Transparency of SDC methods and parameters (post-meeting EAP paper for SDC for Census August 2021). Technical Report EAP168, UK Statistics Authority. <https://uksa.statisticsauthority.gov.uk/wp-content/uploads/2022/02/EAP168-Statistical-Disclosure-Control-for-Census.pdf>.

United States Census Bureau (2019). 2020 Census Barriers, Attitudes, and Motivators Study (CBAMS) Focus Group Final Report. Technical report, US Department of Commerce.

US Bureau of Labor Statistics (2018a). Current Population Survey: Concepts. <https://www.bls.gov/opub/hom/cps/concepts.htm>.

US Bureau of Labor Statistics (2018b). Current Population Survey: Design. <https://www.bls.gov/opub/hom/cps/design.htm>.

US Census Bureau (2010). 2010 Sample Census Form. <https://www.census.gov/history/pdf/2010questionnaire.pdf>.

US Census Bureau (2012). 2010 Census Summary File 1—Technical Documentation. Technical Report SF1/10-4 (RV). <https://www2.census.gov/programs-surveys/decennial/2010/technical-documentation/complete-tech-docs/summary-file/sf1.pdf>.

US Census Bureau (2016). Decennial Census Surname Files (2010, 2000). <https://www.census.gov/data/developers/data-sets/surnames.html>.

US Census Bureau (2018). Protecting the confidentiality of America’s statistics: Adopting modern disclosure avoidance methods at the Census Bureau. https://www.census.gov/newsroom/blogs/research-matters/2018/08/protecting_the_conf.html.

US Census Bureau (2019a). Design and methodology: Current Population Survey. *Technical Paper* 77.

US Census Bureau (2019b). A history of census privacy protections. Technical report.

US Census Bureau (2021a). 2020 Census National Redistricting Data Summary File. Technical report.

US Census Bureau (2021b). 2020 Census State Redistricting Data Summary File. Technical report.

US Census Bureau (2021c). Comparing differential privacy with older disclosure avoidance methods. Factsheet D-FS-GP-EN-0509. <https://www.census.gov/content/dam/Census/library/factsheets/2021/comparing-differential-privacy-with-older-disclosure-avoidance-methods.pdf>.

US Census Bureau (2021d). Disclosure avoidance for the 2020 Census: An introduction. Technical report, US Government Publishing Office, Washington, D.C.

US Census Bureau (2021e). Guidance for geography users: Hierarchy diagrams. <https://www.census.gov/programs-surveys/geography/guidance/hierarchy.html>.

US Census Bureau (2022a). 2010 demonstration data product - Demographic and Housing Characteristics technical document. Technical report, US Census Bureau.

US Census Bureau (2022b). American Community Survey and Puerto Rico Community Survey design and methodology. Technical Report Version 3.0.

US Census Bureau (2022c). Disclosure avoidance protections for the American Community Survey. <https://www.census.gov/newsroom/blogs/random-samplings/2022/12/disclosure-avoidance-protections-ac.html>.

US Census Bureau (2022d). Synthetic sipp data. <https://www.census.gov/programs-surveys/sipp/guidance/sipp-synthetic-beta-data-product.html>.

US Census Bureau (2023a). 2020 Census 118th Congressional District Summary File (CD118). <https://www.census.gov/data/tables/2023/dec/2020-census-CD118.html>.

US Census Bureau (2023b). 2020 Census Demographic and Housing Characteristics (DHC) Demonstration Noisy Measurement File (2023-10-23) README file. Technical report.

US Census Bureau (2023c). 2020 Census Demographic and Housing Characteristics File (DHC). Technical report.

- US Census Bureau (2023d). 2020 Census Demographic and Housing Characteristics File (DHC) technical documentation. Technical report, Washington, DC. <https://www2.census.gov/programs-surveys/decennial/2020/technical-documentation/complete-tech-docs/demographic-and-housing-characteristics-file-and-demographic-profile/2020census-demographic-and-housing-characteristics-file-and-demographic-profile-techdoc.pdf>.
- US Census Bureau (2023e). 2020 Census demographic profile. <https://www.census.gov/data/tables/2023/dec/2020-census-demographic-profile.html>.
- US Census Bureau (2023f). 2020 Census Detailed Demographic and Housing Characteristics File A (Detailed DHC-A) technical documentation. Technical report.
- US Census Bureau (2023g). 2020 Census Disclosure Avoidance System detailed summary metrics. <https://www2.census.gov/programs-surveys/decennial/2020/data/demographic-and-housing-characteristics-file/2020-Census-Disclosure-Avoidance-System-Detailed-Summary-Metrics.xlsx>.
- US Census Bureau (2023h). 2020 Census Redistricting Data (P.L. 94-171) Noisy Measurement File README File. https://www2.census.gov/programs-surveys/decennial/2020/data/01-Redistricting_File--PL_94-171/00-2020-Redistricting-Noisy-Measurement-File/2020%20Redistricting%20NMF%202023-06-15%20README.html.
- US Census Bureau (2023i). 2023-04-03 Privacy-loss budget allocations. Technical report. https://www2.census.gov/programs-surveys/decennial/2020/program-management/data-product-planning/2010-demonstration-data-products/04-Demonstration_Data_Products_Suite/2023-04-03/2023-04-03_Privacy-Loss_Budget_Allocations.pdf [Accessed: 04-25-2023].
- US Census Bureau (2023j). About 2020 Census data products. <https://www.census.gov/programs-surveys/decennial-census/decade/2020/planning-management/release/about-2020-data-products.html>.
- US Census Bureau (2023k). American community survey 2017-2021 5-year PUMS user guide and overview. https://www2.census.gov/programs-surveys/acs/tech_docs/pums/2017_2021ACS_PUMS_User_Guide.pdf.
- US Census Bureau (2023l). Disclosure avoidance and the 2020 Census: How the TopDown Algorithm works. Technical report.
- US Census Bureau (2023m). Disclosure avoidance methods for the detailed demographic and housing

characteristics file a (Detailed DHC-A): How SafeTab-P works. Technical report.

US Census Bureau (2023n). Factsheet on disclosure avoidance for the 2010 demonstration data products suite – Redistricting and Demographic and Housing Characteristics File – production settings (2023-04-03). Technical report. https://www2.census.gov/programs-surveys/decennial/2020/program-management/data-product-planning/2010-demonstration-data-products/04-Demonstration_Data_Products_Suite/2023-04-03/2023-04-03_Factsheet.pdf [accessed 2023-04-25].

US Census Bureau (2023o). List of surveys. <https://www.census.gov/programs-surveys/surveyhelp/list-of-surveys.html>.

US Census Bureau (2023p). Methodology, assumptions and inputs for the 2023 National Population Projections. Technical report. <https://www2.census.gov/programs-surveys/popproj/technical-documentation/methodology/methodstatement23.pdf>.

US Census Bureau (2023q). Methodology for the United States Population Estimates: Vintage 2023. Technical report. <https://www2.census.gov/programs-surveys/popest/technical-documentation/methodology/2020-2023/methods-statement-v2023.pdf>.

US Census Bureau (2023r). Release dates set for next 2020 Census data products; New reader-friendly disclosure avoidance briefs. <https://www.census.gov/programs-surveys/decennial-census/decade/2020/planning-management/process/disclosure-avoidance/newsletters/release-dates-and-da-briefs.html>.

US Census Bureau (2024a). 2020 Census Detailed Demographic and Housing Characteristics File B (Detailed DHC-B) technical documentation. Technical report.

US Census Bureau (2024b). 2020 Census Privacy-Protected Microdata File (PPMF). Technical report.

US Census Bureau (2024c). 2020 Supplemental Demographic and Housing Characteristics File (S-DHC) technical documentation. Technical report.

US Census Bureau (2024d). Census Bureau releases final 2020 Census data product. <https://www.census.gov/newsroom/press-releases/2024/final-2020-census-data-product-s-dhc.html>.

US Census Bureau (2024e). Disclosure avoidance and the 2020 Census: How the SafeTab-H works. Technical report.

US Census Bureau (2024f). Disclosure avoidance and the Supplemental Demographic and Housing Char-

- acteristics File (S-DHC): How PHSafe works. Technical report.
- Vadhan, S. (2017). The complexity of differential privacy. In Lindell, Y., editor, *Tutorials on the Foundations of Cryptography*, pages 347–450. Springer International Publishing, Cham.
- Valliant, R., Dever, J. A., and Kreuter, F. (2018). *Practical tools for designing and weighting survey samples*. Springer, 2 edition.
- van der Geest, S. (2018). Privacy from an Anthropological Perspective. In van der Sloot, B. and de Groot, A., editors, *The Handbook of Privacy Studies: An Interdisciplinary Introduction*, pages 413–444. Amsterdam University Press.
- Van Der Horst, H. and Messing, J. (2006). “It’s not Dutch to close the curtains”: Visual struggles on the threshold between public and private in a multi-ethnic Dutch neighborhood. *Home Cultures*, 3(1):21–37.
- Varian, H. R. (1997). Economic aspects of personal privacy. In National Telecommunications and Information Administration Office of Chief Counsel, editor, *Privacy and Self-Regulation in the Information Age*. US Department of Commerce, Washington, D.C.
- Villa Ross, C. (2023). Uses of Decennial Census programs data in federal funds distribution: Fiscal year 2021. Technical report, US Census Bureau.
- Waldman, A. E. (2021). *Industry unbound: The inside story of privacy, data, and corporate power*. Cambridge University Press.
- Walley, P. (1991). *Statistical Reasoning with Imprecise Probabilities*. Number 42 in Monographs on Statistics and Applied Probability. Chapman and Hall, London ; New York, 1st ed edition.
- Walter, G. and Augustin, T. (2009). Imprecision and prior-data conflict in generalized Bayesian inference. *Journal of Statistical Theory and Practice*, 3(1):255–271.
- Wang, X. (2019). Hundreds of Chinese citizens told me what they thought about the controversial social credit system. *The Conversation*.
- Wang, X. Y. (2020). Chinese people living overseas: Dilemmas during COVID-19. <https://blogs.ucl.ac.uk/assa/2020/03/30/the-dilemma-of-chinese-people-living-over-seas-during-covid-19/>.
- Wang, Y., Balle, B., and Kasiviswanathan, S. P. (2019). Subsampled Rényi differential privacy and analytical moments accountant. In *The 22nd International Conference on Artificial Intelligence and Statistics*,

- AISTATS 2019, 16-18 April 2019, Naha, Okinawa, Japan*, volume 89 of *Proceedings of Machine Learning Research*, pages 1226–1235. PMLR.
- Wang, Y., Basciftci, Y. O., and Ishwar, P. (2017). Privacy-utility tradeoffs under constrained data release mechanisms. <http://arxiv.org/abs/1710.09295>.
- Wang, Y.-X. (2018). Per-instance Differential Privacy. <http://arxiv.org/abs/1707.07708>.
- Wang, Y.-X., Lei, J., and Fienberg, S. E. (2016). Learning with differential privacy: Stability, learnability and the sufficiency and necessity of ERM principle. *J. Mach. Learn. Res.*, 17:183:1–183:40.
- Warner, S. L. (1965). Randomized response: A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association*, 60(309):63–69.
- Warren, S. D. and Brandeis, L. D. (1890). The Right to Privacy. *Harvard Law Review*, 4(5):193.
- Wasserman, L. (1992). Invariance properties of density ratio priors. *The Annals of Statistics*, 20(4):2177–2182.
- Wasserman, L. and Kadane, J. B. (1992). Computing bounds on expectations. *Journal of the American Statistical Association*, 87(418):516–522.
- Wasserman, L. and Zhou, S. (2010). A statistical framework for differential privacy. *Journal of the American Statistical Association*, 105(489):375–389.
- Westin, A. F. (1967). *Privacy and Freedom*. Atheneum, New York.
- Westin, A. F. (2003). Social and Political Dimensions of Privacy. *Journal of Social Issues*, 59(2):431–453.
- Wilk, R. (2018). Internet Privacy Hogwash. *Anthropology News*, 59(3):e153–e156.
- Williams, O. and McSherry, F. (2010). Probabilistic inference and differential privacy. *Advances in Neural Information Processing Systems*, 23:2451–2459.
- Wing, J. M. (2019). The data life cycle. *Harvard Data Science Review*, 1(1).
- Winner, L. (1980). Do Artifacts Have Politics? *Daedalus*, 109(1):121–136.
- Yan, T., Fricker, S., and Tsai, S. (2020). Response Burden: What Is It and What Predicts It? In *Advances in Questionnaire Design, Development, Evaluation and Testing*, chapter 8, pages 193–212. John Wiley & Sons, Ltd.

- Yu, B. (2013). Stability. *Bernoulli*, 19(4):1484–1500.
- Zayatz, L. (2003). Disclosure limitation for Census 2000 tabular data. In *Joint ECE/Eurostat Workshop on Statistical Data Confidentiality*, volume Working Paper #15, page 9, Luxembourg. United Nations Economic Commission for Europe.
- Zayatz, L. (2007). Disclosure avoidance practices and research at the U.S. Census Bureau: An update. *Journal of Official Statistics*, 23(2):253–265.
- Zayatz, L., Lucero, J., Massell, P., and Ramanayake, A. (2010). Disclosure avoidance for Census 2010 and American Community Survey Five-year tabular data products. In *JSM Section on Survey Research Methods*, pages 2279–2288. American Statistical Association.
- Zhang, W., Ohrimenko, O., and Cummings, R. (2022). Attribute privacy: Framework and mechanisms. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability and Transparency, FAccT ’22*, pages 757–766, New York, NY, USA. Association for Computing Machinery.
- Zhou, S., Ligett, K., and Wasserman, L. (2009). Differential privacy with compression. In *Proceedings of the 2009 IEEE International Conference on Symposium on Information Theory - Volume 4, ISIT’09*, pages 2718–2722, Coex, Seoul, Korea. IEEE Press.

ProQuest Number: 32040385

INFORMATION TO ALL USERS

The quality and completeness of this reproduction is dependent on the quality and completeness of the copy made available to ProQuest.



Distributed by
ProQuest LLC a part of Clarivate (2025).
Copyright of the Dissertation is held by the Author unless otherwise noted.

This work is protected against unauthorized copying under Title 17,
United States Code and other applicable copyright laws.

This work may be used in accordance with the terms of the Creative Commons license or other rights statement, as indicated in the copyright statement or in the metadata associated with this work. Unless otherwise specified in the copyright statement or the metadata, all rights are reserved by the copyright holder.

ProQuest LLC
789 East Eisenhower Parkway
Ann Arbor, MI 48108 USA