The Implications of Privacy-Aware Choice

Thesis by

Rachel Cummings

In Partial Fulfillment of the Requirements for the
degree of
Doctor of Philosophy

CALIFORNIA INSTITUTE OF TECHNOLOGY
Pasadena, California

2017
Defended May 15, 2017

ACKNOWLEDGEMENTS

First and foremost, I would like to thank my advisor, Katrina Ligett. This thesis would not have been possible without her support and guidance, and I couldn't dream of a better advisor. She worked tirelessly to mold me into a successful researcher, with exactly the right combination of support, encouragement, pressure, and freedom. She is my role model, and continues to inspire me both as an academic and as a human being.

I spent time at many wonderful institutions during my tenure as a graduate student: Northwestern University, California Institute of Technology, University of Pennsylvania, Technicolor Research, Microsoft Research NYC, the Simons Institute, and the Hebrew University. Thank you to all my hosts, friends, and collaborators from each of these institutions. Aaron Roth and Michael Kearns deserve special mention for hosting me at University of Pennsylvania for my most productive semester yet.

I was told that one should work with as many talented people as possible; you soak up the skills of those around you. I took this advise to heart, and was rewarded with a collection of incredible collaborators. Thank you to all of my coauthors for sharing your skills with me.

I owe thanks to my educators and mentors from my days before graduate school: Joe Talley was the first to encourage me to pursue a career in mathematics; Kristina Pierce was my first badass female mathematician role model; Mark Moore was the first to instill in me a love of economics; Satwindar Sadhal encouraged me to stay the course when my own confidence wavered. An enormous debt of gratitude is owed to my undergraduate advisor, David Kempe, who patiently showed me how to do research and encouraged me to apply to graduate school in computer science.

Thank you to Shaddin Dughmi, who was my best friend and partner for most of this journey. Although our paths have diverged, our time together shaped me into the researcher and the person I am today.

I've been fortunate to have amazing friends and family who kept me sane these past six years. Thank you all for the parties, meals, coffee breaks, cocktails, Skype dates, endless texting, and so much more. Special thanks to Chelsea (Fung) Meier and Lindsey Stiff for being there every step of the way. Thank you to my sister Morgan Cummings who raised the bar for excellence, and was always both my fiercest competitor and my closest ally. Thank you to Jennifer and Ed Kuyper for being my cheering squad and for making the cutest niece and nephew in the world.

Most importantly, I would like to thank my parents for their unwavering love and support, and for always being my biggest fans. They raised me to value education above all else, and to relentlessly pursue my dreams. All of my accomplishments are products of their sacrifices.

ABSTRACT

Privacy concerns are becoming a major obstacle to using data in the way that we want. It's often unclear how current regulations should translate into technology, and the changing legal landscape surrounding privacy can cause valuable data to go unused. In addition, when people know that their current choices may have future consequences, they might modify their behavior to ensure that their data reveal less—or perhaps, more favorable—information about themselves. Given these concerns, how can we continue to make use of potentially sensitive data, while providing satisfactory privacy guarantees to the people whose data we are using? Answering this question requires an understanding of how people reason about their privacy and how privacy concerns affect behavior.

In this thesis, we study how strategic and human aspects of privacy interact with existing tools for data collection and analysis. We begin by adapting the standard model of consumer choice theory to a setting where consumers are aware of, and have preferences over, the information revealed by their choices. In this model of privacy-aware choice, we show that little can be inferred about a consumer's preferences once we introduce the possibility that she has concerns about privacy, even when her preferences are assumed to satisfy relatively strong structural properties. Next, we analyze how privacy technologies affect behavior in a simple economic model of data-driven decision making. Intuition suggests that strengthening privacy protections will both increase utility for the individuals providing data and decrease usefulness of the computation. However, we demonstrate that this intuition can fail when strategic concerns affect consumer behavior. Finally, we study the problem an analyst faces when purchasing and aggregating data from strategic individuals with complex incentives and privacy concerns. For this problem, we provide both mechanisms for eliciting data that satisfy the necessary desiderata, and impossibility results showing the limitations of privacy-preserving data collection.

PUBLISHED CONTENT AND CONTRIBUTIONS

[1] Rachel Cummings, Federico Echenique, and Adam Wierman. “The Empirical Implications of Privacy-Aware Choice”. In: *Operations Research* 64.1 (2016). Preliminary Version appeared in the Proceedings of the 15th ACM Conference on Electronic Commerce (EC 2014), pp. 67–78. doi: 10.1287/opre.2015.1458. url: https://arxiv.org/abs/1401.0336.
Adapted into Chapter 4 of this thesis. R. Cummings contributed to the conception of the project, proving the results, and writing the manuscript.

[2] Rachel Cummings, Stratis Ioannidis, and Katrina Ligett. “Truthful Linear Regression”. In: *Proceedings of The 28th Conference on Learning Theory*. COLT ’15. 2015, pp. 448–483. url: https://arxiv.org/abs/1506.03489.
Adapted into Chapter 6 of this thesis. R. Cummings contributed to the conception of the project, proving the results, and writing the manuscript.

[3] Rachel Cummings, David M. Pennock, and Jennifer Wortman Vaughan. “The Possibilities and Limitations of Private Prediction Markets”. In: *Proceedings of the 17th ACM Conference on Economics and Computation*. EC ’16. 2016, pp. 143–160. doi: 10.1145/2940716.2940721. url: https://arxiv.org/abs/1602.07362.
Adapted into Chapter 7 of this thesis. R. Cummings contributed to the conception of the project, proving the results, and writing the manuscript.

[4] Rachel Cummings et al. “The Strange Case of Privacy in Equilibrium Models”. In: *Proceedings of the 17th ACM Conference on Economics and Computation*. EC ’16. 2016, pp. 659–659. doi: 10.1145/2940716.2940740. url: https://arxiv.org/abs/1508.03080.
Adapted into Chapter 5 of this thesis. R. Cummings contributed to the conception of the project, proving the results, and writing the manuscript.

TABLE OF CONTENTS

Part I

Motivation and Background

Chapter 1

INTRODUCTION

Privacy concerns are becoming a major obstacle to using data in the way that we want. It's often unclear how current regulations should translate into technology, and the changing legal landscape surrounding privacy can cause valuable data to go unused. For well-intentioned companies holding potentially sensitive data, the easiest and safest mode of compliance is simply to not use the data. For example, it is common practice for major companies to limit the sharing of data across departments, or to avoid storing sensitive data altogether. A recent Scientific American article described this occurrence: "as awareness of these privacy concerns has grown, many organizations have clamped down on their sensitive data, uncertain about what, if anything, they can release" [65]. Given these concerns, how can we continue to make use of potentially sensitive data, while providing rigorous privacy guarantees to the people whose data we are using? Answering this question requires tools to guarantee privacy in data analysis, as well as an understanding of how people reason about their privacy and how privacy concerns affect behavior.

In the last decade, a growing literature on *differential privacy* has emerged to address some of these concerns (see Dwork [33] and Dwork and Roth [34] for a survey). First defined by Dwork et al. [36], differential privacy is a parameterized notion of database privacy that gives a mathematically rigorous worst-case bound on the maximum amount of information that can be learned about any one individual's data from the output of a computation. Differential privacy ensures that if a single entry in the database were to be changed, then the algorithm would still have approximately the same distribution over outputs. The privacy community has been prolific in designing algorithms that satisfy this privacy guarantee and maintain usefulness of the computation, resulting in a theoretical toolbox for a wide variety of computational settings, including machine learning, optimization, statistics, and algorithmic economics. Recently, major companies such as Apple [3] and Google [41], and government organizations such as the United Status Census Bureau [70], have announced a commitment to implementing differentially private algorithms. However, many theoretical results cannot be directly applied by practitioners whose problems don't match the underlying assumptions.

In order to achieve a comprehensive solution to the privacy challenges we face today, we need to understand how existing tools for differentially private data analysis interact with *strategic and human aspects* of practical privacy guarantees.

This thesis seeks to bridge this gap between theory and practice in the formal study of privacy. We begin by adapting the standard model of consumer choice theory to a setting where consumers are aware of, and have preferences over, the information revealed by their choices. In this model of privacy-aware choice, we show that little can be inferred about a consumer's preferences once we introduce the possibility that she has concerns about privacy, even when her preferences are assumed to satisfy relatively strong structural properties. Next, we analyze how privacy technologies affect behavior in a simple economic model of data-driven decision making. Intuition suggests that strengthening privacy protections will both increase utility for the individuals providing data and decrease usefulness of the computation. However, we demonstrate that this intuition can fail when strategic concerns affect consumer behavior. Finally, we study the problem an analyst faces when purchasing and aggregating data from strategic individuals with complex incentives and privacy concerns. For this problem, we provide both mechanisms for eliciting data that satisfy the necessary desiderata, and impossibility results showing the limitations of privacy-preserving data collection.

1.1 Overview of the thesis

Privacy concerns affect the interpretability of data

When people know that their current choices may have future consequences, they might modify their behavior to ensure that their data reveal less—or perhaps, more favorable—information about themselves, even if they do not fully understand the consequences of their actions. For example, a person may choose to not download a particular app, change their privacy settings, or even purchase a product from a different website. Indeed, 85% of adult Internet users have taken steps to avoid surveillance by other people or organizations [86]. If analysts use the data of privacy-aware individuals for learning or inference, will the results still be meaningful? The classical *revealed preferences problem* from economics models an observer making inferences about a consumer based upon her choices; in the absence of privacy concerns, an observer can see these choices and, after enough observations, learn the utility function that guides the consumer's decisions. However, when the consumer can take action to change what is revealed, existing tools are ill-equipped to describe such privacy-aware choices.

In Chapter 4, based on joint work with Federico Echenique and Adam Wierman that appeared in EC 2014 and Operations Research in 2016 [21], we initiate the study of the testable implications of choice data, in settings where consumers are privacy-aware. The main message of the paper is that *little can be inferred about a consumer's preferences* once we introduce the possibility that she has concerns about privacy. No matter what her behavior, she always has an "alibi" that can explain her choices as a consequence of privacy concerns. We show that all possible behaviors on the part of the consumer are compatible with all possible preferences she may have over objects, even when her preferences are assumed to satisfy natural economic properties such as separability and monotonicity, which normally place strong restrictions on behavior.

The main result of Chapter 4 is a constructive proof of this claim using tools from economics and graph theory. We adapt the standard model of consumer choice theory to a situation where the consumer is aware of, and has preferences over, the information revealed by her choices. We represent privacy-aware preferences as a directed graph, where vertices of the graph are pairs of choice objects and inferences made by an observer. By imposing edges on this graph according to the constraints of our desiderata—rationalizability, monotonicity, and separability—we show that the graph corresponds to a monotone and separable privacy-aware preference ordering that is consistent with the observed choice behavior.

Impact of privacy policy

Given the promise of differential privacy, one tempting response to privacy concerns is regulation: lawmakers could mandate the use of differentially private algorithms or other privacy technologies, to limit the amount of information that firms can learn about consumers. An implicit assumption in the prior literature is that strengthening privacy protections will both increase utility for the individuals providing data and decrease usefulness of the computation. However, this assumption can fail when strategic concerns affect the impact and guarantees one can get from privacy technologies!

Chapter 5, based on joint work with Katrina Ligett, Mallesh Pai, and Aaron Roth that appeared at EC 2016 [27], serves as a cautionary tale against blindly setting privacy policy in strategic settings: the static effects of adding privacy technologies to a system may be the exact opposite of the effects in equilibrium. In that chapter, we study how privacy technologies affect behavior in a simple economic model of data-driven decision making. A lender would like to use a consumer's past purchases to decide the terms of a new loan, but he is given only a differentially private signal about the consumer's behavior—which can range from no signal at all to a perfect signal, as we vary the differential privacy parameter. Using tools from privacy and game theory, we analyze end-to-end privacy guarantees of this game. We characterize equilibrium behavior as a function of the privacy level promised to the consumer, and show that *the effect of adding privacy in equilibrium can be highly counterintuitive*. Specifically, increasing the level of privacy can actually cause the lender to learn more about the consumer, and can also lead to decreased utility for the consumer and increased utility for the lender. We show that these quantities can generally be non-monotonic and even discontinuous in the privacy level of the signal. These results demonstrate that even in simple models, privacy exhibits much richer behavior in equilibrium than compared to its static counterpart, and suggest that future policy decisions about privacy technologies ought to consider equilibrium effects.

Eliciting data from individuals

Much of the work in private data analysis starts from the premise that an analyst has access to a fixed database that is representative of the underlying population. But how does the analyst acquire this database? If she were to elicit it, why would anyone truthfully report their data? In many practical settings, people can misreport their data (e.g., change browsing behavior or lie on a survey) or refuse to participate (e.g., delete cookies or opt out of a service). These individuals may wish to influence the outcome of a computation performed on their data, or to mask their input due to privacy concerns. If they could potentially come to harm through the use of their private data, they may require additional compensation for this loss. Chapters 6 and 7 study several key challenges an analyst faces when purchasing and aggregating data from strategic individuals with complex incentives and privacy concerns. In these settings, we must design differentially private mechanisms that incentivize players to truthfully share data, allow the analyst to perform her learning task, and minimize the analyst's costs.

In Chapter 6, based on joint work with Stratis Ioannidis and Katrina Ligett that appeared in COLT 2015 [22], we consider a setting where data is unverifiable—such as taste in movies or political beliefs—and *individuals are able to misreport their data to the analyst.* Privacy-aware individuals hold data drawn according to an unknown linear model, which an analyst wishes to learn. The analyst can offer both a privacy guarantee and payments to incentivize players to truthfully report their data, and wishes to minimize her total payments and while still accurately estimating the model. We designed a truthful, individually rational mechanism that produced an asymptotically accurate estimate and allows the analyst's budget to diminish towards zero as the number of participants grows large. The main technical challenge in solving this problem is that differentially private computation of a linear model produces a biased estimate, and existing approaches for eliciting data from privacy-sensitive individuals do not generalize well to biased estimators. We overcome this using tools from peer prediction [74] to design our payment scheme, which leveraged the linear correlation of players' data to induce truthfulness: each player is paid based upon how well her report predicted the reports of other players. Accuracy of the computation followed because the vast majority of players were incentivized to report truthfully, and from a bound on the noise needed to preserve privacy.

Chapter 7, based on joint work with David Pennock and Jennifer Wortman Vaughan that appeared in EC 2016 [23], asks whether *existing techniques for data collection are compatible with differential privacy*. We give both positive and negative results for the design of private prediction markets: financial markets designed to elicit predictions about uncertain events. We first provide a class of private one-shot wagering mechanisms—in which bettors specify a belief about a future event and a monetary wager—that satisfy a number of desirable properties, including truthfulness, budget balance, and differential privacy of the bettors' reported beliefs. We then consider dynamic prediction markets, focusing our attention on the popular cost-function framework in which securities with payments linked to future events are bought and sold by an automated market maker. We show that it is impossible for such a market maker to simultaneously achieve bounded worst-case loss and differential privacy, without allowing the privacy guarantee to degrade extremely quickly as the number of trades grows.

Chapter 2

BACKGROUND ON DIFFERENTIAL PRIVACY

This chapter introduces the relevant terms, tools, and related work from differential privacy that will be used throughout this thesis. We begin in Section 2.1 with formal definitions of differential privacy and joint differential privacy. Section 2.2 presents the properties which make these privacy notions desirable for practical use. Section 2.3 provides a diverse algorithmic toolkit of differentially private mechanisms from the existing privacy literature. For a more comprehensive discussion and additional topics related to differential privacy, the interested reader is referred to Dwork and Roth [34] for a textbook treatment.

2.1 Definitions

We begin with the classic definition of *differential privacy*, introduced by Dwork et al. [36]. In the settings considered here, a database $D \in \mathcal{D}^n$ consists of data from n individuals. We say that two databases are *neighboring* if they differ in at most one entry.

Definition 1 (Differential Privacy [36])**.** *A mechanism* $\mathcal{M} : \mathcal{D}^n \rightarrow \mathcal{R}$ *is* (ϵ,δ)-differentially private *if for every pair of neighboring databases* $D, D' \in \mathcal{D}^n$*, and for every subset of possible outputs* $\mathcal{S} \subseteq \mathcal{R}$*,*

$$\Pr[\mathcal{M}(D) \in \mathcal{S}] \leq \exp(\epsilon)\Pr[\mathcal{M}(D') \in \mathcal{S}] + \delta.$$

If $\delta = 0$*, we say that* $\mathcal{M}$ *is* ϵ-differentially private.

Intuitively, differential privacy bounds the maximum amount that one person's data can change the output of a computation. The parameter ϵ quantifies the privacy guarantee provided by the mechanism: $\epsilon = 0$ provides perfect privacy, as the output must be independent of the input. At the other extreme, $\epsilon = \infty$ guarantees no privacy, as it imposes no binding constraints, and the output of the mechanism is allowed to depend arbitrarily on a single entry in the database. For $\epsilon < 1$, the multiplicative e^ϵ guarantee can be approximated by $(1 + \epsilon)$.

In the context of mechanism design, differential privacy is often too strong of a notion. Suppose, for example, that the mechanism $\mathcal{M}$ outputs a vector of prices that each of n players will pay based on their joint input. While we may want the price that player i pays to be differentially private in the input of the other players, it is natural to allow it to be more sensitive to changes in i's own input. To capture this idea, Kearns et al. [64] defined the notion of *joint differential privacy*. We'll say that two databases are *i-neighbors* if they differ only in the i-th entry, and let $\mathcal{M}(D)_{-i}$ denote the vector of outputs to all players except player i.

Definition 2 (Joint Differential Privacy [64])**.** *A mechanism $\mathcal{M} : \mathcal{D}^n \to \mathcal{R}^n$ is* (ϵ, δ)-jointly differentially private *if for every $i \in [n]$, for every pair of i-neighbors $D, D' \in \mathcal{D}^n$, and for every subset $\mathcal{S} \subseteq \mathcal{R}^{n-1}$,*

$$\Pr[\mathcal{M}(D)_{-i} \in \mathcal{S}] \leq \exp(\epsilon)\Pr[\mathcal{M}(D')_{-i} \in \mathcal{S}] + \delta.$$

If $\delta = 0$, we say that $\mathcal{M}$ is ϵ-jointly differentially private.

Intuitively, joint differential privacy guarantees that the output to all other players excluding player i is insensitive to i's input. It protects the privacy of player i from arbitrary coalitions of other players; even if all other players shared their portion of the output, they would still not be able to infer much about player i's input.

One useful tool for proving joint differential privacy is the *billboard lemma* of Hsu et al. [62]. The idea behind the billboard lemma is quite intuitive and simple. Imagine that a mechanism $\mathcal{M}$ first computes an (ϵ, δ)-differentially private signal based on the data, and then displays that signal publicly to all n agents, as if posted on a billboard. If each player i's portion of the output, $\mathcal{M}(D)_i$, is computable from only this public signal and i's own input D_i, then $\mathcal{M}$ is (ϵ, δ)-jointly differentially private.

Lemma 1 (Billboard Lemma [62])**.** *Suppose $\mathcal{M} : \mathcal{D}^n \to \mathcal{R}$ is (ϵ, δ)-differentially private. Consider any collection of functions $f_i : \mathcal{D}_i \times \mathcal{R} \to \mathcal{R}'$, for $i \in [n]$, where $\mathcal{D}_i$ is the portion of the database containing i's data. Then the composition $\{f_i\left(\prod_i(D_i), \mathcal{M}(D)\right)\}$ is (ϵ, δ)-jointly differentially private, where $\prod_i : \mathcal{D} \to \mathcal{D}_i$ is the projection to i's data.*

This lemma allows a designer to use existing tools from differential privacy — some of which are presented in Section 2.3 — to design jointly differentially private algorithms.

2.2 Properties of Differential Privacy

In this section, we present three properties of differential privacy that make it desirable for use as a privacy notion: post-processing, group privacy, and composition.

First, differential privacy is robust to *post-processing*: no adversary can learn addition information about the database by performing further computations on a differentially private output.

Proposition 1 (Post-processing [36]). *Let $\mathcal{M} : \mathcal{D}^n \to \mathcal{R}$ be an (ϵ, δ)-differentially private mechanism. Let $f : \mathcal{R} \to \mathcal{R}'$ be an arbitrary randomized function. Then $f \circ \mathcal{M} : \mathcal{D}^n \to \mathcal{R}'$ is (ϵ, δ)-differentially private.*

Although the promise of differential privacy is typically phrased as privacy for a single individual, it also provides privacy to arbitrary groups of players, where the level of privacy decreases linearly with the size of the group. This property is known as *group privacy*.

Proposition 2 (Group privacy [36]). *Let $\mathcal{M} : \mathcal{D}^n \to \mathcal{R}$ be an (ϵ, δ)-differentially private mechanism. Then $\mathcal{M}$ is also $(k\epsilon, k\delta)$-differentially private for groups of size k. That is, for all $D, D' \in \mathcal{D}^n$ that differ in at most k entries, and for every subset of possible outputs $\mathcal{S} \subseteq \mathcal{R}$,*

$$\Pr[\mathcal{M}(D) \in \mathcal{S}] \leq \exp(k\epsilon)\Pr[\mathcal{M}(D') \in \mathcal{S}] + k\delta.$$

Differentially private algorithms also *compose*, meaning that the privacy guarantee degrades gracefully as multiple computations are performed on the same database. This allows for the *modular* design of differentially private mechanisms; an algorithm designer can combine several simple differentially private mechanisms as building blocks in a larger, more complicated algorithm. She can then reason about the overall privacy guarantee of her mechanism by reasoning about these simple mechanisms.

We begin with *basic composition* (Theorem 1), which says that the composition of multiple (ϵ, δ)-differentially private mechanisms is also differentially private, where the ϵ and δ parameters "add up."

Theorem 1 (Basic Composition [40]). *Let $\mathcal{M}_i : \mathcal{D}^n \to \mathcal{R}_i$ be (ϵ_i, δ_i)-differentially private for $i = 1, \ldots, k$. Then the composition $\mathcal{M}_{[k]} : \mathcal{D}^n \to \mathcal{R}_1 \times \cdots \times \mathcal{R}_k$, defined as $\mathcal{M}_{[k]}(D) = (\mathcal{M}_1(D), \ldots \mathcal{M}_k(D))$ is $(\sum_{i=1}^k \epsilon_i, \sum_{i=1}^k \delta_i)$-differentially private.*

Advanced composition (Theorem 2) improves on the composition guarantees of Theorem 1 in two ways. First, it allows the ϵ parameter to degrade (asymptotically) as square root of the number of computations performed, rather than linearly, at the cost of a small increase in the δ parameter. Second, it allows for *adaptive composition*: the choice of the i-th mechanism can depend on the previous $i-1$ mechanisms and their outputs. That is, the i-th mechanism of the composition can be written in the following way: $\mathcal{M}_i : \mathcal{D}^n \times (\mathcal{M}_1 \times \cdots \times \mathcal{M}_{i-1}) \times (\mathcal{R}_1 \times \cdots \times \mathcal{R}_{i-1}) \to \mathcal{R}_i$. For a more detailed discussion of k-fold adaptive composition, see Dwork, Rothblum, and Vadhan [35].

Theorem 2 (Adaptive Composition [35])**.** *Let $\mathcal{M} : \mathcal{D}^n \to \mathcal{R}^k$ be a k-fold adaptive composition of (ϵ, δ)-differentially private mechanisms. Then $\mathcal{M}$ is $(\epsilon', k\delta + \delta')$-differentially private for*

$$\epsilon' = \epsilon \sqrt{2k \ln(1/\delta')} + k\epsilon(e^{\epsilon} - 1).$$

Theorem 2 also tells an algorithm designer how to set privacy parameters in subroutines to achieve a desired privacy guarantee for the overall mechanism. This is described in Corollary 1.

Corollary 1. *If $\mathcal{M} : \mathcal{D}^n \to \mathcal{R}^k$ is a k-fold adaptive composition of $(\epsilon / \sqrt{8k \ln(1/\delta)}, 0)$-differentially private mechanisms for any $\epsilon \leq 1$, then $\mathcal{M}$ is (ϵ, δ)-differentially private.*

2.3 Algorithmic Toolkit

In this section, we survey differentially private mechanisms that are commonly used in the privacy literature. This is not intended as a comprehensive list, but is rather a sampling of the algorithmic techniques which underlie many advanced mechanisms.

Randomized Response

Perhaps the simplest differentially private mechanism is that of *randomized response* [100]. Imagine a data analyst wanted to know what fraction of the population engaged in some embarrassing or illegal behavior. Instead of asking participants directly whether they engaged in this behavior — where they may have incentive to lie to avoid legal action or public embarrassment — the analyst can ask participants to complete the following procedure:

1. Flip a coin.

2. If **tails**, then respond truthfully.
3. If **heads**, then flip a second coin and respond "Yes" if heads and "No" if tails.

The privacy of this mechanism comes from the *plausible deniability* it provides. If a participant responds "Yes," then the analyst (or any other observer) cannot be sure whether the participant actually engaged in the behavior, or whether both coins came up heads and this was a random answer. The analyst can still estimate the total number of people engaging in the behavior as twice the number of "Yes" responses minus 1/2, due to the way noise was added.

This simple mechanism demonstrates the essential goal of differential privacy: to allow inferences about a large population, while protecting the privacy of every individual. The version of randomized response described above is $(\ln 3, 0)$-differentially private. Other values of ϵ can be achieved by varying the bias of the coins.

Laplace Mechanism

The Laplace Mechanism is useful for answering numeric queries to a database: functions $f : \mathcal{D}^n \to \mathbb{R}$ that map databases to a real number. Examples of these types of queries are "How many people in the database have blue eyes?" or "What fraction of the entries in the database satisfy property P?"

The Laplace Mechanism first computes the true value of f on the input database, and then adds a noise term drawn according to the *Laplace distribution*, defined below.

Definition 3 (Laplace distribution)**.** *The* Laplace distribution *(centered at 0) with scale b is the distribution with probability density function:*

$$\text{Lap}(x|b) = \frac{1}{2b}\exp\left(-\frac{|x|}{b}\right).$$

We will write Lap(b) *to denote the Laplace distribution with scale b. We may sometimes abuse notation and use* Lap(b) *to denote a random variable drawn from the the Laplace distribution with scale b.*

The *sensitivity* of a function determines the scale of noise that must be added to preserve privacy. Formally defined in Definition 4, sensitivity is the maximum change in the function's output when a single player changes her input.

Definition 4 (Sensitivity)**.** *The* sensitivity *of a function $f : \mathcal{D}^n \to \mathbb{R}$ is:*

$$\Delta f = \max_{D,D',\, neighbors} |f(D) - f(D')|.$$

We now define the *Laplace Mechanism*, first introduced by Dwork et al. [36], which computes the value of f on the input database D, and adds Laplace noise scaled to the sensitivity of f divided by ϵ.

Definition 5 (Laplace Mechanism [36])**.** *Given any function $f : \mathcal{D}^n \to \mathbb{R}$, the* Laplace Mechanism *is defined as:*

$$\mathcal{M}_L(D, f(\cdot), \epsilon) = f(D) + Y,$$

where Y is drawn from $\text{Lap}(\Delta f/\epsilon)$.

Theorem 3 ([36])**.** *The Laplace Mechanism is $(\epsilon, 0)$-differentially private.*

Proof. Let $D, D' \in \mathcal{D}^n$ be neighboring databases, and let $f : \mathcal{D}^n \to \mathbb{R}$ be an arbitrary function.

$$\begin{aligned} \frac{\Pr[\mathcal{M}_L(D, f(\cdot), \epsilon) = x]}{\Pr[\mathcal{M}_L(D', f(\cdot), \epsilon) = x]} &= \frac{\exp\left(-\frac{\epsilon|f(D)-x|}{\Delta f}\right)}{\exp\left(-\frac{\epsilon|f(D')-x|}{\Delta f}\right)} \\ &= \exp\left(\frac{\epsilon\left(|f(D') - x| - |f(D) - x|\right)}{\Delta f}\right) \\ &\leq \exp\left(\frac{\epsilon\left(|f(D) - f(D')|\right)}{\Delta f}\right) \\ &\leq \exp(\epsilon) \end{aligned}$$

The first inequality comes from the triangle inequality and the second inequality comes from the definition of sensitivity. □

The Laplace Mechanism also works for vector-valued queries: functions of the form $f : \mathcal{D}^n \to \mathbb{R}^k$. In this case, the sensitivity of a function should be defined with respect to the appropriate vector norm (typically the ℓ_1-norm), and the mechanism outputs $\mathcal{M}_L(D, f(\cdot), \epsilon) = f(D) + (Y_1, \ldots, Y_k)$, where the Y_i are drawn i.i.d. from $\text{Lap}(\Delta f/\epsilon)$. Other variants of the Laplace Mechanism add noise that is correlated across the k dimensions through a high-dimensional analog of the Laplace distribution. See, e.g., Chaudhuri, Monteleoni, and Sarwate [14], Bassily, Smith, and Thakurta [4], or Section 6.4 of this thesis for more details.

The following theorem provides an accuracy guarantee on the output of the Laplace Mechanism.

Theorem 4. *Let $f : \mathcal{D}^n \to \mathbb{R}^k$. Then for all $D \in \mathcal{D}^n$, all $\epsilon \in \mathbb{R}^+$, and all $\beta \in (0,1]$,*

$$\Pr\left[\|f(D) - \mathcal{M}_L(D, f(\cdot), \epsilon)\|_\infty \geq \ln\left(\frac{k}{\beta}\right) \cdot \left(\frac{\Delta f}{\epsilon}\right)\right] \leq \beta,$$

where Δf is the ℓ_1 sensitivity of f.

Example: Counting Queries Counting queries are queries of the form "How many entries in the database satisfy property P?". They also include generalizations such as the fractional variant ("What fraction of the entries in the database satisfy property P?") and weighted counting queries with weights in $[0,1]$ (i.e., linear queries). Since changing one entry in the database can change the output of a counting query by at most 1, then the sensitivity of these queries is 1, and the Laplace Mechanism adds noise drawn from $\text{Lap}(1/\epsilon)$.

Answering k fixed counting queries can be viewed as a vector-valued query with a k-dimensional output. In this case, the ℓ_1-sensitivity of the query is k, since changing one entry in the database can change the value of each query by at most 1, and thus can change the ℓ_1 norm of the output vector by at most k. To achieve $(\epsilon, 0)$-differential privacy, the Laplace Mechanism should add noise drawn from $\text{Lap}(k/\epsilon)$ to the answer of each query. This can also be seen from Basic Composition.

Example: Histogram Queries A special case of vector-valued counting queries is *histogram queries*, where the data universe $\mathcal{D}$ is partitioned into disjoint cells, and each query counts how many entries fall into a particular cell. Examples of histogram queries include disjoint age brackets or income brackets. In this case, the sensitivity of the function is 1 because changing a single entry can only change the count in one of the cells,[1] so the Laplace Mechanism can guarantee $(\epsilon, 0)$-differential privacy by adding i.i.d. noise draws from $\text{Lap}(1/\epsilon)$ to each dimension of the output.

Example: Gaussian Mechanism The Laplace distribution is particularly well-suited for differential privacy, but one might wonder if other noise distributions would work as well. The *Gaussian Mechanism* is analogous to the Laplace Mechanism, and adds Gaussian noise with mean 0 and variance $\Delta f \ln(1/\delta)/\epsilon$ to the true answer of the query, and guarantees (ϵ, δ)-differential privacy. Due to the tails of

[1] The philosophical question of whether "changing a single entry" means adding or removing an entry, or changing the content of an entry is ignored here. It only affects the privacy guarantees up to a factor of 2.

the Gaussian distribution, the Gaussian Mechanism is not able to achieve $(\epsilon, 0)$-differential privacy. For more details see Appendix A of Dwork and Roth [34].

Exponential Mechanism

The *Exponential Mechanism* [73] is a powerful private mechanism for non-numeric queries with an arbitrary range, such as selecting the best outcome from a set of alternatives. Examples of such queries include "What is the most common eye color in the database?" and "What price should I post to maximize revenue?" The quality of an outcome is measured by a *score function*, relating each alternative to the underlying data. A score function $q \colon \mathcal{D}^n \times \mathcal{R} \to \mathbb{R}$ maps each database and outcome pair to a real-valued score. The quality of an outcome — an eye color or a price, in the example queries given above — will depend on the database, and we assume the analyst wishes to select an outcome with high quality score.

The *sensitivity* of the score function is measured only with respect to the database argument; it can be arbitrarily sensitive in its range argument:

$$\Delta q = \max_{r \in \mathcal{R}} \max_{D,D', \, neighbors} |q(D,r) - q(D',r)|.$$

Definition 6 (Exponential Mechanism [73])**.** *Given a quality score $q : \mathcal{D}^n \times \mathcal{R} \to \mathbb{R}$, the* Exponential Mechanism *is defined as:*

$$\mathcal{M}_E(D, q, \epsilon) = \textit{output } r \in \mathcal{R} \textit{ with probability proportional to } \exp\left(\frac{\epsilon q(D,r)}{2\Delta q}\right).$$

The Exponential Mechanism samples an output from the range $\mathcal{R}$ with probability exponentially weighted by score. Outcomes with higher scores are exponentially more likely to be selected, thus ensuring both privacy and a high quality outcome.

Theorem 5 ([73])**.** *The Exponential Mechanism is $(\epsilon, 0)$-differentially private.*

Proof. Let $D, D' \in \mathcal{D}^n$ be neighboring databases, and let $r \in \mathcal{R}$ be an arbitrary element of the output range.

$$
\begin{aligned}
\frac{\Pr[\mathcal{M}_E(D,q,\mathcal{R}) = r]}{\Pr[\mathcal{M}_E(D',q,\mathcal{R}) = r]} &= \frac{\left(\frac{\exp\left(\frac{\epsilon q(D,r)}{2\Delta q}\right)\mu(r)}{\sum_{r'\in\mathcal{R}}\exp\left(\frac{\epsilon q(D,r')}{2\Delta q}\right)\mu(r')}\right)}{\left(\frac{\exp\left(\frac{\epsilon q(D',r)}{2\Delta q}\right)\mu(r)}{\sum_{r'\in\mathcal{R}}\exp\left(\frac{\epsilon q(D',r')}{2\Delta q}\right)\mu(r')}\right)} \\
&= \left(\frac{\exp\left(\frac{\epsilon q(D,r)}{2\Delta q}\right)}{\exp\left(\frac{\epsilon q(D',r)}{2\Delta q}\right)}\right) \cdot \left(\frac{\sum_{r'\in\mathcal{R}}\exp\left(\frac{\epsilon q(D',r')}{2\Delta q}\right)\mu(r')}{\sum_{r'\in\mathcal{R}}\exp\left(\frac{\epsilon q(D,r')}{2\Delta q}\right)\mu(r')}\right) \\
&= \exp\left(\frac{\epsilon\left(q(D,r) - q(D',r)\right)}{2\Delta q}\right) \cdot \left(\frac{\sum_{r'\in\mathcal{R}}\exp\left(\frac{\epsilon q(D',r')}{2\Delta q}\right)\mu(r')}{\sum_{r'\in\mathcal{R}}\exp\left(\frac{\epsilon q(D,r')}{2\Delta q}\right)\mu(r')}\right) \\
&\leq \exp\left(\frac{\epsilon}{2}\right) \cdot \exp\left(\frac{\epsilon}{2}\right) \cdot \left(\frac{\sum_{r'\in\mathcal{R}}\exp\left(\frac{\epsilon q(D,r')}{2\Delta q}\right)\mu(r')}{\sum_{r'\in\mathcal{R}}\exp\left(\frac{\epsilon q(D,r')}{2\Delta q}\right)\mu(r')}\right) \\
&= \exp(\epsilon)
\end{aligned}
$$

□

Theorem 6 says that the probability of outputting a "bad" outcome decays exponentially quickly in the distance from the optimal output.

Theorem 6 ([73])**.** *Let $r \in \mathcal{R}$ be the output of $\mathcal{M}_E(D,q,\epsilon)$. Then:*

$$\Pr\left[q(D,r) \leq \max_{r'\in\mathcal{R}} q(D,r') - \frac{2\Delta q\,(\ln|\mathcal{R}| + t)}{\epsilon}\right] \leq e^{-t}.$$

Or equivalently:

$$\Pr\left[q(D,r) \geq \max_{r'\in\mathcal{R}} q(D,r') - \frac{2\Delta q\,(\ln(|\mathcal{R}|/\beta))}{\epsilon}\right] \geq 1 - \beta.$$

Above Noisy Threshold

The Above Noisy Threshold Mechanism, first introduced by Dwork et al. [39], takes in a stream of queries and halts after finding the first query with a (noisy) answer above a given (noisy) threshold. It preserves privacy by adding Laplace noise to both the threshold and the answer of each query in the stream.

This mechanism — and its more advanced cousin, the Sparse Vector Mechanism, described in the next subsection — is especially useful if an analyst is facing a stream of queries and believes that only a small number of the queries will have large answers. These mechanisms allow the analyst to identify and answer only the

Algorithm 1 Above Noisy Threshold, **AboveThreshold**$(D, \{f_i\}, T, \epsilon)$

```
Input: A database $D$, an adaptively chosen stream of sensitivity 1 queries $\{f_i\}$, a threshold $T$, a privacy parameter $\epsilon$
Output: A stream of answers $\{a_i\}$
  Let $\hat{T} = T + \text{Lap}\left(\frac{2}{\epsilon}\right)$
  For each query $f_i$ do
    Let $\nu_i = \text{Lap}\left(\frac{4}{\epsilon}\right)$
    If $f_i(D) + \nu_i \geq \hat{T}$ then
      Output $a_i = \top$
      Halt
    Else
      Output $a_i = \bot$
```

"important" queries, without having to incur privacy cost proportional to all queries in the stream.

Theorem 7. *Above Noisy Threshold is* $(\epsilon, 0)$*-differentially private.*

Definition 7 gives an accuracy notion for **AboveThreshold**. It requires that the mechanism provides an output for all k queries, and that its output is approximately correct for all queries with high probability.

Definition 7 (Accuracy)**.** *A mechanism that outputs a stream of answers* $\{a_i\} \in \{\top, \bot\}^*$ *to a stream of k queries* $\{f_i\}$ *is* (α, β)-accurate *with respect to a threshold T if with probability at least* $1 - \beta$*, the mechanism does not halt before* f_k*, and for all* $a_i = \top$*:*

$$f_i(D) \geq T - \alpha,$$

and for all $a_i = \bot$*:*

$$f_i(D) \leq T + \alpha.$$

Theorem 8. *For any sequence of k sensitivity 1 queries* $f_1, \ldots, f_k$ *such that* $|\{i < k : f_i(D) \geq T - \alpha\}| = 0$*, then* **AboveThreshold** *is* (α, β)*-accurate for:*

$$\alpha = \frac{8\left(\ln k + \ln(2/\beta)\right)}{\epsilon}.$$

Note that the qualifier $|\{i < k : f_i(D) \geq T - \alpha\}| = 0$ in the statement of Theorem 8 requires that the only query close to being above threshold is possibly the last one. Without this requirement, the algorithm would be required to halt before the k-th query with high probability, and thus could not satisfy the accuracy guarantee.

Sparse Vector Mechanism

The *Sparse Vector Mechanism* (introduced in Dwork et al. [39], refined in Hardt and Rothblum [56], and abstracted into its current form in Dwork and Roth [34]) takes in a stream of sensitivity 1 queries $\{f_i\}$ on a database D, and a threshold T. The mechanism only outputs answers to those queries with (noisy) answers above the (noisy) threshold, and reports that all other queries were below threshold. There is also an upper bound c on the number of queries that can be answered. The mechanism will halt once it has seen more than c queries with answers above the noisy threshold.

Algorithm 2 Sparse Vector Mechanism, **Sparse**$(D, \{f_i\}, T, c, \epsilon, \delta)$

```
Input: A database $D$, an adaptively chosen stream of sensitivity 1 queries $\{f_i\}$, a
threshold $T$, total number of numeric answers $c$, privacy parameters $\epsilon$ and $\delta$
Output: A stream of answers $\{a_i\}$
  Let $\epsilon_1 = \frac{8}{9}\epsilon$ and $\epsilon_2 = \frac{2}{9}\epsilon$
  If $\delta = 0$ let $\sigma(\epsilon) = \frac{2c}{\epsilon}$. Else let $\sigma(\epsilon) = \frac{\sqrt{32c \ln \frac{2}{\delta}}}{\epsilon}$
  Let $\hat{T}_0 = T + \text{Lap}\left(\sigma(\epsilon_1)\right)$
  Let count = 0
  For each query $f_i$ do
    Let $\nu_i = \text{Lap}\left(2\sigma(\epsilon_1)\right)$
    If $f_i(D) + \nu_i \geq \hat{T}_{\text{count}}$ then
      Output $a_i = f_i(D) + \text{Lap}\left(\sigma(\epsilon_2)\right)$
      Update count = count + 1 and $\hat{T}_{\text{count}} = T + \text{Lap}\left(\sigma(\epsilon_1)\right)$
    Else
      Output $a_i = \perp$
    If count $\geq c$ then
      Halt
```

The Sparse Vector Mechanism works by repeatedly running and restarting Above Noisy Threshold, up to c times. Whenever a query is selected as being above threshold, Sparse Vector answers that query via the Laplace Mechanism. The differential privacy guarantee of Sparse Vector can be analyzed via the privacy of these two subroutines and composition theorems.

Theorem 9 ([39]). *The Sparse Vector Mechanism is (ϵ, δ)-differentially private.*

Proof. When $\delta = 0$, the Sparse Vector Mechanism consists of c runs of Above Noisy Threshold, where each run is $(\frac{8}{9c}\epsilon, 0)$-differentially private, and c runs of the Laplace Mechanism, where each run is $(\frac{1}{9c}\epsilon, 0)$-differentially private. Basic Composition (Theorem 1) gives that Sparse Vector is $(\epsilon, 0)$-differentially private.

When $\delta > 0$, each run of Above Noisy Threshold is $(\frac{8}{9\sqrt{8c\ln(2/\delta)}}\epsilon, 0)$-differentially private. By Advanced Composition (Corollary 1), these runs together are $(\frac{8}{9}\epsilon, \frac{\delta}{2})$-differentially private. Each run of the Laplace Mechanism is $(\frac{1}{9\sqrt{8c\ln(2/\delta)}}\epsilon, 0)$-differentially private. By Advanced Composition, these runs together are $(\frac{1}{9}\epsilon, \frac{\delta}{2})$-differentially private. Basic Composition of these two subroutines gives that Sparse Vector is (ϵ, δ)-differentially private. □

To analyze the accuracy of **Sparse**, we adapt our accuracy notion to a setting the algorithm also outputs numeric answers to some queries. Definition 8 extends Definition 7 to additionally require that when the mechanism outputs a numeric answer, it is within α of the the true answer to that query.

Definition 8 (Numeric Accuracy)**.** *A mechanism that outputs a stream of answers* $\{a_i\} \in (\mathbb{R} \cup \{\perp\})^*$ *to a stream of k queries* $\{f_i\}$ *is* (α, β)-accurate *with respect to a threshold T if with probability at least* $1 - \beta$*, the mechanism does not halt before* f_k*, and for all* $a_i \in \mathbb{R}$*:*

$$|f_i(D) - a_i| \leq \alpha,$$

and for all $a_i = \perp$*:*

$$f_i(D) \leq T + \alpha.$$

We can analyze the accuracy of Sparse Vector using the accuracy guarantees of its subroutines.

Theorem 10 ([39])**.** *For any sequence of k queries* $f_1, \ldots, f_k$ *such that* $|\{i : f_i(D) \geq T - \alpha\}| \leq c$*, if* $\delta = 0$*, then* **Sparse** *is* (α, β)*-accurate for:*

$$\alpha = \frac{9c\left(\ln k + \ln \frac{4c}{\beta}\right)}{\epsilon}.$$

If $\delta > 0$*, then* **Sparse** *is* (α, β)*-accurate for:*

$$\alpha = \frac{9\left(\ln k + \ln \frac{4c}{\beta}\right)\sqrt{8c\ln\frac{2}{\delta}}}{\epsilon}.$$

Proof. The requirement that $f_i(D) \leq T + \alpha$ when $a_i = \perp$ is satisfied by the accuracy guarantees of Above Noisy Threshold. Instantiating Theorem 8 with $\beta = \frac{\beta}{2c}$ and $\epsilon = \frac{8}{9c}\epsilon$ if $\delta = 0$, or $\epsilon = \frac{8}{9\sqrt{8c\ln(2/\delta)}}\epsilon$ if $\delta > 0$, gives that each of the c runs of Above Noisy Threshold is $(\alpha_1, \frac{\beta}{2c})$-accurate for:

$$\alpha_1 = \frac{9c\left(\ln k + \ln \frac{4c}{\beta}\right)}{\epsilon} \quad \text{if } \delta = 0,$$

and

$$\alpha_1 = \frac{9\left(\ln k + \ln \frac{4c}{\beta}\right)\sqrt{8c\ln(2/\delta)}}{\epsilon} \quad \text{if } \delta > 0.$$

We must also show that $|f_i(D) - a_i| \leq \alpha$ if $a_i \in \mathbb{R}$. This is satisfied by the accuracy guarantee of the Laplace Mechanism. Instantiating Theorem 4 with $k = 1$, $\Delta f = 1$, $\beta = \frac{2c}{\beta}$, and $\epsilon = \frac{1}{9c}\epsilon$ if $\delta = 0$, or $\epsilon = \frac{1}{9\sqrt{8c\ln(2/\delta)}}\epsilon$ if $\delta > 0$, gives that $\Pr\left[|f_i(D) - a_i| \geq \alpha_2\right] \leq \frac{\beta}{2c}$ for each $a_i \in \mathbb{R}$, where

$$\alpha_2 = \frac{9c\ln\frac{2c}{\beta}}{\epsilon} \quad \text{if } \delta = 0,$$

and

$$\alpha_2 = \frac{9\ln\frac{2c}{\beta}\sqrt{8c\ln(2/\delta)}}{\epsilon} \quad \text{if } \delta > 0.$$

Noting that $\alpha = \max\{\alpha_1, \alpha_2\}$, and taking a union bound over the failure probabilities completes the proof. □

Chapter 3

APPLICATIONS OF PRIVACY TO ALGORITHMIC ECONOMICS

This chapter presents applications of differentially private algorithms to problems in game theory, mechanism design, and other subfields of algorithmic economics.

Section 3.1 provides a self-contained primer on game theory. Section 3.2 summarizes a body of work that uses differential privacy as a tool in mechanism design. Section 3.3 surveys the literature connecting jointly differentially private algorithms for equilibrium computation to truthfulness in games and equilibrium selection. Section 3.4 considers various models for the loss experienced by a privacy-aware player from participating in a mechanism or game. Parts of this chapter follow a survey of Pai and Roth [81] and its textbook adaptation by Dwork and Roth [34, Chapter 10].

3.1 Game Theory Basics

In this section, we review the standard game theory definitions and properties that will be used throughout this thesis. We first define the terminology of games and review equilibrium notions. We then introduce mechanism design and its desiderata. For a textbook introduction to game theory, see Nisan et al. [77, Chapter 1] or Mas-Colell, Whinston, and Green [71, Chapters 7 and 8].

Before beginning, we introduce some convenient notation. For an arbitrary vector $v = (v_1, \ldots, v_n)$, we use v_i to denote the i-th entry, use v_{-i} to denote the $n-1$ entries of v other than the i-th, and use (v_i', v_{-i}) to denote the vector created by replacing the i-th entry of v with v_i'.

Equilibrium Concepts

A *game* consists of n players, whose actions jointly determine the payoffs of all players. Each player i has a *type* $t_i \in \mathcal{T}$ describing her payoff-relevant information in the game, and can choose an *action* $a_i \in \mathcal{A}_i$. For example in an auction, a player's type could be her value for the good being sold, and her action could be her bid for the good. In the context of data privacy and markets for data, a player's type is often her data, and her action is her (possibly false) data report, so $\mathcal{T} = \mathcal{D}$

or $\mathcal{A}_i = \mathcal{D}$ or both.

Each player has a utility function $u_i : \mathcal{A}_1 \times \cdots \times \mathcal{A}_n \to \mathbb{R}$ describing her payoff in the game, where $u_i(a_i, a_{-i})$ is the payoff of player i when she chooses action a_i and the other players choose actions that form a_{-i}. Players choose their actions strategically to maximize their utility function, which is formalized as a *strategy* $\sigma_i : \mathcal{T} \to \mathcal{A}_i$, mapping from a player's type to her action chosen in the game. If σ_i is a deterministic mapping, it is a *pure strategy*; if σ_i is randomized, it is a *mixed strategy*. A strategy profile $\sigma = (\sigma_1, \ldots, \sigma_n)$ is a collection of strategies, one for each player.

We say that player i is playing an *η-best response* to σ_{-i} if her current strategy maximizes her expected utility, up to an additive η, when the other players play strategy profile σ_{-i}. A strategy profile σ is an *η-Nash equilibrium* if all players are playing an η-best response to σ. When $\eta = 0$, the approximation terms are omitted, and these are simply called a *best response* and a *Nash equilibrium*, respectively. The same naming convention applies when $\eta = 0$ in any of the other equilibrium notions defined in this chapter.

Definition 9 (Nash equilibrium)**.** *A strategy profile σ forms an η-*Nash equilibrium *of the game defined by type vector t, if for every player i, and for all $a_i \in \mathcal{A}_i$,*

$$\mathbb{E}_{\sigma}[u_i(\sigma(t))] \geq \mathbb{E}_{\sigma}[u_i(a_i, \sigma_{-i}(t_{-i}))] - \eta.$$

A mixed strategy Nash equilibrium is guaranteed to exist in every finite game [76]. That is, games with a finite number of players n and finite action spaces $\mathcal{A}_i$ for each player.

If a strategy σ_i is an η-best response to all possible σ_{-i}, we say that it is a *η-dominant strategy*. A strategy profile σ is an *η-dominant strategy equilibrium* if all players are playing an η-dominant strategy. This is stronger equilibrium notion than a Nash equilibrium because it requires a player's strategy to be optimal for any behavior of their opponents, rather than just a fixed strategy profile. As such, many games do not have a dominant strategy equilibrium.

Definition 10 (Dominant strategy equilibrium)**.** *A strategy profile σ forms an η-*dominant strategy equilibrium *of the game defined by type vector t, if for every player i, for all $a_i \in \mathcal{A}_i$, and for all σ_{-i},*

$$\mathbb{E}_{\sigma}[u_i(\sigma(t))] \geq \mathbb{E}_{\sigma}[u_i(a_i, \sigma_{-i}(t_{-i}))] - \eta.$$

The equilibrium notions in Definitions 9 and 10 are defined for *complete information* environments, where the true type vector t is common knowledge to all players in the game.

In settings of *incomplete information*, the types of all players are not commonly known. Instead, each player has a probabilistic belief about the types of others.[1] One common incomplete information environment is the assumption of a *common prior*, where the types of each player are drawn (either jointly or i.i.d.) from a distribution that is commonly known to all players. Each player then realizes her own type and then performs a Bayesian update on her beliefs about the types of other players.

Games of incomplete information with a common prior are also known as *Bayesian games*, and have three informational stages. In the *ex-ante stage*, players only know the prior distribution over types; in the *interim stage*, players know their own type t_i, and have performed a Bayesian update to form a posterior belief about the types of others t_{-i}; in the *ex-post stage*, players have learned the types of all other players and any other previously unknown random quantities in the game.

The appropriate equilibrium notion for incomplete information environments is the *Bayes Nash equilibrium.* A strategy profile σ is an η-*Bayes Nash equilibrium* if all players are playing an η-best response in σ, where the expected utility maximization of the best response is now over both the randomness in σ and a player's interim beliefs about the types of other players.

Definition 11 (Bayes Nash equilibrium)**.** *A strategy profile σ forms an η-*Bayes Nash equilibrium *if for every player i, for all $t_i \in \mathcal{T}$, and for all $a_i \in \mathcal{A}_i$,*

$$\mathbb{E}_{t_{-i},\sigma}[u_i(\sigma(t)) \mid t_i] \geq \mathbb{E}_{t_{-i},\sigma}[u_i(a_i, \sigma_{-i}(t_{-i})) \mid t_i] - \eta.$$

Another equilibrium notion for games of incomplete information is the *ex-post Nash equilibrium*, which requires a strategy profile σ to form a Nash equilibrium for every possible realization of types (i.e., all possible complete information games) that could arise. This is a much stronger solution concept than Bayes Nash equilibrium, and it does not require players to know a prior distribution over types.

Definition 12 (Ex-post Nash equilibrium)**.** *A strategy profile σ forms an η-*ex-post Nash equilibrium *if for every type vector $t \in \mathcal{T}^n$, for every player i, and for all*

[1] These beliefs are often modeled as being a part of the player's type [57].

$a_i \in \mathcal{A}_i$,

$$\mathbb{E}_\sigma[u_i(\sigma(t))] \geq \mathbb{E}_\sigma[u_i(a_i, \sigma_{-i}(t_{-i}))] - \eta.$$

An *ex-post dominant strategy equilibrium* can be defined analogously by requiring σ to form a dominant strategy equilibrium for every possible vector of types.

Mechanism Design

The field of *mechanism design* studies the problem faced by a central designer with the power to influence the behavior of players by changing their utility functions—often through the use of payments that depend on a player's action. Formally, the designer must choose a mechanism $\mathcal{M} : \mathcal{A}^n \to \mathcal{R}$ mapping the actions of players[2] to some range $\mathcal{R}$, which describes the outcome of the game and the utilities to all players.

One example of a mechanism designer is an auctioneer who wishes to sell an item. Players report their bids, and the auctioneer must decide who wins the item and at what price, which in turn determines the utilities of all players. Another example is a government deciding where to build a hospital. It can ask players where they live and work, and then must decide where the hospital will be located. The designer may additionally have his own objective function $f : \mathcal{R} \to \mathbb{R}$ that he wishes to optimize when designing the mechanism: the auctioneer wants to maximize his revenue, and the government wants to serve the most people with the hospital.

A *direct revelation mechanism* is of the form $\mathcal{M} : \mathcal{T}^n \to \mathcal{R}$, where each player's action space is the type space of the game: $\mathcal{A}_i = \mathcal{T}$. It is without loss of generality to consider only the design of direct relation mechanisms. To see this, consider a mechanism $\mathcal{M} : \mathcal{A}^n \to \mathcal{R}$ with an arbitrary action space. A designer implementing the *Revelation Principle* could design a new mechanism $\mathcal{M}' : \mathcal{T}^n \to \mathcal{R}$ that collects type reports t, computes each player's equilibrium strategy $\sigma_i : \mathcal{T} \to \mathcal{A}_i$ from $\mathcal{M}$ on her behalf, and selects the outcome $\mathcal{M}'(t) = \mathcal{M}(\sigma(t))$. This implementation makes truthful reporting of types a Bayes Nash equilibrium [75]; see Section 3.3 for more details.

There are two primary properties that are desired of a mechanism. First, players should be incentivized to truthfully report their type. We say that a mechanism is *incentive compatible* if players can maximize their expected utility by truthfully

[2] The action spaces can still be unique to each player, but this is suppressed here for notional ease.

reporting their type. This property is also referred to as *truthfulness*. A mechanism is *individually rational* if players have a strategy that guarantees them non-negative utility. Without this property, players may prefer to not participate in the mechanism at all.

Definition 13 (Incentive Compatible)**.** *A mechanism $\mathcal{M}$ is* incentive compatible *(IC) if for every player i, for all $t_i \in \mathcal{T}$, and every possible misreport $t_i' \in \mathcal{T}$,*

$$\mathbb{E}_{\mathcal{M},\, t_{-i}}[u_i(t) \mid t_i] \geq \mathbb{E}_{\mathcal{M},\, t_{-i}}[u_i(t_i', t_{-i}) \mid t_i].$$

We say that a mechanism is *strictly incentive compatible* or *strictly truthful* if the inequality in Definition 13 does not hold with equality for any $t_i' \neq t_i$. In a strictly truthful mechanism, the unique maximum of a player's expected utility is achieved by reporting her true type t_i.

Definition 13 requires truthful reporting to maximize expected utility only when other players report truthfully as well. This can be phrased equivalently as requiring that truthtelling forms a Bayes Nash equilibrium under $\mathcal{M}$. For this reason, Definition 13 is also known as *Bayes Nash incentive compatibility* (BNIC). An alternative incentive compatibility condition is *dominant strategy incentive compatibility* (DSIC), which requires that truthful reporting maximizes expected utility regardless of the type reports of other players, and thus that truthtelling forms a dominant strategy equilibrium under $\mathcal{M}$.

The second desirable property of a mechanism is that players are incentivized to participate. A mechanism is *individually rational* if players have a strategy that guarantees them non-negative utility. Without this property, players may prefer to not participate in the mechanism at all.

Definition 14 (Individually Rational)**.** *A mechanism $\mathcal{M}$ is* individually rational *(IR) if for every player i and all $t_i \in \mathcal{T}$, there exists a type report $t_i' \in \mathcal{T}$ such that,*

$$\mathbb{E}_{\mathcal{M},\, t_{-i}}[u_i(t_i', t_{-i}) \mid t_i] \geq 0.$$

If a mechanism is both incentive compatible and individually rational, then the type report t_i' satisfying the individual rationality condition will be her true type t_i.

Definition 14 is also known as *interim individual rationality*, because it assumes player i knows her own type but does not know the types of other players. This is the most commonly used individual rationality condition, because this is typically the information available to players when they choose their action or type report. Other variants are *ex-ante individual rationality*, where player i does not know her own type so the expectation of her utility is also taken over her own type, and *ex post individual rationality*, which requires that players receive non-negative utility for every possible vector of types $t \in \mathcal{T}^n$ and for every realization of $\mathcal{M}$.

3.2 Differential Privacy as a Tool for Mechanism Design

In this section, we survey a body of work that uses differential privacy a tool in mechanism design to achieve the game theoretic desiderata defined in Section 3.1. We begin with an alternative interpretation of $(\epsilon, 0)$-differential privacy, phrased in terms of utility functions. Although Definition 15 is syntactically different from the original Definition 1, it is easily seen to be mathematically equivalent.[3]

Definition 15 (Differential Privacy, Alternative)**.** *A mechanism $\mathcal{M} : \mathcal{T}^n \to \mathcal{R}$ is* ϵ-differentially private *if for every pair of neighboring type vectors $t, t' \in \mathcal{T}^n$, and for all utility functions $u : \mathcal{R} \to \mathbb{R}_+$,*

$$\mathbb{E}_{r \sim \mathcal{M}(t)}[u(r)] \leq \exp(\epsilon)\mathbb{E}_{r \sim \mathcal{M}(t')}[u(r)].$$

Definition 15 considers the change in every possible utility function from a unilateral change in a single type report t_i. It promises that a player's utility cannot change too much from her choice to report truthfully or lie, nor from her choice to participate or opt-out of the mechanism. It also promises that if player i changes her type report, then it won't change player j's utility by too much either.

McSherry and Talwar [73] were the first to observe that differential privacy implies approximate truthfulness.

Proposition 3 ([73])**.** *If a mechanism $\mathcal{M}$ is ϵ-differentially private for $\epsilon \leq 1$, then $\mathcal{M}$ is also 2ϵ-dominant strategy incentive compatible.*

The properties of differential privacy presented in Section 2.2, combined with the truthfulness guarantee of Proposition 3, make differential privacy a desirable solution concept for games. Basic Composition (Theorem 1) implies that a composition of k mechanisms that are each ϵ-differentially private will be $2k\epsilon$-dominant strategy

[3] This equivalence only holds for $\delta = 0$, and requires the utility function u to have strictly positive output range.

incentive compatible. Group privacy (Proposition 2) implies *group strategyproofness*: even for coalitions of k agents who collude on their misreports, truthful reporting is still a $2k\epsilon$-dominant strategy. The incentive properties of other truthful mechanisms are generally not preserved under composition or collusion.

Despite these nice properties, differential privacy has one major drawback as a solution concept in games: misreporting is also an approximate dominant strategy! Differential privacy promises that a player's utility will be insensitive to her own report, so she could achieve approximately the same utility with any other non-truthful report, which intuitively cannot incentivize truthful reporting.

Nissim, Smorodinsky, and Tennenholtz [79] proposed a solution by designing mechanisms where each player's utility is a function of both the mechanism's outcome and the player's reported type: $u_i : \mathcal{R} \times \mathcal{T} \to \mathbb{R}_+$. Specifically, they considered environments where players choose a *reaction* to the outcome — such as buying a good at the price selected by the mechanism, or visiting one of several hospitals with locations selected by the mechanism — and the mechanism designer has the power to restrict reactions.

The *Punishing Mechanism* is a convex combination of the Exponential Mechanism from Section 2.3 and the *Commitment Mechanism*, which is designed to enforce strict truthfulness. The Commitment Mechanism first chooses an outcome $r \in \mathcal{R}$ uniformly at random, and then requires that each player's reaction is a best response to the outcome, according to her reported type. For example, if the mechanism was used to determine the price of a good, players would be forced to buy the good if their reported value was above the chosen price, and would not be allowed to buy the good otherwise.

Definition 16 (Punishing Mechanism [79])**.** *The* Punishing Mechanism $\mathcal{M}_P : \mathcal{T}^n \to \mathcal{R}$ *with parameter* $0 \leq q \leq 1$ *runs the Exponential Mechanism* $\mathcal{M}_E$ *with probability* $1 - q$ *and runs the Commitment Mechanism with complimentary probability* q.

The Exponential Mechanism is known to achieve a high quality outcome (Theorem 6), but suffers from the weak truthfulness guarantees of differential privacy. The Commitment Mechanism, on the other hand, is easily seen to be truthful, but is likely to choose a low quality outcome. By randomizing between the two (with an appropriate choice of q), the Punishing Mechanism enjoys both strict truthfulness and a high quality outcome: truthfulness is enforced by the threat of the Commitment Mechanism, and quality of the outcome is preserved by choosing the Exponential Mechanism most of the time (i.e., relatively large q).

The Punishing Mechanism is also differentially private because it is a composition of two differentially private mechanisms: the Exponential Mechanism by Theorem 5, and the Commitment Mechanism because its choice of outcome is independent of the input.

The following theorem summarizes the properties of the Punishing Mechanism, when the quality score used in the Exponential Mechanism is the *social welfare*, or total utility of all players, of each outcome on the input type reports.

Theorem 11 ([79]).[4] *The Punishing Mechanism is ϵ-differentially private, strictly truthful, and for sufficiently large n achieves social welfare at least*

$$\mathrm{OPT} - O\left(\sqrt{\frac{\log |\mathcal{R}|}{2\epsilon n}}\right).$$

3.3 Joint Differential Privacy and Equilibrium Computation

In this section we show that joint differential privacy (Definition 2) provides another avenue for strengthening the truthfulness guarantees of private mechanisms in game-theoretic settings. Recall that joint differential privacy requires player i's input to have only a small effect on the output to other players, but still allows i's portion of the output to be arbitrarily sensitive in her own report. This allows the mechanism designer more power to enforce truthfulness, while still preserving the desirable properties of differentially private mechanisms, such as composition and group strategyproofness.

Joint differential privacy can be brought to bear for implementing equilibrium behavior via a "mediated" direct revelation mechanism. A *mediator* $\mathcal{M} : (\mathcal{T} \cup \{\perp\})^n \to \mathcal{A}^n$ collects type reports (or the null report $\perp$) from all players, computes an equilibrium of the reported (complete information) game, and suggests back to each player an action corresponding to her part of the equilibrium. A mediator is weaker than the mechanisms of Section 3.1 and 3.2 because it does not have the power to enforce actions or outcomes. Each player is free to opt-out of the mediator (i.e., report $\perp$), to misreport her type, or to deviate from the mediator's suggested action.

[4] For simplicity of presentation, this statement ignores some minor technical conditions on the environment.

In a mediated game, the designer would like players to truthfully report their type, and then faithfully follow the suggested action of the mediator. We call this strategy *good behavior*. The following theorem shows that if the mediator computes an equilibrium under the constraint of differential privacy, then the suggested actions are jointly differentially private (by the Billboard Lemma), and good behavior forms an ex-post Nash equilibrium of the mediated game.

Theorem 12 ([87]). *Let $\mathcal{M} : (\mathcal{T} \cup \{\perp\})^n \to \mathcal{A}^n$ be a mediator satisfying (ϵ,δ)-joint differential privacy, and on any input type vector t, with probability $1 - \beta$ computes an α-pure strategy Nash equilibrium of the complete information game defined by type vector t. Then the good behavior strategy profile forms an η-ex-post Nash equilibrium of the mediated game for*

$$\eta = \alpha + 2(2\epsilon + \beta + \delta).$$

The intuition behind this result lies in the use of joint differential privacy, which ensures that if player i changes her type report, then the joint actions suggested to other players will remain approximately unchanged. Player i's best response to this action profile will be suggested by the mediator if she reports truthfully, but if she misreports then she may receive an arbitrarily worse suggestion. Further, when other players follow the good behavior strategy, then deviating from the mediator's suggestion is equivalent to deviating from a Nash equilibrium. Thus each player can maximize her expected utility by truthfully reporting her type to the mediator and faithfully following the suggested action.

Theorem 12 reduces the problem of truthful mechanism design to the problem of jointly differentially private equilibrium computation. Kearns et al. [64] was the first to show that this kind of reduction could be accomplished and gave an algorithm to privately compute correlated equilibria[5] in arbitrary *large games*, where the effect of any single player's action on the utility of others is diminishing with the number of players in the game. The private computation of correlated equilibrium turns out to give the desired reduction to a direct revelation mechanism only when the mediator has the power to verify types, i.e., players are not able to opt-out or misreport. Rogers and Roth [87] relaxed this requirement of the mediator

[5] A *correlated equilibrium* is a generalization of Nash equilibrium, where a correlation device — such as a mediator — sends a signal to each player sampled from a joint distribution, and each player's strategy is a best response conditioned upon the realization of her signal. A real-world example is a traffic light; it signals green or red to each driver, and each driver's best response is to stop if they see red and go if they see green. A Nash equilibrium is a special case of correlated equilibria, where the signal to each player is sampled independently.

by giving a mechanism to privately compute Nash equilibria in a special subclass of large games, namely unweighted congestion games. Cummings et al. [26] generalized the class of games in which we can privately compute Nash equilibria to include all large aggregative games, and additionally allowed the mediator to select a Nash equilibrium that optimizes any linear objective function, thereby solving the *equilibrium selection* problem.

This line of work [64, 87, 26] provides a more robust version of the Revelation Principle of Myerson [75], because it requires weaker informational assumptions, yet provides a stronger solution concept. The Revelation Principle enables a mediator to implement a Bayes Nash equilibrium and requires a prior on types, whereas joint differential privacy allows the mediator to implement a Nash equilibrium without the need of a prior. Additionally, since good behavior is an ex-post Nash equilibrium of the mediated game, players will be incentivized to follow this strategy even when types are arbitrary or worst-case.

It is an open question whether it's possible to extend the weak mediators of Rogers and Roth [87] and Cummings et al. [26] to the class of all large games, as Kearns et al. [64] did with strong mediators. As a step towards answering this question, Cummings et al. [25] defined the *coordination complexity* of a game to be the maximum length message the mediator must communicate to players via the Billboard Lemma in order to implement a high quality equilibrium. The authors show that any game with low coordination complexity also has a jointly differentially private algorithm for equilibrium computation — and the lower the coordination complexity, the better the utility guarantee of the private algorithm. They use the Exponential Mechanism (Definition 6) to select from the mediator's message space (i.e., all messages of length at most the coordination complexity). Recall from Theorem 6 that the utility guarantee of the Exponential Mechanism has a logarithmic dependence on the size of the output range $\mathcal{R}$, so if the message space is small then the mechanism will select a higher quality output. This result, combined with Theorem 12, further reduces the problem of truthful mechanism design to the problem of upper bounding coordination complexity. If a class of games has low coordination complexity, then the mediator can efficiently and privately find a message that will allow players to coordinate on a high quality equilibrium, and good behavior will be an ex-post Nash equilibrium of the mediated game.

3.4 Modeling Costs for Privacy

A line of work on strategic data revelation and mechanism design for privacy-aware agents, starting from Ghosh and Roth [47], uses differential privacy as a way to quantify the privacy cost incurred by a player who participates in a mechanism. A player's privacy cost is assumed to be a function of the mechanism's privacy parameter ϵ and her personal cost parameter $c_i \in \mathbb{R}_+$, which determines her sensitivity to a privacy violation. In particular, each player i has a privacy cost function $f_i(c_i, \epsilon)$ that describes the cost she incurs when her data is used in an ϵ-differentially private computation.

The game-theoretic interpretation of differential privacy in Definition 15 guarantees that a player's utility cannot change by more than a multiplicative $\exp(\epsilon)$ factor from participating in an ϵ-differentially private mechanism. For $\epsilon < 1$, this multiplicative factor can be approximated as $\approx 1 + \epsilon$, so a player can lose at most at additive $\epsilon \, \mathbb{E}_{r \sim \mathcal{M}}[u(r)]$ in her utility from participating in the mechanism. Motivated by this fact, Ghosh and Roth [47] used Assumption 1, and suggested an interpretation of c_i as the player's expected utility from the mechanism's outcome.

Assumption 1 ([47]). *The privacy cost function of each player i satisfies*

$$f_i(c_i, \epsilon) = c_i \epsilon.$$

Nissim, Orlandi, and Smorodinsky [78] subsequently relaxed Assumption 1 to an inequality: $f_i(c_i, \epsilon) \leq c_i \epsilon$. The authors pointed out that although the privacy cost of a player could be upper bounded by a linear function of ϵ, this linear function was not justified as a lower bound.

Xiao [101] proposed that a player's privacy cost should reflect the amount of information leaked about her type through her participation in the mechanism. He modeled a player's privacy cost function as the mutual information between her type and the outcome of the mechanism. Informally, the *mutual information* between two random variables (formally defined in Chapter 5, Definition 24) measures the reduction in uncertainty about the value of one variable after seeing the value of the other.

Assumption 2 ([101]). *The privacy cost function of each player i participating in mechanism $\mathcal{M}$ is*

$$f_i(c_i, \epsilon) = \mathcal{I}\left(t_i; \mathcal{M}(\sigma_i(t_i), t_{-i})\right),$$

where $\mathcal{I}(\cdot;\cdot)$ is mutual information.

This model of Xiao [101] achieves the intuitive goal of capturing how much a mechanism's publicly observable output reveals about a player's private input, but has the slight drawback that it requires a prior distribution on types for the mutual information to be well-defined. However, Nissim, Orlandi, and Smorodinsky [78] point out a paradox arising from the fact that privacy costs in Assumption 2 are strategy dependent, using the following *Rye and Wholewheat Game*.

Imagine a world with only two types of bread: Rye (R) and Wholewheat (W). A player's type t is her favorite type of bread, and the prior on types is uniform. A mechanism $\mathcal{M} : \{R, W\} \to \mathcal{R}$ takes in a player's reported type, and outputs a sandwich on her reported favorite bread. Now consider two strategies a player could employ in this game, $\sigma_{truthful}$ and σ_{random}. With the former strategy, a player always reports her true sandwich preferences; with the latter, she randomizes equally between the two possible type reports. Under Assumption 2, the privacy cost of using the truthful strategy is $\mathcal{I}\left(t; \mathcal{M}(\sigma_{truthful}(t))\right) = 1$, and the privacy cost of using the random strategy is $\mathcal{I}\left(t; \mathcal{M}(\sigma_{random}(t))\right) = 0$. However to an outside observer, these strategies are indistinguishable, because he will see each type of sandwich with probability $1/2$.

Chen et al. [18] proposed that each player additionally has an *outcome-dependent* privacy cost function g_i that measures her loss for participating in a particular instantiation of a differentially private mechanism. The authors assume (Assumption 3) that the value of g_i is upper-bounded by a function that depends on the effect that player i's report has on the mechanism's output. This assumption leverages the functional relationship between player i's data and the output of the mechanism. For example, if a particular mechanism ignores the input from player i, then her privacy cost should be 0 since her data is not used. Additionally, this model does not require a prior distribution on types.

For a mechanism $\mathcal{M} : \mathcal{D}^n \to \mathcal{R}$, define $g_i(\mathcal{M}, r, D_i, D_{-i})$ to be the privacy cost to player i for reporting $D_i \in \mathcal{D}$ when all other players report $D_{-i} \in \mathcal{D}^{n-1}$ and the output of $\mathcal{M}$ is $r \in \mathcal{R}$.

Assumption 3 ([18]).[6] *For any mechanism $\mathcal{M} : \mathcal{D}^n \to \mathcal{R}$, for all players i, for all*

[6] The assumption proposed in Chen et al. [18] allows privacy costs to be bounded by an arbitrary function of the log probability ratio that satisfies certain natural properties. We restrict to this particular functional form for simplicity, following [48].

outputs $r \in \mathcal{R}$, and for all input databases $D \in \mathcal{D}^n$,

$$g_i(\mathcal{M}, r, D_i, D_{-i}) \leq c_i \ln \left(\max_{D_i', D_i'' \in \mathcal{D}} \frac{\Pr[\mathcal{M}(D_i', D_{-i}) = r]}{\Pr[\mathcal{M}(D_i'', D_{-i}) = r]} \right).$$

Note that when a mechanism $\mathcal{M}$ is ϵ-differentially private, Assumption 3 implies that $g_i(\mathcal{M}, r, D_i, D_{-i}) \leq c_i\epsilon$ in the worst case, which returns to the model where privacy costs are bounded above by a linear function of ϵ.

The following lemma shows that if we instead consider the expected value of g_i, we can achieve a tighter bound on privacy cost that is quadratic in ϵ. At a high level, Lemma 2 says that a differentially private algorithm cannot achieve its worst-case privacy leakage ϵ on all outputs, and the privacy leakage is more like ϵ^2 on the "average" output, sampled according to the mechanism's distribution.

Lemma 2 ([35]). *In settings that satisfy Assumption 3 and for mechanisms $\mathcal{M} : \mathcal{D}^n \to \mathcal{R}$ that are ϵ-differentially private for $\epsilon \leq 1$, then for all players i with data D_i, for all inputs of other players D_{-i}, and for all possible misreports D_i' by player i,*

$$\mathbb{E}_{r \sim \mathcal{M}(D)}[g_i(\mathcal{M}, r, D_i, D_{-i})] - \mathbb{E}_{r \sim \mathcal{M}(D_i', D_{-i})}[g_i(\mathcal{M}, r, D_i', D_{-i})] \leq 2c_i\epsilon(e^{\epsilon} - 1) \leq 4c_i\epsilon^2.$$

To combine this framework with the utility model of, e.g., Ghosh and Roth [47], Nissim, Orlandi, and Smorodinsky [78], and Xiao [101], we need only to interpret $f_i(c_i, \epsilon) = \frac{1}{4}\mathbb{E}[g_i(\mathcal{M}, r, D_i, D_{-i})]$. That is, $f(c_i, \epsilon)$ is player i's expected cost for participating in the mechanism (up to a scaling constant). This interpretation motivates Assumption 4.

Assumption 4 ([18]). *The privacy cost function of each player i satisfies*

$$f_i(c_i, \epsilon) \leq c_i\epsilon^2.$$

The quadratic bound in Assumption 4 was introduced by Chen et al. [18], adopted by Ghosh et al. [48] and Cummings, Ioannidis, and Ligett [22], and is used in Chapter 6 of this thesis.

Part II

Economic Foundations of Privacy-Aware Choice

Chapter 4

THE INTERPRETABILITY OF PRIVACY-AWARE CHOICE DATA

4.1 Introduction

In this chapter we study what an observer can learn about a consumer's preferences and behavior when the consumer has concerns for her privacy and knows that she is being observed. The basic message of our results is that very little can be learned without strong assumptions on the form of the consumer's privacy preferences.

To motivate the problem under study, consider the following story. Alice makes choices on the internet. She chooses which websites to visit, what books to buy, which hotels to reserve, and which newspapers to read. She knows, however, that she is being watched. An external agent, "Big Brother" (BB), monitors her choices. BB could be a private firm like Google, or a government agency like the NSA. As a result of being watched, Alice is concerned for her privacy; and this concern affects her behavior.

Alice has definitive preferences over the things she chooses among. For example, given three political blogs a, b, and c, she may prefer to follow a. But, BB will observe such a choice, and infer that she prefers a over b and c. This is uncomfortable to Alice, because her preferences are shaped by her political views, and she does not like BB to know her views or her preferences. As a result, she may be reluctant to choose a. She may choose b instead because she is more comfortable with BB believing that she ranks b over a and c.[1]

Now, the question becomes, given observations of Alice's behavior, what can we learn about her preferences? We might conjecture that her behavior must satisfy some kind of rationality axiom, or that one could back out, or reverse-engineer, her preferences from her behavior. After all, Alice is a fully rational consumer (agent), meaning that she maximizes a utility function (or a transitive preference relation). She has a well-defined preference over the objects of choice, meaning that if she could fix what BB learns about her—if what BB learns about her were independent from her choices—then she would choose her favorite object. Further,

[1] Like Alice, 85% of adult internet users have take steps to avoid surveillance by other people or organizations, see Rainie et al. [86].

Alice's preferences over privacy likely satisfy particular structural properties. For example, she has well-defined preferences over the objects of choice, and she cares about the preference revealed by her choices: she always prefers revealing less to revealing more. In economics, preferences of this form are called separable and monotonic; and such preferences normally place strong restrictions on agents' behavior.

However, contrary to the above discussion, the results in this paper prove that nothing can be inferred about Alice's preferences once we introduce the possibility that she has concerns about privacy. No matter what her behavior, it is compatible with some concerns over privacy, i.e., she always has an "alibi" that can explain her choices as a consequence of privacy concerns. The strongest version of this result is that *all possible behaviors on the part of Alice are compatible with all possible preferences that Alice may have over objects*: postulate some arbitrary behavior for Alice, and some arbitrary preference over objects, and the two will always be compatible.

So BB's objective is hopeless. He can never learn anything about Alice's true preferences over political blogs, or over any other objects of choice. If BB tries to estimate preferences from some given choices by Alice, he finds that all preferences could be used to explain her choices. He cannot narrow down the set of preferences Alice might have, no matter what the observed behavior. The result continues to hold if BB adversarially sets up scenarios for Alice to choose from. That is, even if BB offers Alice menus of choices so as to maximize what he can learn from her behavior, the result is still that nothing can be learned.

The results in this paper have a variety of implications.

First, they motivate the use of specific parametric models of preferences over privacy. Our main result makes strong qualitative assumptions about preferences (separability, monotonicity). Given that such assumptions lack empirical bite, one should arguably turn to stronger assumptions yet. The paper proposes an additive utility function that depends on the chosen object and on what is revealed by the consumer's choices. If Alice chooses x then she obtains a utility $u(x)$ and a "penalty" $v(x,y)$ for not choosing y, for all non-chosen y, as she reveals to BB that she ranks x over y. This additive model does have restrictions for the consumer's behavior, and could be estimated given data on Alice's choices. The model is methodologically close to models used in economics to explain individual choices, and could be econometrically estimated using standard techniques. The paper discusses a test for the additive model based on a linear program.

Second, while the paper's main motivation is consumers' behavior on the internet, the results have implications for issues commonly discussed in behavioral economics. Some behavioral "anomalies" could be the consequence of the presence of an outside observer. For example (elaborating on a laboratory experiment by Simonson and Tversky [93]), consider a consumer who is going to buy a technical gadget, such as a phone or a camera. The consumer might prefer a simple camera over a more complex one which they might not know how to operate; but when presented with a menu that has a simple, an intermediate and an advanced camera, they might choose the intermediate one because they do not want to reveal to the world that they do not know how to use a complex camera. Of course, the results show that this line of reasoning may not be very useful, as anything can be explained in this fashion. The results suggest, however, that a stronger parametric model may be useful to explain various behavioral phenomena.

Third, the results explain why BB may want to be hide the fact that consumer behavior is being observed. The NSA or Google seem to dislike openly discussing that they are monitoring consumers' online behavior. One could explain such a desire to hide by political issues, or because the observers wish to maintain a certain public image, but here we point to another reason. The observations simply become ineffective when the consumer is aware that she is being observed.

Related Work

The growing attention to privacy concerns has led to a growing literature studying privacy, see Heffetz and Ligett [59] for a survey. Within this literature, an important question is how to model the preferences or utilities of privacy-aware agents in a way that describes their behavior in strategic settings.

One approach toward this goal, surveyed in Section 3.4, is to use differential privacy in mechanism design as a way to quantify the privacy loss of an agent from participating the mechanism. Within this literature, each of Ghosh and Roth [47], Nissim, Orlandi, and Smorodinsky [78], and Xiao [101] assume that the utility of a privacy-aware agent is her gain from the outcome of the interaction minus her loss from privacy leakage. Note that this is a stronger condition than separability, as defined in Section 4.3, and a weaker condition than additivity, as defined in Section 4.3. In contrast, Chen et al. [18] and Nissim, Vadhan, and Xiao [80] make the same separability assumption as used in this paper, but Chen et al. [18] allows for non-monotone privacy preferences and Nissim, Vadhan, and Xiao [80] uses a relaxed version of monotonicity.

Perhaps the model closest to ours is Gradwohl [52], which also considers privacy-aware agents with preferences over outcome-privacy pairs. However, the technical quantification of privacy is different in the two models, as Gradwohl [52] considers multiple agents engaging in a single interaction instead of multiple choices by a single agent as in our paper. Another related model is that of Gradwohl and Smorodinsky [53], where a single privacy-aware agent must make decisions and knows that her choices are being observed. The model differs from ours in that the agent has a (cardinal) utility function over outcome-privacy pairs, rather than ordinal preferences as in this paper. Further, Gradwohl and Smorodinsky [53] works in a Bayesian setting — the observer maintains a distribution over types that the agent may have, and performs a Bayesian update after each observed choice — whereas our results are in a prior-free setting. In addition, the nature of the results in both papers differ from ours. Gradwohl [52] studies implementation from a mechanism design perspective, and Gradwohl and Smorodinsky [53] studies the existence and uniqueness of equilibria between the agent and the observer. To contrast, we study the testable implications of privacy-aware preferences through the lens of revealed preference analysis.

Our paper is also related to the literature on psychological games. This line of work was initiated by Geanakoplos, Pearce, and Stacchetti [45], and studies settings where players have preferences over their opponents' beliefs about their *actions*. This makes for an interesting interaction between the endogenous actions taken by players in a game, and each player's beliefs over those actions. In our paper, in contrast, there is a single agent making decisions (hence, we are not in a game theoretic setup), and this agent cares about what an outside observer infers about her *preferences*. The outside observer does not make decisions in our paper.

More broadly, there is a literature within economics that studies privacy in other game theoretic models. For example, Daughety and Reinganum [29] study a signaling model where an agent's action may convey information about her type. In their setting, the agent may have privacy concerns, and therefore cares about what others learn about her type. Conitzer, Taylor, and Wagman [19] study a setting in which consumers make repeat purchases from the same seller, and show that price discrimination based on past purchase decisions can harm consumer welfare. In a similar setting, Taylor [94] shows that when consumers are privacy-oblivious, they suffer losses in welfare, but privacy-aware consumers may even cause firms to lower prices in equilibrium. Other studies on the relation between privacy and economics are surveyed by Acquisti, Taylor, and Wagman [2].

Another related stream of work consists of empirical studies of people making privacy-aware choices in practice. For example, Goldfarb and Tucker [50] examined the changing willingness of people to fill in personal data on income for market research surveys. The authors concluded that the recent decrease in willingness can be explained in part by an increased preference for privacy, and suggest that this desire for privacy may result from the now ubiquitous practice of data-driven price discrimination. In addition to pecuniary losses from price discrimination, simple embarrassment may also cause agents to become privacy-aware. Indeed, Goldfarb et al. [51] demonstrates empirically that people behave differently when their choices are recorded by a human being, which the authors attribute to embarrassment. These studies suggest that people do indeed have privacy-aware preferences, and thus privacy concerns should be considered when analyzing empirical data.

4.2 Modeling Privacy Preferences

The goal of this paper is to study the testable implications of choice data in a context where agents have privacy preferences. To this end, we adapt the standard conceptualization of consumer choice theory in economics (see e.g. the textbook treatments in Mas-Colell, Whinston, and Green [71] or Rubinstein [89]) to a situation where the consumer is aware of, and has preferences over, the information revealed by her choices.

The Setting

We focus on a situation where there is an outside observer (he), such as Google or the NSA, that is gathering data about the choices of a consumer (she) by observing her choices. We assume that the consumer is presented with a set of alternatives A and then makes a choice $c(A)$, which the outside observer sees. The observer then infers from this choice that $c(A)$ is preferred to all other alternatives in A.

The above parallels the classical revealed preference theory framework; however our model differs when it comes to the the behavior of the consumer, which we model as *privacy-aware*. We assume that the consumer is aware of the existence of an outside observer, and so she may care about what her choices reveal about her. Specifically, her choices are motivated by two considerations. On the one hand, she cares about the actual chosen alternative. On the other hand, she cares about what those choices reveal about her preferences over alternatives, i.e., her revealed preferences. We capture this by assuming that the consumer has preferences over pairs (x, B), where x is the chosen object and B is the information revealed about the consumer's preferences.

An important point about the setting is that the inferences made by the observer do not recognize that the consumer is privacy aware. This assumption about the observer being naive is literally imposed on the behavior of the observer, but *it is really an assumption about how the agent thinks that the observer makes inferences.* The agent thinks that the observer naively uses revealed preference theory to make inferences about her preferences. The observer, however, could be as sophisticated as any reader of this paper in how they learn about the agent's preferences. The upshot of our results is that such a sophisticated observer could not learn anything about the agent's behavior.

It is natural to go one step further and ask "What if the agent knows that the observer knows that the agent is privacy-aware?" Or, "what if the agent knows that the observer knows that the agent knows that the observer knows that the agent is privacy-aware?" The problem naturally lends itself to a discussion of the role of higher order beliefs. We formalize exactly this form of a cognitive hierarchy in Section 4.4, and we discuss how our results generalize.

Preliminaries

Before introducing our model formally, there are a few preliminaries that are important to discuss. Let $\mathbf{B}(X) = 2^{X \times X}$ denote the set of all binary preference relations on a set X and recall that a binary relation $\succeq$ is a weak order if it is complete (total) and transitive. We say that $x \succ y$ when $x \succeq y$ and it is not the case that $y \succeq x$. Finally, a linear order is a weak order such that if $x \neq y$ then $x \succ y$ or $y \succ x$.

We shall often interpret binary relations as graphs. For $B \in \mathbf{B}(X)$, define a graph by letting the vertex set of the graph be equal to X and the edge set be B. So, for each element $(x, y) \in B$, we have a directed edge in the graph from x to y. We say that a binary relation B is acyclic if there does not exist a directed path that both originates and ends at x, for any $x \in X$. The following simple result, often called Spilrajn's Lemma, is useful.

Lemma 3. *If $B \subseteq \mathbf{B}(X)$ is acyclic, then there is a linear order $\succeq$ such that $B \subseteq \succeq$.*

The Model

Given the setting described above, our goal is to characterize the testable implications of choice data, and to understand how the testable implications change when consumers are privacy-aware as opposed to privacy-oblivious. To formalize this we denote a *choice problem* by a tuple $(X, \mathcal{A}, c)$ where X is a finite set of alternatives, $\mathcal{A}$ a collection of nonempty subsets of X, and $c : \mathcal{A} \to X$ such that $c(A) \in A$ for all $A \in \mathcal{A}$.

In choice problem $(X, \mathcal{A}, c)$, the consumer makes choices for each $A \in \mathcal{A}$ according to the function c. Further, given $A \in \mathcal{A}$ and $x = c(A)$, the observer infers that the consumer prefers x to any other alternative available in A. That is, he infers that the binary comparisons (x, y) $\forall$ $y \in A \setminus \{x\}$ are part of the consumer's preferences over X. Such inferences lie at the heart of revealed preference theory (see e.g. [97] or [96]).

A *privacy preference* is a linear order $\succeq$ over $X \times 2^{X \times X}$. A privacy preference ranks objects of the form (x, B), where $x \in X$ and $B \in \mathbf{B}(X)$. If a consumer's choices are guided by a privacy preference, then she cares about two things: she cares about the choice made (i.e. x) and about what her choices reveal about her preference (hence B).

Our approach differs from the standard model in that the consumer has preferences not only over objects, but also over the choice data. Other papers have broken from the standard model to allow for preferences over menus (see, e.g., Dekel, Lipman, and Rustichini [30] and Dekel, Lipman, and Rustichini [31]) or over beliefs (see Geanakoplos, Pearce, and Stacchetti [45]).

Given the notions of a choice problem and privacy preferences defined above, we can now formally define the notion of rationalizability that we consider in this paper.

Definition 17. *A choice problem* $(X, \mathcal{A}, c)$ *is* rationalizable (via privacy preferences) *if there is a privacy preference* $\succeq$ *such that if* $x = c(A)$ *and* $y \in A \setminus \{x\}$ *then*

$$(x, \{(x, z) : z \in A \setminus \{x\}\}) \succ (y, \{(y, z) : z \in A \setminus \{y\}\}),$$

for all $A \in \mathcal{A}$. *In this case, we say that* $\succeq$ rationalizes $(X, \mathcal{A}, c)$.

This definition requires that for every observation of an element x chosen from a set A, and for every alternative $y \in A$ that was available but not chosen, the consumer prefers x *paired with the inferences* made from her choice of x, to the alternative

y paired with the *counterfactual* inferences that would have been made if she had chosen y instead. We shall sometimes use the notation $A_x = \{(x,z) : z \in A \setminus \{x\}\}$ to denote the set of binary comparisons inferred by the observer from the consumer's choice of x from set A.

Thus, a choice problem is rationalizable when there exists a privacy preference that "explains" the data, i.e., when there exists a privacy preference for which the observed choices are maximal.

4.3 The Rationalizability of Privacy-Aware Choice

In this section, we present our main results, which characterize when choice data from privacy-aware consumers is rationalizable. Our results focus on the testable implications of structural assumptions about the form of the privacy preferences of the consumer. While a consumer's preferences may, in general, be a complex combination of preferences over the choices and revealed preferences, there are some natural properties that one may expect to hold in many situations. In particular, we focus on three increasingly strong structural assumptions in the following three subsections: monotonicity, separability, and additivity.

Monotone Privacy Preferences

A natural assumption on privacy preferences is *monotonicity*, i.e., the idea that revealing less information is always better. Monotonicity of privacy preferences is a common assumption in the privacy literature, e.g., see Xiao [101] and Nissim, Orlandi, and Smorodinsky [78], but of course one can imagine situations where it may not hold, e.g., see Chen et al. [18] and Nissim, Vadhan, and Xiao [80].

In our context, we formalize monotone privacy preferences as follows.

Definition 18. *A binary relation $\succeq$ over $X \times 2^{X\times X}$ is a* monotone privacy preference *when*

(i) $\succeq$ is a linear order, and

(ii) $B \subsetneq B'$ implies that $(x,B) \succ (x,B')$.

This definition formalizes the idea that revealing less information is better. In particular, if $B \subsetneq B'$, then fewer comparisons are being made in B than in B', so (x,B) reveals less information to the observer than (x,B').

Given the above definition, the question we address is "what are the empirical implications of monotone privacy preferences?" That is, "Is monotonicity refutable via choice data?" The following proposition highlights that monotonicity is *not* refutable, so any choice data has a monotone privacy preference that explains it.

Proposition 4. *Any choice problem is rationalizable via monotone privacy preferences.*

Proof. We shall use the following notation:

$$A_x = \{(x, y) : y \in A \setminus \{x\}\}.$$

Define a binary relation E on $X \times \mathbf{B}(X)$ as follows: $(x, B)\ E\ (x', B')$ if either $x = x'$ and $B \subsetneq B'$, or $x \neq x'$ and there is $A \in \mathcal{A}$ with $x = c(A)$, $x' \in A$ and $B = A_X$ while $B' = A_{x'}$. It will be useful for our proof to think of E as the edges of a directed graph $G = (V, E)$, where $V = X \times \mathbf{B}(X)$. The edges where $x = x'$ result from the monotonicity requirement, and the edges where $x \neq x'$ result from the requirement that observed choices be rationalized. For shorthand, we will call these edges "monotone" and "rationalizing," respectively. It should be clear that any linear order that extends B (i.e any linear order $\succeq$ with $E \subseteq \succeq$) is a monotone privacy preference that rationalizes $(X, \mathcal{A}, c)$. By Lemma 3, we are done if we show that E is acyclic. To prove that E is acyclic, it is equivalent to show that the graph is acyclic.

By the definition of E, for any pair $(x, B)\ E\ (x', B')$, the cardinality of B must be at most that of B', and if $x = x'$ then the cardinality must be strictly smaller due to monotonicity. Therefore, there can be no cycles containing monotone edges.

Thus any cycle must contain only rationalizing edges $(x, B)\ E\ (x', B')$ with $x \neq x'$. Each such edge arises from some $A \in \mathcal{A}$ for which $B = A_X$ while $B' = A_{x'}$, and for each such A there is a unique $x \in A$ with $x = c(A)$. If the graph were to contain two consecutive rationalizing edges, it would contradict uniqueness of choice. Therefore there cannot be any cycles in E. □

Proposition 4 provides a contrast to the context of classical revealed preference theory, when consumers are privacy-oblivious. In particular, in the classical setting, choice behavior that violates the strong axiom of revealed preferences (SARP) is not rationalizable, and thus refutes the consumer choice model. However, when privacy-aware consumers are considered, such a refutation of monotonic preferences is impossible. Interestingly, this means that while one may believe that preferences are non-monotonic, the form of data considered in this paper does not have the power to refute monotonicity.[2]

Note that the question addressed by Proposition 4 is only whether the consumer's choice behavior is consistent with rational behavior, and is not about whether the consumer's underlying preferences over outcomes in X can be learned. In particular, these underlying preferences may not even be well defined for the general model considered to this point. We address this issue in the next section after imposing more structure on the privacy preferences.

Separable Privacy Preferences

That all choice behavior is rationalizable via monotone privacy preferences can be attributed to the flexibility provided by such preferences. Here we turn to a significant restriction on the preferences one might use in rationalizing the consumer's behavior.

It is natural to postulate that the consumer would have some underlying, or intrinsic, preferences over possible options when her choices are not observed. Indeed, the observer is presumably trying to learn the agent's preferences *over objects*. Such preferences should be well defined: if outcome x is preferred to outcome y when both are paired with the same privacy set B, then it is natural that x will always be preferred to y when both are paired with the same privacy set B', for all possible B'. This property induces underlying preferences over items in X, as well as the agent's privacy-aware preferences.

We formalize the notion of separable privacy preferences as follows.

Definition 19. *A binary relation $\geq$ over $X \times 2^{X \times X}$ is a* separable privacy preference *if it is a monotone privacy preference and additionally satisfies that for all $x, y \in X$ and $B \in \mathbf{B}(X)$,*

$$(x, B) \geq (y, B) \implies (x, B') \geq (y, B') \; \forall B' \in \mathbf{B}(X).$$

That is, whenever (x, B) is preferred to (y, B) for some preference set B, then also (x, B') is preferred to (y, B') for all other sets B'.

[2] This phenomenon is common in the consumer choice formulation of the revealed preference problem, but it comes about for completely different reasons.

Separable privacy preferences have an associated preference relation over X. If $\geq$ is a separable privacy preference, then define $\geq|_X$ as $x \geq |_X y$ if and only if $(x,B) \gtrsim (y,B)$ for all $B \in \mathbf{B}(X)$. Note that $\geq|_X$ is a linear order over X. We can interpret $\geq|_X$ as the projection of $\geq$ onto X.

There are two questions we seek to answer: "What are the empirical implications of separability?" and "When can an observer learn the underlying choice preferences of the consumer?" The following proposition addresses both of these questions. Note that Proposition 5 follows from a more general result, Theorem 13, which is presented in Section 4.4.

Proposition 5. *Let $(X,\mathcal{A},c)$ be a choice problem, and let $\geq$ be any linear order over X. Then there is a separable privacy preference $\geq^*$ that rationalizes $(X,\mathcal{A},c)$ such that the projection of $\geq^*$ onto X is well defined and coincides with $\geq$, i.e., $\geq^*|_X = \geq$.*

Think of $\geq$ as a conjecture that the observer has about the agent. Proposition 5 implies that *no matter the nature of such a conjecture, and no matter what choice behavior is observed, the two are compatible.*

This proposition carries considerably more weight than Proposition 4. Separability imposes much more structure than monotonicity alone and, further, Proposition 5 says much more than simply that separability has no testable implications, or that it is not refutable via choice data. Proposition 5 highlights that the task of the observer is hopeless in this case – regardless of the choice data, there are preferences over revealed information that allow all possible choice observations to be explained.

That is, the choice data does not allow the observer to narrow his hypothesis about the consumer preferences at all. This is because the consumer always has an alibi available (in the form of preferences over revealed information) which can allow her to make the observed data consistent with any preference ordering over choices.

In some sense, our result is consistent with the idea that secrecy is crucial for observers such as the NSA and Google. If the consumer is not aware of the fact that she is being observed then the observer can learn a considerable amount from choice data, while if the consumer is aware that she is being observed then the choice data has little power (unless more structure is assumed than separability).

One way out of the negative conclusions from our result is to impose additional structure on the consumer's preferences. For example, one could require that the consumer cares more about correct inferences than about incorrect ones. We look next at a specific family of privacy-aware utility functions with an additive structure, that do have empirical content. These functions also lend themselves nicely to imposing additional assumptions on the form of preferences (such as penalizing correct inferences more than incorrect ones).

Additive Privacy Preferences

So far, we have seen that monotonicity and separability do not provide enough structure to allow choice data to have testable implications or to allow the observer to learn *anything* about consumer preferences over choices. This implies that further structure must be imposed for choice data to have empirical power. To that end, we now give an example of a model for privacy preferences where choice data does have testable implications. The model we consider builds on the notion of separable privacy preferences and additionally imposes additivity.

Definition 20. *A binary relation* $\succeq$ *over* $X \times 2^{X \times X}$ *is an* additive privacy preference *if there are functions* $u : X \to \mathbb{R}^+$ *and* $v : X \times X \to \mathbb{R}^+$ *such that* $(x, B) \succ (x', B')$ *iff*

$$u(x) - \sum_{(z,z') \in B} v(z, z') > u(x') - \sum_{(z,z') \in B'} v(z, z').$$

While monotonicity and separability are general structural properties of privacy preferences, the definition of additivity is much more concrete. It specifies a particular functional form, albeit a simple and natural one. In this definition, the consumer experiences utility $u(x)$ from the choice made and disutility $v(x, y)$ from the privacy loss of revealing that $x \succ y$ for every pair $(x, y) \in X \times X$. Note that this form is an additive extension of the classical consumer choice model, which would include only u and not v.

Moreover, this definition also satisfies both monotonicity and separability, making it a strictly stronger restriction. Monotonicity is satisfied because the agent always experiences a *loss* from each preference inferred by the observer. Namely, the range of v is restricted to non-negative reals, so for a fixed choice element, the agent will always prefer fewer inferences to be made about her preferences.[3] Separability is satisfied because utilities u determine the linear ordering over X, so for a fixed

[3] Monotonicity restricts to the case where people want to keep their preferences private. It may be interesting to explore in future work, the case where people are happy to reveal their information, e.g., conspicuous consumption. Under additive preferences, this would correspond to allowing the range of v to be all of $\mathbb{R}$.

set of inferences made by the observer, privacy preferences will correspond to the preferences determined by u.

Of course there are a number of variations of this form that could also make sense, e.g., if the disutility from a revealed preference (x, y) was only counted once instead of (possibly) multiple times due to multiple revelations in the choice data. This would correspond to a consumer maximizing a "global" privacy loss rather than optimizing online for each menu. However, this modeling choice requires the agent to know ex ante the set $\mathcal{A}$ of menus from which she will choose, and additional assumptions about the order in which the she faces these menus. For our analysis we restrict to additive preferences as defined above.

Rationalizability of additive privacy preferences corresponds to the existence of functions u and v, such that the observed choice behavior maximizes the consumer's utility under these functions. Here, it turns out the imposed structure on privacy preferences is enough to allow the model to have testable implications, as shown in the following proposition.

Proposition 6. *There exists a choice problem $(X, \mathcal{A}, c)$ that is not rationalizable with additive privacy preferences.*

Proposition 6 highlights that, while monotonicity and separability cannot be refuted with choice data, additivity can be refuted. To show this, we construct a simple example of choice data that cannot be explained with any functions u and v.

Proof of Proposition 6. To construct an example that is not rationalizable via additive privacy preferences, we begin by defining the set of alternatives as $X = \{x, y, z, w\}$ and the choice data as follows. It includes six observations: $z = c(\{x, z\})$, $x = c(\{x, y, z\})$, $w = c(\{w, z\})$, $z = c(\{w, y, z\})$, $x = c(\{x, w\})$, $w = c(\{x, y, w\})$.

To see that this choice data is not rationalizable suppose, towards a contradiction, that the pair (u, v) rationalizes c. Then $z = c(\{x, z\})$ implies that

$$u(z) - v(z, x) > u(x) - v(x, z),$$

while $x = c(\{x, y, z\})$ implies that

$$u(z) - v(z, x) - v(z, y) < u(x) - v(x, z) - v(x, y).$$

Therefore $v(z, y) > v(x, y)$.

Similarly, we can argue that $w = c(\{w,z\})$ and $z = c(\{w,y,z\})$ together imply that $v(w,y) > v(z,y)$, and $x = c(\{x,w\})$ and $w = c(\{x,y,w\})$ together imply that $v(x,y) > v(w,y)$. This gives us a contradiction and so proves that the choice data is not rationalizable. □

Given that the structure imposed by additive privacy preferences is testable, the next task is to characterize data sets that are consistent with (or refute) the additive privacy preference model. The example given in the proof of Proposition 6 already suggests an important feature of choice data that must hold for it to be rationalizable.

Given a choice problem $(X,\mathcal{A},c)$ and an element $y \in X$, define the binary relation R^y by $x\ R^y\ z$ if there is $A \in \mathcal{A}$ with $z = c(A)$ and $x = c(A \cup \{y\})$. Our next result gives a test for additively rational preferences. It says that, if there are cycles in the binary relation R^y, then the choice data cannot be rationalized by additive privacy preferences.

Proposition 7. *A choice problem can be rationalized by additive privacy preferences only if R^y is acyclic, for all y.*

Proof. Let c be rationalizable by the additive privacy preferences characterized by (u,v). For each $x,z \in X$ such that $x\ R^y\ z$, then there is some $A \in \mathcal{A}$ such that $z = c(A)$ and $x \in A$, so

$$u(z) - \sum_{t \in A} v(z,t) > u(x) - \sum_{t \in A} v(x,t).$$

Similarly, $x = c(A \cup \{y\})$ and $z \in A \cup \{y\}$, so

$$u(z) - \sum_{t \in A} v(z,t) - v(z,y) > u(x) - \sum_{t \in A} v(x,t) - v(x,y).$$

For both inequalities to be true simultaneously, we need $v(z,y) > v(x,y)$. Thus,

$$x\ R^y\ z \Longrightarrow v(z,y) > v(x,y). \tag{4.1}$$

Now assume there exists a cycle in binary relation R^y: $a_1\ R^y\ a_2\ R^y \cdots R^y\ a_k\ R^y\ a_1$. Then by Equation (4.1), it must be that $v(a_1,y) > v(a_2,y) > \cdots > v(a_k,y) > v(a_1,y)$. In particular, $v(a_1,y) > v(a_1,y)$ which is a contradiction. Then for choices to be rationalized, acyclicity of R^y for all $y \in X$ is a necessary condition. □

Of course, one would like to develop a test for rationalizability that is both necessary and sufficient. We do this next. Unfortunately, the test we develop takes super-exponential time to even write down. This suggests that acyclicity of R^y, despite being only a necessary condition, is likely a more practical condition to use when testing for rationalizability.

To describe the test for rationalizability, first observe that when an object x is chosen from a set, the observer infers that x (with its associated privacy) is preferred to y (and its associated privacy), for all $y \in A \setminus \{x\}$. Since we have assumed these preferences to have a specific functional form as in Definition 20, the observer can also infer the corresponding inequality in terms of functions u and v. We initialize a large matrix to record the inequalities that are inferred from choice behavior, and ask if there exist values of $u(x)$ and $v(x, x')$ for all $x, x' \in X$ for which all inferred inequalities hold. If so, these values of u and v form additive privacy preferences that rationalize choices. If not, then no such preferences exist and the observed choice behavior is not rationalizable.

Remark 1. *A choice problem $(X, \mathcal{A}, c)$ is rationalizable if and only if there exists functions $u : X \to \mathbb{R}^+$ and $v : X \times X \to \mathbb{R}^+$ satisfying the matrix inequality given by Equation* (4.4)*, below.*

To explicitly state the matrix inequality, let us index the elements of $X = \{x_1, \ldots, x_n\}$. Then for each $A \in \mathcal{A}$, the agent chooses some $x_i = c(A) \in A$. By the definition of additive preferences, every $x_j \in A$ for $j \neq i$ was *not* chosen because

$$u(x_i) - \sum_{z \in A \setminus \{x_i\}} v(x_i, z) > u(x_j) - \sum_{z \in A \setminus \{x_j\}} v(x_j, z).$$

Rearranging terms gives the following inequality:

$$u(x_i) - u(x_j) + \sum_{z \in A \setminus \{x_j\}} v(x_j, z) - \sum_{z \in A \setminus \{x_i\}} v(x_i, z) > 0. \tag{4.2}$$

To record all inequalities implied by observed choices, we instantiate a matrix T with n^2 columns, where the first n columns correspond to elements $x_1, \ldots, x_n \in X$, and the remaining $n^2 - n$ columns correspond to ordered pairs (x_i, x_j) of elements in X, for $i \neq j$.[1] T will have a row for each triple (A, x_i, x_j), where $A \in \mathcal{A}$, and $x_i, x_j \in A$. If the agent is observed to choose $x_i = c(A)$, then Equation (4.2) must be true for each $x_j \in A$ for $j \neq i$. To encode this inequality for each such x_j, we fill in the row corresponding to (A, x_i, x_j) as follows: enter +1 in the i^{th} column, −1 in the j^{th} column, $+1$ in columns corresponding to pairs (x_j, z) where $z \in A$, -1 in columns corresponding to pairs (x_i, z) where $z \in A$, and zeros elsewhere.

To complete the encoding, we also instantiate a vector $\vec{u}$, which represents the values of $u(\cdot)$ and $v(\cdot,\cdot)$ evaluated on all elements of X. The first n entries of $\vec{u}$ will contain variables for $u(x_1), \ldots, u(x_n)$, and the remaining $n^2 - n$ entries will contain variables for $v(x_i, x_j)$ for $i \neq j$, in the same order in which the pairs appear in the columns of T.

Each row of the matrix product $T\vec{u}$ would equal

$$u(x_i) - u(x_j) + \sum_{z \in A \setminus \{x_j\}} v(x_j, z) - \sum_{z \in A \setminus \{x_i\}} v(x_i, z) \tag{4.3}$$

for some set $A \in \mathcal{A}$, observed choice $x_i = c(A)$, and not-chosen element $x_j \in A$. Note that Equations (4.2) and (4.3) are identical, so the observed choices are rationalizable if and only if there exists an assignment of the variables in $\vec{u}$ such that each row of $T\vec{u}$ is greater than zero. That is,

$$T\vec{u} > \vec{0}. \tag{4.4}$$

Any such $\vec{u}$ would specify functions $u : X \to \mathbb{R}^+$ and $v : X \times X \to \mathbb{R}^+$ which correspond to additive privacy preferences that are optimized by the observed choices.

4.4 Higher Order Privacy Preferences

The results we have discussed so far are predicated on the notion that the agent thinks that the observer is naive. We shall now relax the assumption of naivete. We are going to allow the agent to believe that the observer thinks that she is privacy aware.

Going back to Alice, who is choosing among political blogs, suppose that she reasons as follows. Alice may realize that her observed choices violate the strong axiom of revealed preference and therefore cannot correspond to the choices of a rational agent. This could tip off the observer to the fact that she is privacy aware. We have seen that privacy awareness is a plausible explanation for violations of the revealed preference axioms. So Alice could now be concerned about the observer's inference about her preferences over objects and over revealed preference. Perhaps she thinks that the observer will infer that she is avoiding blog a because of what it reveals about her, and that fact itself is something she does not wish be known. After all, if Alice has a preference for privacy, perhaps she has something to hide.

More generally, an agent may be concerned not only about what her behavior reveals about her preferences over X, but also about what her behavior reveals of her preferences for privacy. She may then make choices to minimize inferences the observer is able to make about her preferences for privacy, as well as her preferences over X.

To provide a model that incorporates such issues, we define a hierarchy of higher order preferences, called *level-k preferences*, where a level-k consumer is aware that the observer may make inferences about her level-$(k-1)$ privacy preferences, and has preferences over the information the observer can infer. In our construction, level-0 corresponds to the classical privacy-oblivious setting, and the setting we have considered to this point is that of a level-1 consumer (Sections 4.2 and 4.3).

The meaning of such levels should be clear. If Alice is concerned about facing an observer who makes level k inferences, then her behavior will be dictated by the level $k+1$ model. To emphasize a point we have made repeatedly, *the real observer may be as sophisticated as one wants,* but Alice thinks that the observer thinks that Alice thinks that the observer thinks that Alice thinks ... that the observer makes inferences based on revealed preferences.

Level-k Privacy Preferences

To formally define a "cognitive hierarchy" for privacy-aware consumers we use the following sequence of sets, $\mathcal{Y}^k$ for $k \geq 0$. $\mathcal{Y}^0 = X$, $\mathcal{Y}^1 = X \times \mathbf{B}(\mathcal{Y}^0)$, and let $\mathcal{Y}^k = X \times \mathbf{B}(\mathcal{Y}^{k-1})$. A level-$k$ privacy preference can then be defined as a binary relation $\succeq^k$ over $\mathcal{Y}^k = X \times \mathbf{B}(\mathcal{Y}^{k-1})$. That is, $\succeq^k$ describes preferences over pairs of objects $x \in X$ and the set of level-$(k-1)$ preferences that are revealed from the choice of x.

Given the results in Section 4.3, our focus is on monotone, separable privacy preferences, and so we can extend the notion of monotonicity discussed in Section 4.3 to level-k privacy preferences as follows.

Definition 21. *A* monotone level-k privacy preference *is a binary relation $\succeq^k$ over $\mathcal{Y}^k = X \times \mathbf{B}(\mathcal{Y}^{k-1})$ such that*

1. *$\succeq^k$ is a linear order, and*
2. *$B \subsetneq B'$ implies that $(x, B) \succ (x, B')$, for all $B, B' \in \mathbf{B}(\mathcal{Y}^{k-1})$.*

For this definition to hold for level-0, we define $\mathcal{Y}^{-1}$ to be the empty set.

Similarly, we extend the notion of separability to level-k privacy preferences as follows.

Definition 22. *A* separable level-k privacy preference *is a binary relation $\succeq^k$ over $\mathcal{Y}^k = X \times \mathbf{B}(\mathcal{Y}^{k-1})$ such that it is monotone and additionally satisfies for any $B \in \mathbf{B}(\mathcal{Y}^{k-1})$,*

$$(x, B) \succeq^k (y, B) \Longrightarrow (x, B') \succeq^k (y, B') \; \forall B' \in \mathbf{B}(\mathcal{Y}^{k-1}).$$

Given the notion of level-k privacy preferences, we need to characterize how an observer will make inferences from observed choices. Naturally, the exact information inferred will depend on the level which the observer believes the privacy preferences to be. For example, if the observer believes the consumer to have level-0 preferences, the information inferred by the observer is the set

$$A_x = \{(x, y) : y \in A \setminus \{x\}\},$$

which is a binary relation over X. So $A_x \in \mathbf{B}(\mathcal{Y}^0)$. However, if the observer believes the consumer to have level-1 preferences, the information inferred by the observer is the set

$$\{((x, A_x), (y, A_y)) : y \in A \setminus \{x\}\} \in \mathbf{B}(\mathcal{Y}^1).$$

More generally, to describe the observer's inferences under the the belief that the consumer is level-k, we introduce the following notation. Consider the functions $T^k : \mathcal{A} \times X \to \mathbf{B}(\mathcal{Y}^k)$, for $k \geq 0$. Let

$$\begin{aligned}
T^0(A, x) &= \{(x, y) : y \in A \setminus \{x\}\} \in \mathbf{B}(\mathcal{Y}^0) \\
T^1(A, x) &= \Big\{\Big((x, T^0(A, x)), (y, T^0(A, y))\Big) : y \in A \setminus \{x\}\Big\} \in \mathbf{B}(\mathcal{Y}^1) \\
\vdots \quad & \quad \vdots \\
T^k(A, x) &= \Big\{\Big((x, T^{k-1}(A, x)), (y, T^{k-1}(A, y))\Big) : y \in A \setminus \{x\}\Big\} \in \mathbf{B}(\mathcal{Y}^k).
\end{aligned}$$

In words, $T^k(A, x)$ are the level-k preferences (over alternatives in A and set of level-$(k-1)$ preferences that will be inferred from each choice) that would cause the agent to choose x from the set A. Then generally, a level-k agent making choice $x = c(A)$ must have $T^k(A, x)$ as a subset of her level-k preferences.

Example: Level-2 Privacy Preferences

In order to illustrate the cognitive hierarchy more concretely it is useful to describe the case of level-2 privacy preferences in detail. Recall that the level-0 privacy preferences are the classical setting of privacy-oblivious consumers and level-1 privacy preferences are the case we study in Sections 4.2 and 4.3. As we shall see, there is a sense in which level-2 is all that is needed.

Continuing with the story about Alice, we remarked how she could come to question her level-1 behavior because she should realize that there is something suspicious about her choices violating the revealed preference axioms. As the result of such a realization, she might entertain level-2 behavior. She might think that the observer thinks that she is level-1. Now, there is no reason for her to go any further because, in contrast with level-1, *nothing could give her away.*

While her violations of the revealed preference axioms indicate that she cannot be level-0, given our Proposition 4, nothing about her behavior could contradict that she is level-1. She has no reason to think that reasoning beyond level-2 will afford her more privacy—we have already seen that nothing in her behavior that could prove to the observer that she is not level-1.

More concretely, suppose that x is chosen from set A. The observer, who thinks the consumer is at level-1, infers the level-1 preferences

$$(x, A_x) \succ (z, A_z) \;\; \forall\, z \in A \setminus \{x\},$$

or, more specifically, that her level-1 privacy preferences correspond to the binary relation,

$$\bigcup \{[(x, A_x), (z, A_z)] : z \in A \setminus \{x\}\}\,. \tag{4.5}$$

Now the agent who believes that the observer will make such an inference, will only choose x when this choice *together with inferences revealed by the choice* is better than the choice of another alternative in A with its accompanying inferences. That is, she will choose x over y in A whenever the choices of x *together* with the release of the information in Equation (4.5) is preferred to the choice of y together with the information,

$$\bigcup \Big\{[(y, A_y), (z, A_z)] : z \in A \setminus \{y\}\Big\}\,.$$

That is, if a level 2 agent chooses x from set A, she knows that observer will make inferences according to Equation (4.5). Then her choice of x must maximize her preferences over outcomes *and* these known inferences that will be made. Specifically, she will choose x if her level-2 preferences are, for all available $y \in A$,

$$(x, \cup\{[(x, A_x), (z, A_z)] : z \in A \setminus \{x\}\}) \succ (y, \cup\{[(y, A_y), (z, A_z)] : z \in A \setminus \{z\}\}). \tag{4.6}$$

Using the notation defined earlier in this section, we can re-write Equation (4.6) as a binary relation,

$$[(x, T^1(A, x)), (y, T^1(A, y))].$$

Since the same can be said for *every* available alternative $y \in A$ that was not chosen, the following must be a part of the agent's level-2 preferences

$$T^2(A, x) = \left\{ \left[(x, T^1(A, x)), (y, T^1(A, y)) \right] : y \in A \setminus \{x\} \right\}$$

Note, however, that the observer does not get to infer $T^2(A, x)$. He believes the agent to have level-1 preferences, and upon seeing $x = c(A)$, he infers $T^1(A, x)$. This is why the agent chooses $x \in A$ to optimize her preferences over X *and* sets of the form $T^1(A, \cdot)$.

The Rationalizability of Level-k Preferences

Given the notion of a privacy-aware cognitive hierarchy formalized by level-k privacy preferences, we are now ready to move on to the task of understanding the empirical implications of higher order reasoning by privacy-aware consumers. To do this, we must first adapt the notion of rationalizability to level-k reasoning. For this, the natural generalization of Definition 17 to higher order reasoning is as follows. This definition reduces to Definition 17 when level-1 is considered, and to the classical definition of rationalizable in the privacy-oblivious case when level-0 is considered.

Definition 23. *A choice $(X, \mathcal{A}, c)$ is* level-k rationalizable *if there is a level-k privacy preference $\succeq^k \in \mathbf{B}(\mathcal{Y}^k)$ such that for all $A \in \mathcal{A}$, $T^k(A, c(A)) \subseteq \succeq^k$.*

Given this definition, we can now ask the same two questions we considered in Section 4.3 about level-k privacy preferences: "What are the empirical implications of level-k privacy preferences?" and "When can the observer learn the underlying choice preferences of consumers?" Our main result is the following theorem, which answers these questions.

Theorem 13. *Let $(X, \mathcal{A}, c)$ be a choice problem. Let $k > 0$ and $\succeq$ be any linear order over X. Then there is a monotone, separable level-k privacy preference $\succeq^*$ that level-k rationalizes $(X, \mathcal{A}, c)$ and such that:*

$$x \succeq y \textit{ iff } (x, B) \succeq^* (y, B) \textit{ for all } B \in \mathbf{B}(\mathcal{Y}^{k-1}).$$

This result implies that the conclusions of Proposition 5 are not just an anomaly due to restrictions on Alice's reasoning. Theorem 13 says that for *any* level-k at which Alice chooses to reason, Big Brother cannot test if she is behaving rationally, and cannot learn anything about her preferences over X.

Another interpretation is that Alice always has an "alibi," in terms of other preference orderings over objects that are also consistent with her choices. In the case where Big Brother deems one particular preference ordering to be the most desirable, Alice's choices can never reveal that her preferences differ from Big Brother's desired ordering, for any level of reasoning that she may use.

Proof of Theorem 13. Let $T^{k-1} : \mathcal{A} \times X \to \mathbf{B}(\mathcal{Y}^{k-1})$ be as defined in Section 4.4. For shorthand, write $\mathcal{Y}$ for $\mathcal{Y}^{k-1}$ and T for T^{k-1}. Then T describes the set of level-$(k-1)$ preferences inferred by the observer as a result of the agent's choice behavior. That is, when the agent chooses $x = c(A)$, the observer will infer all preferences in the set $T(A, x)$. Note that T is one-to-one and satisfies the following property: for all $A \in \mathcal{A}$ and all $x, x' \in A$,

$$|T(A, x)| = |T(A, x')|. \tag{4.7}$$

Property (4.7) follows because the number of pairs $\{((x, T(A, x)), (y, T(A, y))) : y \in A \setminus \{x\}\}$ is the same for any $x \in A$.

We now construct a binary relation E over $X \times \mathbf{B}(\mathcal{Y})$. As in the proof of Proposition 4, it will be useful to think of E as the edges of a directed graph $G = (V, E)$, where $V = X \times \mathbf{B}(\mathcal{Y})$. We create edges in E according to the desiderata of our privacy-aware preferences: monotone, separable, and rationalizing choice behavior. Define E as follows: $(x, B)\ E\ (x', B')$ if either (1) $x = x'$ and $B \subsetneq B'$, (2) $B = B'$ and $x \succ x'$ according to linear order $\succeq$, or (3) $x \neq x'$ and there is $A \in \mathcal{A}$ with $x = c(A)$, $x' \in A$ and $B = T(A, x)$ while $B' = T(A, x')$. We will call these edges types respectively "monotone," "separable," and "rationalizing," as a reference to the property they are meant to impose. By Lemma 3, we are done if we show that E is acyclic.

Assume towards a contradiction that there is a cycle in this graph. Then there exists a sequence $j = 1, \ldots, K$ such that

$$(x^1, B^1)\ E\ (x^2, B^2)\ E \cdots E\ (x^K, B^K) \text{ and } (x^K, B^K)\ E\ (x^1, B^1).$$

For any monotone edge $(x^i, B^i)\ E\ (x^{i+1}, B^{i+1})$, it must be the case that $|B^i| < |B^{i+1}|$ since $B^i \subset B^{i+1}$. If this were a separable edge, then $B^i = B^{i+1}$, so $|B^i| = |B^{i+1}|$. Similarly, for any rationalizing edge, $|B^i| = |T(A, x)| = |T(A, x')| = |B^{i+1}|$. Thus as we traverse any path in this graph, the size of the second component is non-increasing along all edges, and strictly decreasing along monotone edges. This implies that there can be no cycles containing monotone edges, and our assumed cycle must consist entirely of rationalizing and separable edges.

If there are two sequential rationalizing edges in this cycle, then there exists a j and some $A \in \mathcal{A}$ such that

$$(x^j, T(A, x^j))\ E\ (x^{j+1}, T(A, x^{j+1}))\ E\ (x^{j+2}, T(A, x^{j+2})),$$

where $x^j, x^{j+1}, x^{j+2} \in A$. From the first edge, $x^j = c(A)$ in some observation, and from the second edge, $x^{j+1} = c(A)$ in another observation. If $x^j \neq x^{j+1}$ then $c(A) = x^j \neq x^{j+1} = c(A)$ which contradicts the uniqueness of choice imposed by the linear ordering. If $x^j = x^{j+1}$, then $(x^j, T(A, x^j)) = (x^{j+1}, T(A, x^{j+1}))$, which implies that an element of $X \times \mathbf{B}(\mathcal{Y})$ is strictly preferred to itself, which is a contradiction. Thus no cycle can contain two sequential rationalizing edges.

If there are two sequential separable edges in our cycle, then there exists a j such that

$$(x^j, B^j)\ E\ (x^{j+1}, B^{j+1})\ E\ (x^{j+2}, B^{j+2}),$$

where $B^j = B^{j+1} = B^{j+2}$ and $x^j \succ x^{j+1} \succ x^{j+2}$. By transitivity, $x^j \succ x^{j+2}$, so there must also be a separable edge in the graph $(x^j, B^j)\ E\ (x^{j+2}, B^{j+2})$. If the cycle we have selected contains two sequential separable edges, then there must exist another cycle that is identical to the original cycle, except with the two sequential separable edges replaced by the single separable edge. Thus we can assume without loss of generality that the cycle we have selected does not contain two sequential separable edges. Note that the only time this is with some loss of generality is when there is a cycle containing only separable edges. By assumption, $\geq$ is a linear order over X and must be acyclic, so this is not possible.

Given the previous two observations, we can assume without loss of generality that the cycle $j = 1, \dots, K$ contains alternating rationalizing and separable edges. This includes the endpoints $j = K$ and $j = 1$ since they too are connected by an edge.

If there is a path in the graph containing sequential rationalizing, separable, and rationalizing edges, then there exists $A, A' \in \mathcal{A}$ and a j such that

$$(x^j, T(A, x^j)) \; E \; (x^{j+1}, T(A, x^{j+1})) \; E \; (x^{j+2}, T(A', x^{j+2})) \; E \; (x^{j+3}, T(A', x^{j+3})),$$

where $x^j, x^{j+1} \in A$, $x^{j+2}, x^{j+3} \in A'$, and choices $x^j = c(A)$ and $x^{j+2} = c(A')$ are observed. Since the middle edge is separable, it must be that $T(A, x^{j+1}) = T(A', x^{j+2})$, so $x^{j+1} = x^{j+2}$ and $A = A'$. However, this means that $(x^{j+1}, T(A, x^{j+1}))$ is strictly preferred to itself, which is a contradiction, so no such path in the graph can exist.

Since edges must alternate between rationalizing and separable, this leaves the only possible cycles to be of length two, containing one rationalizing edge and one separable edge. However, if such a cycle existed, then traversing the cycle twice would yield a path containing sequential rationalizing, separable, and rationalizing edges, which has been shown to not exist in this graph.

Thus we can conclude that E must be acyclic, which completes the proof. □

4.5 Concluding Remarks

We conclude by describing what our results mean for Alice's story, and for future research on privacy.

Alice makes choices knowing that she is being observed. She thinks that the observer uses revealed preference theory to infer her preferences. She might think that the observer is not sophisticated, and uses revealed preferences naively to infer her preferences over objects. Alternatively, she might think that the observer is sophisticated, and knows that she has preferences for privacy; in this case, the observer tries to infer (again using revealed preferences) Alice's preferences for objects and privacy.

The story of Alice, however, is more "The Matrix" than "in Wonderland." Alice believes that she is one step ahead of the observer, and makes choices taking into account what he learns about her from her choices. In reality, however, the observer is us: the readers and writers of this paper.

We are trying to understand Alice's behavior, and to infer what her preferences over objects might be. The main result of our work is that such a task is hopeless.

Any behavior by Alice is consistent with any preferences over objects one might conjecture that she has (and this is true for any degree of sophistication that Alice may have in her model of what the observer infers from her). Other observers on the internet, such as Google or the NSA, would have to reach the same conclusion.

One way out is to impose additional structure on Alice's preferences. The main result uses separability and monotonicity, which are strong assumptions in many other environments, but that is not enough. We have suggested additive privacy preferences (Section 4.3) as a potentially useful model to follow. Additive preferences do impose observable restrictions on choice, and its parameters could be learned or estimated from choice data. Privacy researchers looking to model a utility for privacy should consider the additive model as a promising candidate.

Another source of structure is the observer's possible objectives. Our model is one of *intrinsic* preferences for privacy; Alice has an innate, exogenous desire for privacy. One could instead imagine an *instrumental* preference for privacy, which arises from Alice's prediction of how the (adversarial) observer will use the information he collects about her preferences. By imposing assumptions on the observer's possible actions and objectives, one can restrict the universe of possible privacy-aware preferences. If one does not impose any assumptions on the observer's future actions, one is back in the framework of our paper, even if privacy concerns are instrumental.

Chapter 5

THE IMPACT OF PRIVACY POLICY IN EQUILIBRIUM MODELS

5.1 Introduction

As advertising becomes increasingly targeted and data driven, there is growing concern that the algorithms driving these personalization decisions can be *discriminatory*. For example, Datta, Tschantz, and Datta [28] highlighted potential gender bias in Google's advertising targeting algorithms, giving evidence that male job seekers were more likely to be shown ads for high paying jobs than female job seekers. Similarly, the FTC has expressed concern that data mining of user online behavior, which data brokers use to categorize users into categories such as "ethnic second-city struggler" and "urban scrambler", is used to selectively target users for high interest loans [90].

One tempting response to such concerns is regulation: for example, we could mandate the use of privacy technologies which would explicitly limit the amount of information advertisers could learn about users past behavior in a quantifiable way.[1] If advertisers are only able to see differentially private signals about user behavior, for example, then we can precisely quantify the amount of information that the advertiser's signal contains about the actions of the user. As we increase the level of privacy, we would naively expect to see several effects: first, the amount of information that the advertiser learns about the user should decrease. Second, the utility of the advertiser should decrease, since she is now less able to precisely target her advertisements. Finally, if the user really was experiencing disutility from the way that the advertiser had been targeting her ads, the user's utility should increase.

These expectations are not necessarily well grounded, however, for the following reason: in strategic settings, as the information content of the signal that the advertiser receives changes, he will change the way he uses the information he receives to target ads. Similarly, given the way that user behavior is used in ad targeting, a sophisticated user may change her browsing behavior. Therefore it is not enough to statically consider the effect of adding privacy technologies to a system, but instead

[1] An alternative approach to limiting the information that advertisers can collect is to explicitly try to limit how they use that information to avoid unfairness. See Dwork et al. [38] for work in this direction.

we must consider the effect in equilibrium. Each privacy level defines a different strategic interaction which results in different equilibrium behavior, and, a priori, it is not clear what effect the privacy technology will have on the equilibrium outcome.

In this paper, we consider a very simple two-stage model of advertising that may be targeted based on past user behavior. The model has three rational agents: a consumer, a seller, and an advertiser. The consumer has a value for the good being sold by the seller, and also a type which determines which of several ads the advertiser benefits most from showing the consumer. The consumer's value and type are drawn from a joint distribution, and the two are correlated—hence, information about the consumer's value for the good being sold is relevant to the advertiser when determining what ad to show her.

The game proceeds in two stages. In the first stage, the seller determines a price to set for the good he is selling—then the consumer determines whether or not she wishes to buy the good. In the second stage, the advertiser receives some signal about the purchase decision of the consumer in the first round. The signal may be noisy; the correlation of the signal with the consumer's purchase decision reflects the level of privacy imposed on the environment, and is quantified via differential privacy. As a function of the signal, the advertiser performs a Bayesian update to compute his posterior belief about the consumer's type, and then decides which ad to show the consumer. The consumer has a preference over which ad she is shown—for example, one might be for a credit card with a lower interest rate, or for a job with a higher salary. As such, the consumer does not necessarily act myopically in the first round when deciding whether to purchase the seller's good or not, and instead takes into account the effect that her purchase decision will have on the second round.

We characterize the equilibria of this model as a function of the level of differential privacy provided by the signal the advertiser receives. We show that in this model, several counter-intuitive phenomena can arise. For example, the following things may occur in equilibrium as we increase the privacy level (i.e., decrease the correlation between the user's purchase decision and the signal received by the advertiser):

1. The signal received by the advertiser can actually contain *more* information about the agent's type, as measured by mutual information (Figure 5.3). Similarly, the difference in the advertiser's posterior belief about the true type of

the agent after seeing the purchase/ did not purchase noisy bits can increase (Figure 5.2). (Interestingly, this difference does not necessarily peak at the same privacy level as the mutual information between the consumer's type and the advertiser's signal).

2. Consumer utility can decrease, and advertiser utility can increase (Figure 5.4).

3. More generally, these quantities can behave in complicated ways as a function of the privacy parameter ϵ: they are not necessarily monotone in ϵ, and can even be discontinuous, and equilibrium multiplicity can vary with ϵ (Figure 5.2, 5.6, 5.7).

Our work also gives a precise way to derive the functional form of the *value of privacy* for players (at least in our simple model), in contrast to a large prior literature surveyed in Section 3.4, that has debated the right way to impose exogenously a functional form on player privacy cost functions [47, 101, 18, 80]. In contrast to these assumed privacy cost functions, which increase as privacy guarantees are weakened, we show that players can actually sometimes have a negative marginal cost (i.e., a positive marginal utility) for *weakening* the privacy guarantees of a mechanism.

In summary, we show that even in extremely simple models, privacy exhibits much richer behavior in equilibrium compared to its static counterpart, and that decisions about privacy regulation need to take this into account. Our results serve as a call-to-arms — policy decisions about privacy technologies ought to consider *equilibrium effects*, rather than just static effects, because otherwise it is possible that the introduction of a new privacy technology or regulation could have exactly the opposite effect as was intended.

Related Literature

It is well known that when the accuracy of a differentially private mechanism must be traded off against the level of data owner participation, the "optimal" level of privacy will in general be some intermediate level between zero and full privacy, and that agents with "beneficial" types might prefer to reveal their type with fewer privacy protections [61]. This is distinct from the phenomenon that we study in this paper, in which, in equilibrium, the "value" of higher privacy levels can be appropriated by a third party who has the ability to set prices.

Blum et al. [8] study a sequential coordination game inspired by financial markets, and show that when players play un-dominated strategies, the game can have substantially higher social welfare when players are given differentially private signals of each other's actions, as compared to when they are given no signal at all. Here, like in other work, however, privacy is viewed as a constraint on the game (i.e. it is added because of its own merits), and does not increase welfare compared to the full information setting. Ghosh and Ligett [46] study a simple model of data procurement in which user costs for privacy are a function of the number of other users participating — and study the ability of a mechanism to procure data in equilibrium. In Ghosh and Ligett [46], agents have explicitly encoded values for their privacy loss. In contrast, in our model, agents do not care about privacy except insofar as it affects the payoff-relevant outcomes of the game they are playing — differential privacy in our setting is instead a parameter defining the game we analyze.

A small literature in economics and marketing has looked to understand the effect of privacy in repeated sales settings. The earliest paper is by Taylor [94], who studies a setting where buyers purchase from firm 1 in period 1 and firm 2 in period 2. The author shows that counter-intuitively, strategic consumers may prefer that their purchase decision be made public, while strategic sellers may prefer to commit to keep purchase decisions private.

More recently, Conitzer, Taylor, and Wagman [19] consider a setting where a buyer purchases twice from the same firm, and the firm cannot commit in period 1 to the future price, and may condition on the consumer's purchase decision. They consider the effect of allowing the buyer to purchase privacy, i.e., "hide" his first period purchase decision from the seller, and show that in equilibrium, having this ability may make buyers worse off (and the firm better off) than in its absence.

We build on these papers by here modeling privacy as a continuous choice variable in the spirit of differential privacy, rather than the discrete choice (purchase decision revealed or not) considered in previous papers. This allows us to analyze the quantitative effect of privacy on welfare and profit as a continuous quantity, rather than a binary effect, and in particular lets us show for the first time that *increasing* privacy protections (i.e., decreasing the correlation between the advertiser's signal and the buyer's action) can actually increase the information contained in the signal about the buyer's type.

Related in spirit is Calzolari and Pavan [11], who derive similar results in a general contracting setting. However, their paper considers direct revelation mechanisms rather than a fixed, natural mechanism such as posted prices. Therefore, in their setting, the transaction that the first principal actually conducts with the agent is disjoint from the information she learns about the agent (i.e., the agent's exact type). In our setting, like the other works cited above, the first principal's choice of posted price jointly determines both what she learns about the agent's type and the transaction that occurs among them (if at all).

Finally, the present work is part of a larger literature that uses models of games with incomplete information to understand how agents' concerns about others' beliefs about them may affect their behavior relative to the myopic optimal. Specific relevant examples include repeated contracting with persistent private information ([43], [58], [92]) and signaling ([53]). Chen, Sheffet, and Vadhan [16] connect this approach to the differential privacy literature by considering games in which players are incentivized to act in ways that are differentially private in their own types.

5.2 Model and Preliminaries

We study a two period game with a single consumer and two firms. The first firm has a single good to sell to the consumer and wishes to maximize its expected profit. The second firm is an advertiser who wishes to show a targeted ad to the consumer. We will also refer to the first firm as the *seller*, and the second firm as the *advertiser*.

1. In period 1, the consumer has a privately known value $v \in [0,1]$ for the good, drawn from a distribution with CDF F and density f. The seller posts a take-it-or-leave-it price p for the good, and the consumer makes a purchase decision. We assume that the seller's price is not observed by the advertiser.

2. In period 2, the consumer can have one of two types, t_1 and t_2, where the probability of having each type depends on her value v from period 1. Specifically, $\Pr(t_1) = g(v)$ and $\Pr(t_2) = 1 - g(v)$, for a known function $g : [0,1] \to [0,1]$. The advertiser may show the buyer one of two ads, A and B. He gets payoff s_{1A} and s_{2A} respectively from showing ad A to a consumer of type t_1 and t_2, and payoffs s_{1B} and s_{2B} from showing ad B to a consumer of type t_1 and t_2 respectively. The consumer gets additional utility δ from being shown ad A over B.[2]

[2] Since the customer's preferences in period 2 are independent of her type, it does not matter whether she knows her period 2 type from the beginning or learns it at the start of period 2.

Assumption 5. *The following two assumptions are made regarding the distribution of buyer's value and type:*

1. *The distribution of the buyer's value satisfies the non-decreasing hazard rate assumption, i.e.* $\frac{f(v)}{1-F(v)}$ *is non-decreasing in* v*.*
2. *The probability that a buyer of value* v *is of type* t_1*,* $g(v)$*, is non-decreasing in* v*.*
3. *Buyers prefer ad A, i.e.* $\delta > 0$*.*

Regarding the payoffs to the advertiser, we assume that $s_{1A} > s_{1B}$ *and* $s_{2B} > s_{2A}$*.*

The payoff assumption corresponds to the case where type t_1 is the "high type" and ad A is the "better" ad. For example, the ad could be for a credit card. The advertiser can offer either a card with a low interest rate (ad A) or high interest rate (ad B), and the consumer's purchase history may reveal his creditworthiness. The former distributional assumption is standard in mechanism design. The latter assumption amounts to saying that high-value buyers are more likely to be the ones the advertiser wants to target with the "good" ad.

The advertiser neither observes the consumer's type, nor directly observes his purchase decision. Following the consumer's decision, the advertiser will learn (partial) information about the consumer's action in the first period. We write b to denote the bit encoding the consumer's decision in the first period — that is, $b = 1$ if the consumer purchased the good in period 1, and $b = 0$ otherwise. The advertiser does not learn b exactly, but rather a noisy version $\hat{b}$ that has been flipped with probability $1 - q$, for $q \in [1/2, 1]$. The advertiser observes the noisy bit $\hat{b}$ and then performs a Bayesian update on his beliefs about the consumer's type, and displays the ad that maximizes his (posterior) expected payoff. The consumer knows that her period 1 purchase decision will affect the ad she sees in period 2. She seeks to maximize her total utility over both periods, and thus is not myopic.

The parameter q measures the correlation of the reported bit with the actual purchase decision of the consumer, which can also be quantified via differential privacy. In our setting, it is easy to translate the parameter q into an ϵ-differential privacy guarantee for the Period 1 purchase decision as follows:

$$\frac{q}{1-q} = e^{\epsilon} \iff q = \frac{e^{\epsilon}}{1+e^{\epsilon}}.$$

In the two extremal cases, full privacy ($q = 1/2$) corresponds to ϵ-differential privacy with $\epsilon = 0$, and no privacy ($q = 1$) corresponds to ϵ-differential privacy with $\epsilon = \infty$. By varying q from $1/2$ to 1, we will be able to measure changes to the equilibrium outcomes for all possible privacy levels.

5.3 Equilibrium analysis

To begin, we observe that any equilibrium must be such that the consumer follows a cutoff strategy in Period 1: there exists a marginal consumer with value v^*, such that any consumer with value $v \in [0, v^*)$ does not buy the good, and any consumer with value $v \in [v^*, 1]$ does. We formally verify this in the Appendix (Proposition 14).

Period 2

The advertiser sees $\hat{b}$, the noisy purchase decision bit, and performs a Bayesian update on its prior over types. This allows us to define the advertiser's posterior given an observed $\hat{b} = j$, for $j \in \{0, 1\}$. Recall that $\hat{b} = b$ with probability q, and $\hat{b} = 1 - b$ with probability $1 - q$.

Plugging in the probabilities for each $j \in \{0, 1\}$, we see that the advertiser's posterior when he understands the consumer is following a threshold strategy with cutoff v^* is as follows:

$$r(\hat{b} = 1, v^*, q) = \frac{(1-q)\int_0^{v^*} g(v)f(v)dv + q\int_{v^*}^1 g(v)f(v)dv}{(1-q)F(v^*) + q(1 - F(v^*))},$$

$$r(\hat{b} = 0, v^*, q) = \frac{q\int_0^{v^*} g(v)f(v)dv + (1-q)\int_{v^*}^1 g(v)f(v)dv}{qF(v^*) + (1-q)(1 - F(v^*))}.$$

Here $r(\hat{b}, v^*, q)$ is the advertiser's posterior belief that the consumer is of type t_1 after seeing the noisy bit $\hat{b}$, given that v^* is the marginal consumer in Period 1 and given the noise level q.

The advertiser wishes to maximize his expected payoff, so his Bayesian optimal decision rule will be to show ad A if and only if,

$$s_{1A}r(\hat{b}, v^*, q) + s_{2A}(1 - r(\hat{b}, v^*, q)) > s_{1B}r(\hat{b}, v^*, q) + s_{2B}(1 - r(\hat{b}, v^*, q)).$$

That is, he will show ad A if it maximizes his expected payoff. Rearranging this in terms of his posterior, he will show ad A if and only if,

$$r(\hat{b}, v^*, q) > \frac{s_{2B} - s_{2A}}{s_{1A} - s_{2A} - s_{1B} + s_{2B}} := \eta.$$

We define this fraction to be η for shorthand. Notice that η does not depend on any game parameters other than the advertiser's payoff for each outcome, s_{1A}, s_{2A}, s_{1B}, and s_{2B}. Further, by our assumptions on the ranking of these four (Assumption 5) we have ensured that $\eta \in [0, 1]$.

The following lemma will be useful throughout. It says that fixing a cutoff strategy v^*, the seller's posterior on seeing a noisy "purchased" bit is increasing as the amount of noise decreases (i.e. q increases), and similarly the seller's posterior on seeing a noisy "did not purchase" bit is decreasing.

Lemma 4. *Fixing v^*, $r(1, v^*, q)$ is increasing in q and $r(0, v^*, q)$ is decreasing in q.*

Proof. Define $\alpha_1 = \frac{\int_0^{v^*} g(v)f(v)dv}{F(v^*)}$, $\alpha_2 = \frac{\int_{v^*}^1 g(v)f(v)dv}{1-F(v^*)}$. Note that since $g(\cdot)$ is non-decreasing, $\alpha_1 \le \alpha_2$. Therefore, $r(1, v^*, q)$ can be written as:

$$r(1, v^*, q) = \frac{(1-q)F(v^*)\alpha_1 + q(1-F(v^*))\alpha_2}{(1-q)F(v^*) + q(1-F(v^*))}.$$

This is the convex combination of α_1 and α_2 with weights $\frac{(1-q)F(v^*)}{(1-q)F(v^*)+q(1-F(v^*))}$ and $\frac{q(1-F(v^*))}{(1-q)F(v^*)+q(1-F(v^*))}$ respectively. Next, note that for $q \in [1/2, 1]$ the weight on α_2 is increasing in q, and the weight on α_1 correspondingly decreasing. To see this, differentiate the weight with respect to q and observe that it is always positive. Therefore $r(1, v^*, q)$ is increasing in q.

Finally, note that $r(0, v^*, q) = r(1, v^*, 1-q)$, so the latter claim follows. □

The following proposition is an important property of the advertiser's posterior, i.e. that in any equilibrium, seeing a noisy "purchased" bit always results in a higher assessment of type t_1 than a noisy non-purchased bit.

Lemma 5. *For any period 1 cutoff value v^* and any noise level q, the advertiser's posterior probability of the consumer having type t_1 given noisy bit $\hat{b} = 1$ is higher than his posterior belief of type t_1 given noisy bit $\hat{b} = 0$. Formally, for all v^*, q, it holds that $r(1, v^*, q) \ge r(0, v^*, q)$.*

Proof. The proposition follows from Lemma 4, the fact that $r(0, v^*, q) = r(1, v^*, 1-q)$, and $q \ge 0.5$. □

In light of this, there are only three different strategies that the advertiser could use in equilibrium:

1. Show ad A to a consumer with noisy bit $\hat{b} = 1$ and ad B to consumer with noisy bit $\hat{b} = 0$. This is characterized by the following inequalities:

$$r(1,v^*,q) > \eta \quad \text{and} \quad r(0,v^*,q) < \eta \tag{5.1}$$

2. Always shows ad A, regardless of the observed noisy bit $\hat{b}$. This is optimal for the advertiser when the parameters are such that:

$$r(1,v^*,q) > \eta \quad \text{and} \quad r(0,v^*,q) > \eta \tag{5.2}$$

3. Always shows ad B, regardless of the observed noisy bit. This is optimal for the advertiser when the parameters are such that:

$$r(1,v^*,q) < \eta \quad \text{and} \quad r(0,v^*,q) < \eta \tag{5.3}$$

In the latter two cases, consumers will behave myopically in the first round because their purchase decision doesn't affect their payoff in the next round. The seller can then maximize period 1 profits by posting the monopoly price for the distribution F. Thus cases 2 and 3 can only occur when the posterior induced by the monopoly price satisfies (5.2) or (5.3).

We call the equilibrium when the advertiser follows the first strategy a *discriminatory advertising equilibrium*. The latter two are referred to as *uniform advertising equilibria A* and *B* respectively

Define the myopic monopoly price as p_M, i.e. p_M solves:

$$p_M - \frac{1 - F(p_M)}{f(p_M)} = 0. \tag{5.4}$$

The following proposition discusses existence and properties of uniform advertising equilibria.

Proposition 8. *Fixing other parameters of the game:*

1. *For $q = \frac{1}{2}$ there is either a uniform advertising equilibrium A or B, but never both.*

2. *In the former case: uniform advertising equilibria A exist for all $q \in [\frac{1}{2}, \bar{q}_2]$, where $\bar{q}_2$ is the largest solution to $r(0, p_M, q) = \eta$ in $[\frac{1}{2}, 1]$, if any. Further, there are no uniform advertising equilibria B for any q.*

3. *In the latter, conversely, uniform advertising equilibria B exist for all $q \in [\frac{1}{2}, \bar{q}_3]$, where $\bar{q}_3$ is the largest solution to $r(1, p_M, q) = \eta$ in $[\frac{1}{2}, 1]$, if any. Further there are no uniform advertising equilibria A for any q.*

Proof. At $q = 1/2$, the signal $\hat{b}$ is complete noise, and the advertiser's posterior on types will be exactly his prior. This means that the advertiser must show the same ad to all consumers, which corresponds exactly to a uniform advertising equilibrium A or B, depending on whether his prior probability of type t_1 is larger or smaller than η. The corner case where the prior exactly equals η is ignored.

Note that $r(1, p_M, \frac{1}{2}) = r(0, p_M, \frac{1}{2})$, and both are either $> \eta$ or $< \eta$. Further, by Lemma 4,$r(1, p_M, q)$ is increasing in q, while the $r(0, p_M, q)$ is decreasing in q.

Therefore the system of equations (5.2) or (5.3) can only be satisfied on some interval $[\frac{1}{2}, \bar{q}_j]$ if at all for $v^* = p_M$. □

To collect everything we have shown so far, there are three kinds of possible equilibria in this game:

1. *Discriminatory Equilibrium*: Advertiser shows ad A on seeing a noisy purchase bit and ad B and seeing a noisy non-purchase bit. The cutoffs followed by the consumer, v^* is such that (5.1) is satisfied.

2. *Uniform Advertising Equilibrium A*: Advertiser always shows ad A, regardless of bit. In this case buyer purchases myopically, and seller charges the myopic monopoly price p_M. Further, $r(\cdot)$ evaluated at v^* equaling the myopic monopoly price p_M satisfies (5.2). This equilibrium exists for all q on some interval $[\frac{1}{2}, \bar{q}_2]$, if at all.

3. *Uniform Advertising Equilibrium B*: Advertiser always shows ad B, regardless of bit. In this case buyer purchases myopically, and seller charges the myopic monopoly price p_M. Further, $r(\cdot)$ evaluated at v^* equaling the myopic monopoly price p_M satisfies (5.3). This equilibrium exists for all q on some interval $[\frac{1}{2}, \bar{q}_3]$, if at all.

By observation, the two types of uniform advertising equilibria cannot coexist in the same game.

In a uniform advertising equilibrium, the period 1 behavior is straightforward. We now finish the analysis by characterizing period 1 behavior under discriminatory advertising equilibria.

Period 1 Behavior in Discriminatory Advertising Equilibria

In this kind of equilibrium the consumer is aware that her period 1 purchasing decisions will affect the ad she sees in period 2. She will buy in period 1 if and only if her surplus from purchasing at price p plus her continuation payoff from having purchased (i.e. expected utility from the ad that will be shown) is greater than the continuation payoff from not having purchased. Formally, a consumer with value v will purchase in period 1 if the following holds:

$$(v - p) + q\delta \geq (1 - q)\delta. \tag{5.5}$$

For the marginal consumer with value v^*, this inequality must hold with equality.

$$v^* = p + (1 - 2q)\delta.$$

Define $p_1(q)$ to be the seller's optimal price charged at noise level q. Further define $v^*(q)$ as the implied cutoff type at noise level q, i.e. $v^*(q) = p_1(q) + (1 - 2q)\delta$. Note that both $p_1(q)$ and $v^*(q)$ are continuous functions of q.

Lemma 6. *Assuming a discriminatory advertising equilibrium exists for a neighborhood of $q \in [\frac{1}{2}, 1]$, the optimal price $p_1(q)$ is increasing in q while the cutoff type $v^*(q)$ is decreasing.*

Proof. From equation (5.5) above, if the seller charges a price of p in a discriminatory advertising equilibrium, then the buyer purchases if her value exceeds $p + (1 - 2q)\delta$. Therefore the seller chooses p to maximize his net profit,

$$p(1 - F(p + (1 - 2q)\delta)).$$

Differentiating with respect to p, the optimal price $(p_1(q))$ in a discriminatory advertising equilibrium solves:

$$p_1(q) - I(p_1(q) + (1 - 2q)\delta) = 0, \tag{5.6}$$

where $I(v) = \frac{1-F(v)}{f(v)}$. Applying the implicit function theorem, we see that

$$p_1'(q) - I'(p_1 + (1 - 2q)\delta)(p_1'(q) - 2\delta) = 0.$$

Therefore, since I' is negative, it follows that $p_1'(\cdot)$ must be positive. Further, we have that $p_1'(q) - 2\delta$ is negative, i.e. the cutoff value who buys at the optimal price charged is decreasing in q. □

The following proposition says that whenever a discriminatory advertising equilibrium exists at a given q, the price is higher and more customers buy than in either uniform advertising equilibrium.

Lemma 7. *For any $q \in [1/2, 1]$, the period 1 price in a discriminatory advertising equilibrium (if it exists), is higher than the monopoly price (i.e. the price charged in a uniform advertising equilibrium), which is higher than the purchase cutoff employed by a consumer in a discriminatory advertising equilibrium. Formally,*

$$v^*(q) \leq p_M \leq p_1(q).$$

Proof. Note that at $q = \frac{1}{2}$, $p_1(q) = v^*(q) = p_M$. The result now follows since $p_1'(\cdot)$ is positive, while $v^{*\prime}(\cdot)$ is negative. □

Finally, to resolve existence, which follows easily from the definitions.

Observation 1. *A discriminatory advertising equilibrium exists at all q such that $r(1, v^*(q), q) \geq \eta$ and $r(0, v^*(q), q) \leq \eta$.*

Next, note that $v^*(q)$ is a continuous function of q. Therefore, equilibria of type 1 may exist for possibly multiple disjoint intervals in $(\frac{1}{2}, 1]$.

5.4 Illustrations via an Example

In this section we highlight some of the counter-intuitive effects that result from changing the level of differential privacy constraining the advertiser's signal, by means of an explicit family of simple examples. The phenomena we highlight are quite general, and are generally not brittle to the choice of specific parameters in the game. For simplicity of exposition, we highlight each of these phenomena in the simplest example in which they arise. For the remainder of this section, we take the distribution of buyer values, F, to be the uniform distribution on $[0, 1]$. We also set the buyer's probability of having type t_1 to be exactly his value — i.e. we take $g(v) = v$ for all $v \in [0, 1]$. Finally, we set the additional utility that a buyer gets from being shown ad A to be $\delta = 1$. The value of parameter η (along with its defining parameters s_{1A}, s_{1B}, s_{2A}, and s_{2B}) will vary by example, and will be specified when relevant.

The static monopoly price in this game is $p_M = 1/2$, and the price and cutoff value in a discriminatory equilibrium (if it exists at a given q) are

$$p_1(q) = \frac{1}{2} + (2q-1)\frac{\delta}{2} \quad \text{and} \quad v^*(q) = \frac{1}{2} - (2q-1)\frac{\delta}{2}.$$

Below we plot these values as a function of q. Note that in this particular example, the discriminatory price and cutoff value are linear in q (because values are distributed uniformly), although this need not be the case in general.

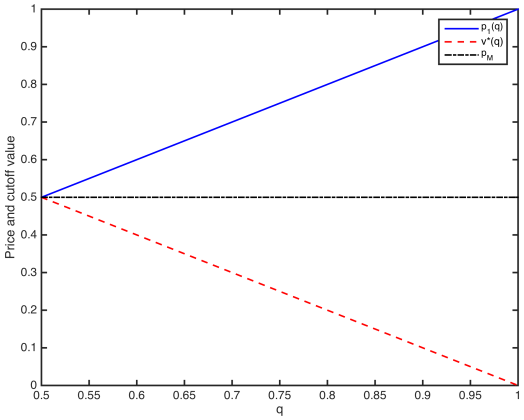


Figure 5.1: A plot of the monopoly price p_M and the equilibrium discriminatory price p_1 and cutoff value v^* as a function of q. In a uniform equilibrium, the cutoff value is equal to the monopoly price because consumers behave myopically.

Figure 5.1 shows that as q increases, the discriminatory equilibrium price increases and cutoff value (value of marginal consumer that purchases in period 1) decreases. Relative to a uniform advertising equilibrium, more consumers purchase the good in period 1, and at a higher price in a discriminatory equilibrium. Observe further that at $q = 1$ (i.e. no privacy), all consumers purchase in period 1, regardless of their value. This is because the value in period 2 of hiding information relevant to their type exceeds the loss that they take by buying at a loss in period 1. Here, when the consumers are offered no privacy protections, they in effect change their behavior to guarantee their own privacy.

The existence and types of equilibria in this game depend on the advertiser's posterior, given that the prices and cutoff values above will arise in period 1. The advertiser's posterior beliefs about the consumer's type for each realization of $\hat{b}$ are given below for both the discriminatory and uniform advertising equilibria.

$$
\begin{aligned}
r(1, v^*(q), q) &= \frac{-2q^3 + 5q^2 - 3q + 1}{4(q^2 - q + \frac{1}{2})} \\
r(0, v^*(q), q) &= \frac{-2q^3 + 5q^2 - 3q}{4(q^2 - q)} = \frac{3 - 2q}{4} \\
r(1, p_M, q) &= \frac{1 + 2q}{4} \\
r(0, p_M, q) &= \frac{3 - 2q}{4}
\end{aligned}
$$

To illustrate the existence of various equilibria, we plot the advertiser's possible posterior beliefs below as a function of q.

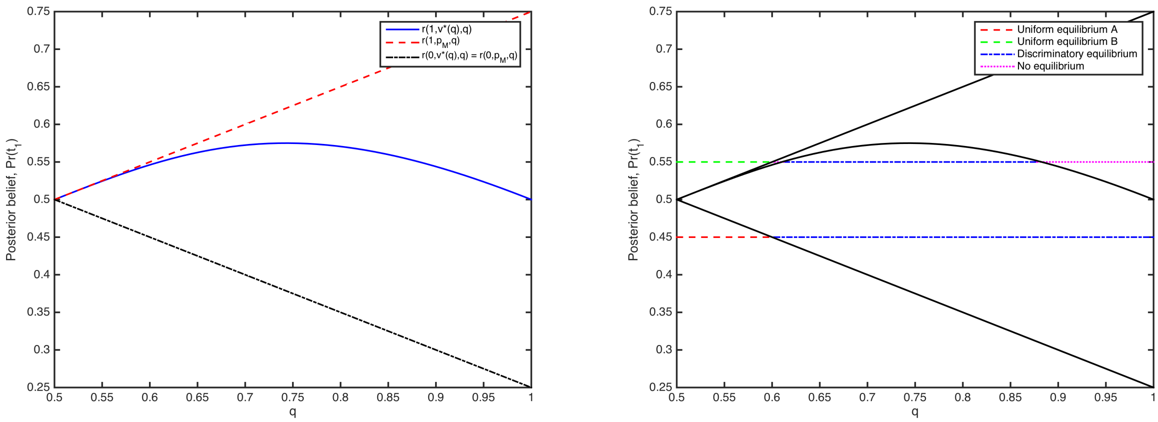


Figure 5.2: On the left is a plot of the advertiser's possible posteriors in both discriminatory and uniform equilibria, given the corresponding prices and cutoff values of Period 1 and an observation of $\hat{b}$. In this example, $r(0, v^*(q), q) = r(0, p_M, q)$ for all q. On the right, we show for $\eta = 0.45$ and $\eta = 0.55$ how these posteriors correspond to different equilibrium types as q varies.

Note the following counter-intuitive fact: the advertiser's posterior, having seen a noisy "purchased" bit, i.e., $\hat{b} = 1$, in a discriminatory equilibrium is non-monotone in the noise level q. Statically, if we were to increase the privacy level (i.e. decrease q), we should always expect the advertiser's posterior to be *less informative* about the consumer's type, but as we see here, in equilibrium, increasing the privacy level can sometimes make the advertiser's posterior more accurate. This can occur because of the two competing implications of adding less noise (i.e. increasing q): on the one hand, the observed bit $\hat{b}$ is less noisy, and is thus a more accurate indicator of the consumer's purchase decision in period 1. On the other hand, as q increases, a larger fraction of consumers buy in period 1; the pool of consumers who do purchase the item is "watered down" by low-valued consumers who are unlikely to have type t_1. This can be viewed as a larger fraction of consumers modifying their behavior to guarantee their own privacy, as the privacy protections inherent in the market are weakened. The dominating effect on the posterior depends on q and the game parameters.

Similar non-monotonicities can never be observed when the advertiser sees $\hat{b} = 0$ in a discriminatory equilibrium because these two implications are no longer at odds. As we reduce the amount of noise added, the noisy bit $\hat{b}$ is still more likely to be accurate. In addition, the maximum value $v^*(q)$ of a consumer who did not purchase is decreasing in q, ensuring that only the lowest valued (and thus the least likely to have type t_1) consumers do not purchase in period 1. These two effects conspire to ensure that $r(0, v^*(q), q)$ is monotonically decreasing in q.

Figure 5.2 can be also be used to illustrate to existence of equilibria in this family of games. Specifying a value of η for the game determines the type of equilibria (if any) that exist at each q, according to Conditions (5.1), (5.2), and (5.3). This can be easily visualized using Figure 5.2.

Equilibria need not exist for all ranges of q: it is possible for none of Conditions (5.1), (5.2), or (5.3) to be satisfied for a given η and q. However, in all games, there is a uniform equilibrium at $q = 1/2$, where consumers behave myopically in period 1 and then no information is shared with the advertiser. Equilibria also need not be unique for a given q; a discussion of equilibrium multiplicity is deferred to Section 5.5.

Next we show that in settings for which a discriminatory equilibrium exists for a range of q, the *mutual information* between the noisy bit $\hat{b}$ and the consumer's type can be non-monotone in q. In particular, as we increase the privacy protections of the market (i.e. decrease q), we can sometimes end up *increasing* the amount of information about the consumer's type present in the advertiser's signal! This is for similar reasons to those that lead to non-monotonicity of the advertiser's posterior belief—as the market's privacy protections decrease, consumers change their behavior in order to guarantee their own privacy.

Mutual information (Definition 24) is a standard information theoretic measure which quantifies the amount of information revealed about one random variable, by observing a realization of the other random variable. Here, we use it to quantify how much the advertiser is able to learn about the consumer's type from the noisy signal $\hat{b}$.

Definition 24 (Mutual Information)**.** *The* mutual information *between two discrete random variables X and Y with joint probability distribution $p(x, y)$ and marginal probability distributions $p(x)$ and $p(y)$ respectively, is defined as,*

$$I(X; Y) = \sum_{y \in Y} \sum_{x \in X} p(x, y) \log \left(\frac{p(x, y)}{p(x)p(y)} \right).$$

Figure 5.3 plots the mutual information between $\hat{b}$ and the consumer's type as a function of q in a discriminatory equilibrium.[3] Note that although both the advertiser's posterior and the mutual information between the consumer's purchase decision and the signal exhibit similar non-monotonicities in q, they do not peak at the same value of q!

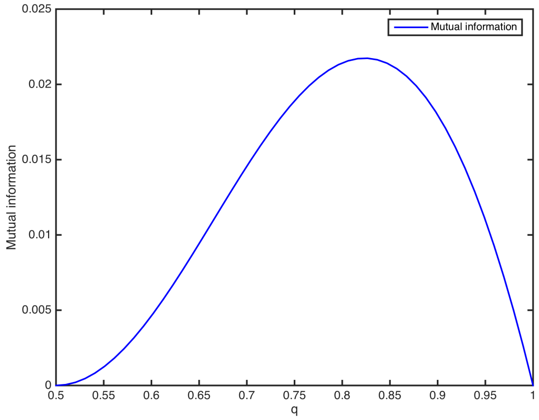


Figure 5.3: A plot of the mutual information between $\hat{b}$ and the consumer's type in a discriminatory equilibrium.

These phenomena together suggest that in the range of q in which the mutual information is decreasing, the advertiser might actually prefer that the market include stronger privacy guarantees. Indeed, Figure 5.4 plots the advertiser's utility in a discriminatory equilibrium as a function of q, where $s_{1A} = s_{2B} = 1$ and $s_{1B} = s_{2A} = 0$. This setting of parameters gives $\eta = 1/2$, where a discriminatory equilibrium exists for all q.

[3] As illustrated by Figure 5.2, there are games for which a discriminatory equilibria exist for the relevant range of q, e.g., when $\eta = 1/2$, a discriminatory equilibrium exists for all $q \in [1/2, 1]$.

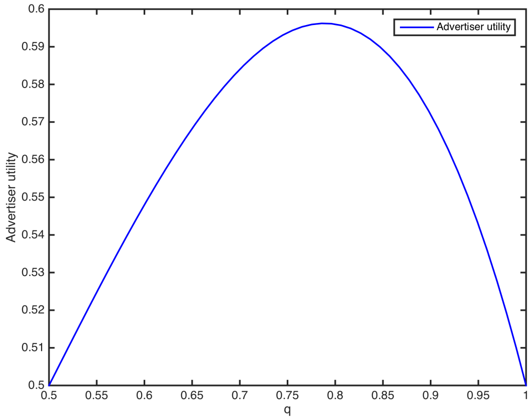


Figure 5.4: A plot of the advertiser's utility in a discriminatory equilibrium, where $s_{1A} = s_{2B} = 1$ and $s_{1B} = s_{2A} = 0$. These parameters correspond to $\eta = 1/2$, which ensures that a discriminatory equilibrium exists for all $q \in [1/2, 1]$.

As predicted, the advertiser's utility is non-monotone in q. However, the advertiser's utility is not maximized at the same value of q that maximizes the mutual information. Thus, the advertiser's interests are not necessarily aligned with the goal of learning as much information about the consumer as possible — and are certainly not incompatible with privacy technologies being introduced into the market. Indeed, the ideal level of q for the advertiser is strictly on the interior of the feasible set $[1/2, 1]$.

We would also like to understand how consumer surplus and profit vary with q. These are confounded by the fact that for a fixed level of η, an equilibrium type may or may not exist for a given q. To simplify the analysis to not account for this existence problem, consider the following though experiment: for each equilibrium type, and any q, pick η such that the appropriate one of Conditions (5.1), (5.2), or (5.3) is satisfied. Fixing the advertiser's equilibrium behavior, η only affects the advertiser's payoff, not the seller's or buyer's. Figure 5.5 plots the consumer's surplus and the seller's profit as a function of q for each equilibrium type, under this artificial thought experiment.

In a discriminatory equilibrium, the consumer surplus is decreasing in q, while the seller's profit is increasing in q. This suggests that the preferences of the consumer and seller are misaligned: the consumer fares best with full privacy and the seller prefers no privacy. As q increases, more consumers purchase the good at a higher price in equilibrium. Unsurprisingly, consumer surplus and revenue are constant

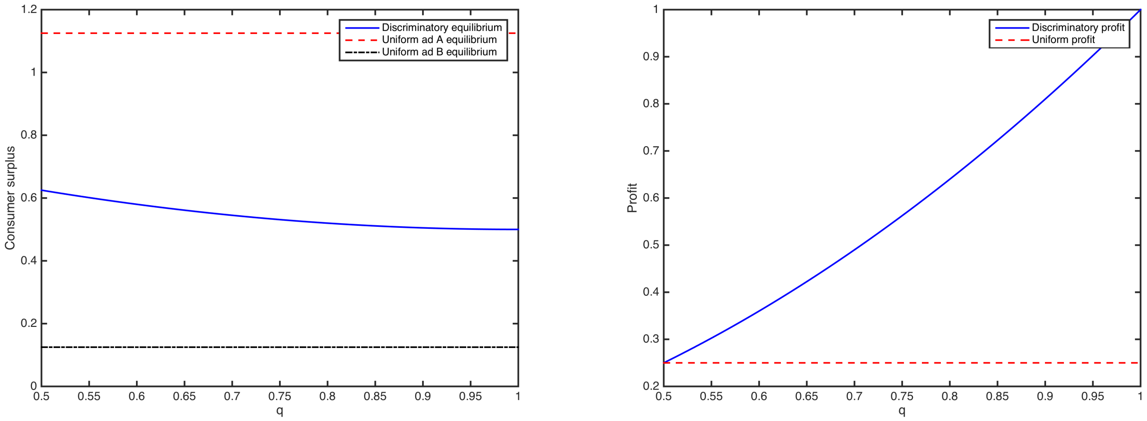


Figure 5.5: On the left is a plot of consumer surplus in all three equilibrium types. On the right is a plot of the seller's profit in both discriminatory and uniform equilibria. The seller's profit in a uniform equilibrium doesn't depend on the ad shown in period 2.

across q in uniform advertising equilibria.

In addition to changes for a fixed equilibrium type, varying q can also change the type of equilibria that exist in the game. Figure 5.2 showed that small changes in q can cause new equilibria to spring into existence or disappear. Thus consumer surplus, seller's profit, and advertiser's utility can jump discontinuously in q. We illustrate this phenomenon with two examples: the first example illustrates changes between a uniform advertising equilibrium B, a discriminatory equilibrium, and no equilibrium. The second shows a change from a uniform advertising equilibrium A to a discriminatory equilibrium.

For the first example, set $s_{1A} = .5$, $s_{2B} = .6$, and $s_{1B} = s_{2A} = .05$, which implies $\eta = .55$. As illustrated in Figure 5.2, as we increase q from $1/2$ to 1, there is first a uniform advertising equilibrium B, then equilibria briefly cease to exist, then discriminatory equilibria exist, and finally no equilibria exist for large q. Each change of equilibrium existence results in discrete jumps in the consumer surplus, profit, and advertiser utility. Figure 5.6 illustrates this below.

For the second example, set $s_{1A} = .6$, $s_{2B} = .5$, and $s_{1B} = s_{2A} = .05$, which implies $\eta = .45$. In this game, as q increases from $1/2$ to 1, the equilibrium type changes discretely from a uniform advertising equilibrium A to a discriminatory equilibrium, also illustrated in Figure 5.2. This change also causes discontinuities in the consumer surplus, profit, and advertiser utility. Note here that a tiny decrease

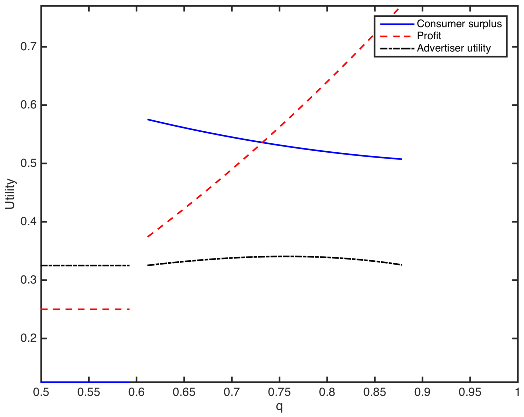


Figure 5.6: A plot of consumer surplus, profit, and advertiser utility in equilibrium when $s_{1A} = .5$, $s_{2B} = .6$, and $s_{1B} = s_{2A} = .05$. Discontinuities correspond to changes of the equilibrium type that exists, or an absence of equilibria altogether.

in the level of privacy (i.e. increase in q) can cause a precipitous drop in welfare. Figure 5.7 illustrates this effect.

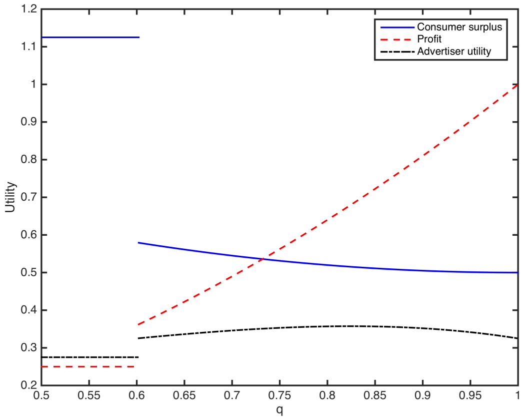


Figure 5.7: A plot of consumer surplus, profit, and advertiser utility in equilibrium when $s_{1A} = .6$, $s_{2B} = .5$, and $s_{1B} = s_{2A} = .05$. Discontinuities correspond to changes of the equilibrium type that exists.

5.5 Multiplicty of Equilibria and Other Results

In this section we build on the example of the previous section to provide some formal results about equilibrium multiplicity, welfare, etc. as a function of the promised level of privacy offered to the consumer. Proofs are deferred to the Appendix.

Multiplicity

Proposition 9. *A discriminatory equilibrium and a uniform advertising equilibrium A can coexist in the same game for the same level of q.*

Proposition 10. *If a discriminatory equilibrium and a uniform advertising equilibrium A coexist at a given noise level q, then the buyer prefers the uniform equilibrium A to the discriminatory equilibrium, regardless of her value v. On the other hand, the seller prefers the discriminatory equilibrium.*

Proposition 10 is intuitive, so we omit a formal proof. To see the first claim, observe that a buyer faces both a lower price in period 1 and sees a better ad in period 2, so she is always better off in the uniform advertising equilibrium, regardless of her value. To see the latter, observe that the discriminatory equilibrium allows the seller to sell to more consumers ($v^*(q) < p_M$) at a higher price ($p_1(q) > p_M$).

Takeaway Small changes in q can make uniform advertising equilibria cease to exist, and can therefore have a discrete impact on welfare and revenue, as they cause the equilibrium to shift discontinuously from uniform to discriminatory. At these boundaries, the buyer may (strictly, discontinuously) prefer slightly less privacy while the seller may (strictly, discontinuously) prefer more privacy!

Proposition 11. *A discriminatory equilibrium and a uniform advertising equilibrium B can coexist in the same game for the same level of q.*

Proposition 12. *In any game where both a discriminatory equilibrium and a uniform advertising equilibrium B exist for the same value of q, average consumer welfare is always higher under the discriminatory equilibrium, but individual consumers may have different preferences for these two equilibria.*

Takeaway In a game where a discriminatory equilibrium and uniform advertising equilibrium B coexist, buyers prefer the discriminatory equilibrium. Since the uniform advertising equilibrium B exists for an interval of "low" q, a buyer therefore may prefer less privacy!

Welfare Comparative Statics and Preferences over Levels of Privacy

Proposition 13. *In settings where a discriminatory equilibrium exists for a range of q, the equilibrium consumer surplus can be increasing in q.*

Takeaway The buyer's preferences over different levels of privacy may be complex, and her welfare is not necessarily monotonically increasing in her privacy as is often assumed. Consumers may have negative marginal utility for increased privacy (i.e. a smaller value of q), and would prefer that their purchase decision be revealed more accurately.

Observation 2. *The seller prefers the largest q consistent with discriminatory equilibrium — he gets a higher price and more demand. If a discriminatory equilibrium ceases to exist at some interior q, the seller will at that point prefer that the market provide* more *privacy.*

Takeaway The seller prefers discriminatory equilibria over uniform advertising equilibria, and the least privacy that is consistent with a discriminatory equilibrium if given the choice over privacy levels.

Observation 3. *For any level of noise q, the advertiser always prefers a discriminatory equilibrium to a uniform equilibrium if both exist.*

To see this, note that in any uniform equilibrium, the advertiser's net utility is the same as his ex-ante utility from showing that ad (since he ignores information from period 1). Since he chooses to act on the information he gets in a discriminatory equilibrium, his net utility must exceed the ex-ante utility of showing the same ad.

Takeaway The advertiser will always prefer levels of q consistent with discriminatory equilibria. Among these, however, he may prefer strictly interior levels of q, for example, as demonstrated by Figure 5.4 previously.

5.6 Concluding Remarks

A rich body of work on differential privacy has developed in the computer science literature over the past decade. Broadly caricatured, this literature provides algorithms for accurate data analyses (of various sorts), subject to guaranteeing ϵ-differential privacy to individual entries in the dataset. Typically, ϵ can be set to be any value (implicitly to be chosen by the entity that controls access to the dataset), and mediates a trade-off between the strength of the privacy guarantee offered to the data owners and the accuracy of the analysis promised to the user of the data. This trade-off is typically viewed in simple terms: higher values of ϵ (i.e. less privacy) are better for the analyst because they allow for higher accuracy, the reasoning goes, and worse for the privacy-desiring individual. Of course, if the dataset is already gathered, this reasoning is correct.

In this paper, we add some caveats to this folk wisdom in a simple stylized model. The dataset here consists of the individuals' purchase decision of a good. The data analyst in our model is an advertiser. Individuals do not care about the privacy of their purchase decision for its own sake, but rather, care about privacy only insofar as it affects how ads are targeted at them. A crucial point is that in our model, at the time of choosing the privacy policy, these purchase decisions have not yet been made. As a result, the price of the good, the purchase decision of the individual, and the advertising policy of the advertiser all depend on the announced privacy policy. Evaluations of the privacy preferences of seller, advertiser, and buyers must take into account everyone's equilibrium incentives. As we demonstrated, these can be the opposite of the simple static trade-offs we are used to, and reasoning about them correctly can be complex.

As the literature expands from (the already hard) questions of privately analyzing existing datasets, to thinking of setting privacy policies that influence future user behavior and the datasets that result from this behavior, the equilibrium approach we espouse here will be important. We hope this paper serves as a call-to-arms to reasoning about privacy policy in such settings, while also highlighting the difficulties.

5.7 Appendix: Omitted Proofs

This appendix contains all proofs that were omitted in the body of the paper.

Proposition 14. *All equilibria have the property that in period 1, there exists a threshold value v^* such that the consumer buys if and only if $v > v^*$.*

Proof. Assume not. Then there exists v, v' such that $v' < v$, and in equilibrium consumers with value v' buy in period 1, while consumers with value v do not. We consider the three possible equilibrium types, and show that each one leads to a contradiction.

If this is a discriminatory equilibrium, then consumers with $\hat{b} = 1$ are shown ad A, and consumers with $\hat{b} = 0$ are shown ad B. Then it must be the case that the consumer's utility satisfies:

$$u(v', \text{buy}) \geq u(v', \text{not buy}) \quad \text{and} \quad u(v, \text{not buy}) \geq u(v, \text{buy}).$$

This implies:

$$\begin{aligned} & u(v',\text{buy}) + u(v,\text{not buy}) \geq u(v',\text{not buy}) + u(v,\text{buy}) \\ \iff & [v' - p + q\delta] + [(1-q)\delta] \geq [(1-q)\delta] + [v - p + q\delta] \\ \iff & v' \geq v \end{aligned}$$

This is a contradiction because $v > v'$.

If this is a uniform advertising equilibrium A, then all consumers are shown ad A and receive utility δ in period 2. It must be the case that:

$$\begin{aligned} & u(v',\text{buy}) \geq u(v',\text{not buy}) \\ \iff & v' - p + \delta \geq \delta \\ \iff & v' \geq p. \end{aligned}$$

Also,

$$\begin{aligned} & u(v,\text{not buy}) \geq u(v,\text{buy}) \\ \iff & \delta \geq v - p + \delta \\ \iff & p \geq v. \end{aligned}$$

These two facts above imply that $v' \geq v$, which is a contradiction since we have assumed that $v > v'$.

If this is a uniform advertising equilibrium B, then all consumers are shown ad B and receive zero utility in period 2.

$$\begin{aligned} u(v',\text{buy}) \geq u(v',\text{not buy}) & \iff v' \geq p \\ u(v,\text{not buy}) \geq u(v,\text{buy}) & \iff p \geq v \end{aligned}$$

Again, these facts imply $v' \geq v$, which is a contraction. □

Proposition 9 *A discriminatory equilibrium and a uniform advertising equilibrium A can coexist in the same game for the same level of q.*

Proof. Suppose the distribution F of buyers' values is uniform on $[0,1]$, and suppose the distribution of period 2 types is such that $g(v)$ is the step function that is 0 below $\frac{1-\delta}{2}$ and 1 at or above $\frac{1-\delta}{2}$. We will show that in this game, both uniform A and discriminating equilibria co-exist for a continuous range of q as well, and derive sufficient conditions on η, δ, for both to exist. Recall from Lemma 5 that it is sufficient for a uniform advertising equilibrium A to exist if $r(0, p_M, q) > \eta$. Given myopic behavior on the part of the consumers and the seller setting the monopoly price in period 1, the advertiser's posterior having seen noisy bit $\hat{b} = 0$ is as follows:

$$\begin{aligned} r(0, p_M, q) &= \frac{q\int_0^{1/2} \mathbb{1}_{v > \frac{1-\delta}{2}} dv + (1-q)\int_{1/2}^1 dv}{qF(1/2) + (1-q)(1-F(1/2))} \\ &= \frac{q\frac{\delta}{2} + (1-q)\frac{1}{2}}{q\frac{1}{2} + (1-q)\frac{1}{2}} \\ &= \frac{q\frac{\delta}{2} + (1-q)\frac{1}{2}}{\frac{1}{2}} \\ &= \delta q + (1-q). \end{aligned}$$

A uniform advertising equilibrium A will exist whenever $r(0, p_M, q) > \eta$, that is, for any $q \in [1/2, 1]$ satisfying

$$\delta q + (1-q) > \eta \iff q < \frac{1-\eta}{1-\delta}.$$

Thus a uniform advertising equilibrium A will exist for all $q \in [1/2, \frac{1-\eta}{1-\delta}]$. Restricting $\eta < \frac{1+\delta}{2}$ and $0 < \delta < 1$ will ensure that $\frac{1-\eta}{1-\delta} > 1/2$ so this range is non-empty.

We now verify that there is a continuous range of q for which there also exists a discriminatory equilibrium of this game. By equation (5.6), the discriminatory equilibrium price $p_1(q)$ must satisfy $p_1(q) - I(p_1(q) + (1-2q)\delta) = 0$, where $I(v) = \frac{1-F(v)}{f(v)}$. Plugging in the distribution F as $U[0,1]$,

$$\begin{aligned} &p_1(q) - (1 - p_1(q) - (1-2q)\delta) = 0 \\ &\qquad\qquad \iff \quad p_1(q) = \frac{1}{2} - (1-2q)\frac{\delta}{2}. \end{aligned}$$

The Period 1 cutoff value is then

$$v^*(q) = p_1(q) + (1-2q)\delta = \frac{1}{2} + (1-2q)\frac{\delta}{2}.$$

There exists a discriminatory equilibrium at $q \in [1/2, 1]$ if both $r(1, v^*(q), q) > \eta$ and $r(0, v^*(q), q) < \eta$.

We first compute the advertiser's posterior $r(1, v^*(q), q)$ and show that it is monotone increasing in q.

$$\begin{aligned}
r(1, v^*(q), q) &= \frac{(1-q)\int_0^{v^*(q)} \mathbb{1}_{v > \frac{1-\delta}{2}} dv + q\int_{v^*(q)}^1 dv}{qF(v^*(q)) + (1-q)(1-F(v^*(q)))} \\
&= \frac{(1-q)(v^*(q) - \frac{1-\delta}{2}) + q(1-v^*(q))}{(1-q)v^*(q) + q(1-v^*(q))} \\
&= 1 - \frac{(1-q)\frac{1-\delta}{2}}{(1-q)v^*(q) + q(1-v^*(q))} \\
&= 1 - \frac{\frac{1}{2}(1-q)(1-\delta)}{q + (1-2q)[\frac{1}{2} + (1-2q)\frac{\delta}{2}]} \\
&= 1 - \frac{\frac{1}{2}(1-q)(1-\delta)}{q + \frac{1}{2}(1-2q) + \frac{1}{2}(1-2q)^2\delta} \\
&= 1 - \frac{\frac{1}{2}(1-q)(1-\delta)}{\frac{1}{2}[1 + (1-2q)^2\delta]} \\
&= 1 - \frac{(1-q)(1-\delta)}{1 + (1-2q)^2\delta}
\end{aligned}$$

We take the partial derivative of $r(1, v^*(q), q)$ with respect to q.

$$\begin{aligned}
\frac{\partial r(1, v^*(q), q)}{\partial q} &= -\frac{(-1+\delta)[1 + (1-2q)^2\delta] - [-2\delta(1-2q)(-2)q(1-\delta)]}{[1 + (1-2q)^2\delta]^2} \\
&= \frac{(1-\delta)[1 + (1-2q)^2\delta + 4\delta(2q-1)q]}{[1 + (1-2q)^2\delta]^2}
\end{aligned}$$

Due to our restrictions that $\delta \in (0, 1)$ and $q \in [1/2, 1]$, this expression is strictly positive, so $r(1, v^*(q), q)$ is strictly increasing in q.

Next note that $r(1, v^*(1/2), 1/2) = 1 - \frac{\frac{1}{2}(1-\delta)}{1+0} = \frac{1+\delta}{2}$. Thus for $\eta < \frac{1+\delta}{2}$, then $r(1, v^*(q), q) > \eta$ for any $q \in [1/2, 1]$. Fortunately, this is the same condition on η that was required for a uniform advertising equilibrium A to exist.

We now compute the advertiser's posterior $r(0, v^*(q), q)$.

$$\begin{aligned}
r(0, v^*(q), q) &= \frac{q \int_0^{v^*(q)} \mathbb{1}_{v > \frac{1-\delta}{2}} dv + (1-q) \int_{v^*(q)}^1 dv}{qF(v^*(q)) + (1-q)(1 - F(v^*(q)))} \\
&= \frac{q(v^*(q) - \frac{1-\delta}{2}) + (1-q)(1 - v^*(q))}{qv^*(q) + (1-q)(1-v^*(q))} \\
&= 1 - \frac{q\frac{1-\delta}{2}}{qv^*(q) + (1-q)(1-v^*(q))} \\
&= 1 - \frac{\frac{1}{2}q(1-\delta)}{(1-q) - (1-2q)[\frac{1}{2} + (1-2q)\frac{\delta}{2}]} \\
&= 1 - \frac{\frac{1}{2}q(1-\delta)}{(1-q) - \frac{1}{2}(1-2q) - \frac{1}{2}(1-2q)^2\delta} \\
&= 1 - \frac{\frac{1}{2}q(1-\delta)}{\frac{1}{2}[1 - (1-2q)^2\delta]} \\
&= 1 - \frac{q(1-\delta)}{1 - (1-2q)^2\delta}
\end{aligned}$$

For there to be a discriminatory equilibrium at q, it must be the case that

$$r(0, v^*(q), q) = 1 - \frac{q(1-\delta)}{1 - (1-2q)^2\delta} < \eta.$$

Although there is not a nice closed form description of the $q \in [1/2, 1]$ satisfying this condition, we note that this expression is differentiable (and thus continuous), everywhere except when

$$1 = (1-2q)^2\delta \quad \Longleftrightarrow \quad q = \frac{1 \pm \sqrt{\frac{1}{\delta}}}{2}.$$

From our restriction of $\delta < 1$, this q will fall outside of our range of interest, and this expression $r(0, v^*(q), q)$ is differentiable on $[1/2, 1]$.

We now take the derivative of $r(0, v^*(q), q)$ with respect to q, and see that it is negative, so the function is monotone decreasing.

$$\frac{\partial r(0, v^*(q), q)}{\partial q} = -\frac{(1-\delta)[1 - (1-2q)^2\delta] - [-2\delta(1-2q)(-2)q(1-\delta)]}{[1 - (1-2q)^2\delta]^2}$$

We are only interested in the sign of this expression, and the denominator is clearly positive, so we will proceed only with the numerator (without the negative sign in front of it).

$$\begin{aligned}(1-\delta)[1-(1-2q)^2\delta-4\delta(1-2q)q] &= (1-\delta)[1-\delta[(1-2q)^2+4q(1+2q)]] \\ &= (1-\delta)[1-\delta(1-2q)(1+2q)] \\ &= (1-\delta)[1-\delta(1-4q^2)] \\ &= (1-\delta)^2+(1-\delta)4\delta q^2\end{aligned}$$

Since $0 < \delta < 1$, this expression is positive, so plugging it back into the original expression (with a negative in front) means that $\frac{\partial r(0,v^*(q),q)}{\partial q}$ is negative, so $r(0,v^*(q),q)$ is a monotone decreasing function. Thus if $r(0,v^*(\bar{q}),\bar{q}) = \eta$ for some distinguished $\bar{q}$, then $r(0,v^*(q),q) < \eta$ for all $q \in (\bar{q},1]$.

We complete the proof by showing that there exist values of η such that $\bar{q} < \frac{1-\eta}{1-\delta}$. Thus both types of equilibria will exist for a non-empty range of $q \in (\bar{q}, \max\{\frac{1-\eta}{1-\delta},1\}]$.

Note that $r(0,v^*(1/2),1/2) = 1 - \frac{\frac{1}{2}(1-\delta)}{1-0} = \frac{1+\delta}{2}$, so from our restriction of $\eta < \frac{1+\delta}{2}$ and because $r(0,v^*(q),q)$ is strictly decreasing q, we can always choose an η such that $r(0,v^*(\bar{q}),\bar{q}) = \eta$ for some $\bar{q} \in (1/2,1]$.

Finally, we show that $r(0,v^*(q),q) < r(0,p_M,q)$ for all $q \in (1/2,1]$.

$$\begin{aligned}r(0,v^*(q),q) &= 1-\frac{q(1-\delta)}{1-(1-2q)^2\delta} \\ &> 1-q(1-\delta) \quad \text{because } q \in (1/2,1] \text{ and } \delta \in [0,1] \\ &= \delta q+(1-q) \\ &= r(0,p_M,q)\end{aligned}$$

In particular, this holds at $\bar{q}$, meaning that $\eta = r(0,v^*(\bar{q}),\bar{q}) < r(0,p_M,\bar{q})$, so a uniform advertising equilibrium A exists as $\bar{q}$, so $\bar{q} < \frac{1-\eta}{1-\delta}$.

To summarize, when $0 < \delta < 1$ and $\eta < \frac{1+\delta}{2}$ in this game, a uniform advertising equilibrium A exists for $q \in [1/2, \max\{\frac{1-\eta}{1-\delta},1\}]$, and a discriminatory equilibrium exists for $q \in (\bar{q},1]$, and both of these ranges are non-empty. Further, these ranges overlap and there is a non-empty set of $q \in (\bar{q}, \max\{\frac{1-\eta}{1-\delta},1\}]$ where both equilibria types coexist. □

Proposition 11 *A discriminatory equilibrium and a uniform advertising equilibrium B can coexist in the same game for the same level of q.*

Proof. Consider the same example as in Proposition 9, where the distribution F of buyers' values is uniform on $[0,1]$ and $g(v)$ is 1 if $v \geq \frac{1-\delta}{2}$ and 0 otherwise. Let $\delta \in (0,1)$.

There will exist a uniform advertising equilibrium B in this game at q if and only if $r(1, p_M, q) < \eta$, where $p_M = \frac{1}{2}$. The advertiser's posterior in this setting after seeing $\hat{b} = 1$ is

$$\begin{aligned} r(1, p_M, q) &= \frac{(1-q)\int_0^{1/2} \mathbb{1}_{v > \frac{1-\delta}{2}} dv + q\int_{1/2}^1 dv}{qF(1/2) + (1-q)(1-F(1/2))} \\ &= \frac{(1-q)\frac{\delta}{2} + q\frac{1}{2}}{q\frac{1}{2} + (1-q)\frac{1}{2}} \\ &= \frac{(1-q)\frac{\delta}{2} + q\frac{1}{2}}{\frac{1}{2}} \\ &= (1-q)\delta + q. \end{aligned}$$

A uniform advertising equilibrium B exists for all q satisfying

$$(1-q)\delta + q < \eta \iff q < \frac{\eta - \delta}{1-\delta}.$$

Restricting $\eta > \frac{1+\delta}{2}$ ensures that there is a non-empty interval of $q \in [1/2, 1]$ satisfying this condition.

A discriminatory equilibrium exists for any q such that $r(1, v^*(q), q) > \eta$ and $r(0, v^*(q), q) < \eta$. Recall from the proof of Proposition 9 that the equilibrium price is $p_1(q) = \frac{1}{2} - (1-2q)\frac{\delta}{2}$ and the equilibrium cutoff value is $v^*(q) = \frac{1}{2} + (1-2q)\frac{\delta}{2}$. Also recall that $r(0, v^*(q), q) = 1 - \frac{q(1-\delta)}{1-(1-2q)^2\delta}$, and is monotone decreasing in q, while $r(1, v^*(q), q) = 1 - \frac{(1-q)(1-\delta)}{1+(1-2q)^2\delta}$ and is monotone increasing in q. Finally, recall that $r(0, v^*(1/2), 1/2) = r(1, v^*(1/2), 1/2) = \frac{1+\delta}{2}$. From these facts and our restriction that $\eta > \frac{1+\delta}{2}$, the condition that $r(0, v^*(q), q) < \eta$ will be satisfied for all $q \in [1/2, 1]$.

What remains to be shown is that $r(1, v^*(q), q) > \eta$ for some range of q overlapping with $[1/2, \max\{\frac{\eta-\delta}{1-\delta}, 1\}]$. We do this by proving that $r(1, v^*(q), q) > r(1, p_M, q)$ for all $q \in (1/2, 1]$. Assume not.

$$\begin{aligned} 1 - \frac{(1-q)(1-\delta)}{1+(1-2q)^2\delta} &< (1-q)\delta + q \\ \frac{(1-2q)^2\delta + (1-q)\delta + q}{1+(1-2q)^2\delta} &< (1-q)\delta + q \\ (1-2q)^2\delta + (1-q)\delta + q &< [(1-q)\delta + q][1+(1-2q)^2\delta] \\ 1 &< (1-q)\delta + q \end{aligned}$$

This is a contradiction because $q \in [1/2, 1]$ and $\delta \in (0, 1)$, so it must be that $r(1, v^*(q), q) > r(1, p_M, q)$. In particular, fix any $\frac{1+\delta}{2} < \eta < 1$, and let $\bar{q}$ be the value of q such that $r(1, v^*(\bar{q}), \bar{q}) = \eta$. Then $\eta = r(1, v^*(\bar{q}), \bar{q}) > r(1, p_M, \bar{q})$, so a uniform advertising equilibrium B exists at $\bar{q}$, so $\bar{q} < \frac{\eta-\delta}{1-\delta}$.

To summarize, when $0 < \delta < 1$ and $\frac{1+\delta}{2} < \eta < 1$ in this game, both a uniform advertising equilibrium B and a discriminatory equilibrium exist for a non-empty range of $q \in (\bar{q}, \max\{\frac{\eta-\delta}{1-\delta}, 1\}]$. □

Proposition 12 *In any game where both a discriminatory equilibrium and a uniform advertising equilibrium B exist for the same value of q, average consumer welfare is always higher under the discriminatory equilibrium, but individual consumers may have different preferences for these two equilibria.*

Proof. In any uniform advertising equilibrium B, the myopic monopoly price is charged, and by assumption, all consumers are shown the inferior ad. Therefore the ex-ante consumer welfare is $\int_{p_M}^1 (v - p_M) f(v) dv$.

At any q where a discriminatory equilibrium exists, the ex-ante consumer welfare is:

$$\begin{aligned} &\int_{v^*(q)}^1 (v - p_1(q)) f(v) dv + [q(1 - F(v^*(q))) + (1-q)F(v^*(q))]\delta \\ = &\int_{v^*(q)}^1 [v - p_1(q) - (1-2q)\delta] f(v) dv + (1-q)\delta. \end{aligned}$$

Note that the latter term is positive. Further, recall from Lemma 7, $v^*(q) = p_1(q) + (1-2q)\delta$, and that $v^*(q) \le p_M$. Therefore $\int_{v^*(q)}^1 (v - p_1(q) - (1-2q)\delta) \ge \int_{p_M}^1 (v - p_M) f(v) dv$.

We will now see that if both a discriminatory equilibrium and a uniform advertising equilibrium B exist, then consumers with values $v \in [0, p_M]$ will prefer the discriminatory equilibrium, while the preferences of consumers with values $v \in [p_M, 1]$ depend on the game parameters δ and η, as well as the noise level q.

For consumers with values $v \in [0, v^*(q)]$, they prefer the discriminatory equilibrium because they still don't purchase the good in period 1, but they have a chance at the better ad in period 2. For consumers with values $v \in [v^*(q), p_M]$, under the discriminatory equilibrium, they receive utility $(v - p_1(q)) + \delta q$ for buying the good at price $p_1(q)$ and being shown the better ad with probability q. Under the uniform advertising equilibrium B, they receive utility 0 for not buying in period 1, and then being shown ad B with probability 1. For all consumers with value in this range, the discriminatory equilibrium is preferred because

$$v - p_1(q) + q\delta > (1 - q)\delta > 0,$$

where the first inequality is because the consumer maximized her utility by buying in the discriminatory equilibrium, and the second inequality is because $\delta > 0$ and $q < 1$.

For consumers with values $v \in [p_M, 1]$, under the discriminatory equilibrium, they receive utility $v - p_1(q) + \delta q$ for buying the good at price $p_1(q)$ and being shown the better ad with probability q. Under the uniform advertising equilibrium B, they receive utility $v - p_M$ for buying at price p_M in period 1, and then being shown ad B with probability 1. These consumers will prefer the discriminatory equilibrium if and only if

$$v - p_1(q) + \delta q > v - p_M \iff \delta q > p_1(q) - p_M.$$

Intuitively, the term δq captures the consumer's bonus in period 2 from the possibility of being shown the better ad. The term $p_1(q) - p_M$ is the additional amount the consumer must pay in period 1 to get the good. Then the consumer will prefer discrimination whenever the increase in utility from seeing the better ad outweighs the increased price she must pay in period 1. □

Proposition 13 *In settings where a discriminatory equilibrium exists for a range of q, the equilibrium consumer surplus can be increasing in q.*

Proof. To simplify calculations, suppose again that valuations are distributed on the entire positive real line. Next consider ex-ante equilibrium welfare of the buyer as a function of q in the discriminatory equilibrium:

$$\int_{v^*(q)}^{\infty} (v - p_1(q)) f(v) dv + q\delta(1 - F(v^*(q))) + (1 - q)\delta F(v^*(q)).$$

Differentiating with respect to q and collecting terms, we have:

$$\begin{aligned} &- f(v^*(q)) v^{*\prime}(q) \left(v^*(q) - p_1(q) - (1 - 2q)\delta\right) \\ &\qquad + \delta(1 - 2F(v^*(q))) - p_1'(q)(1 - F(v^*(q))). \end{aligned}$$

Recall that by the definitions of $v^*(q)$, we have:

$$v^*(q) - p_1(q) - (1 - 2q)\delta = 0$$

Therefore, the derivative of ex-ante welfare in a discriminatory equilibrium w.r.t. q equals

$$\delta(1 - 2F(v^*(q))) - p_1'(q)(1 - F(v^*(q))).$$

Recall from the proof of Lemma 6 that we have:

$$\begin{aligned} & p_1'(q) - I'(p_1 + (1 - 2q)\delta)(p_1'(q) - 2\delta) = 0, \\ \Longrightarrow\ & p_1'(q) = 2q \left(\frac{I'(p_1 + (1 - 2q)\delta)}{I'(p_1 + (1 - 2q)\delta) - 1} \right). \end{aligned}$$

Substituting in, the derivative of welfare w.r.t. q equals:

$$\delta(1 - 2F(v^*(q))) - 2q \left(\frac{I'(p_1 + (1 - 2q)\delta)}{I'(p_1 + (1 - 2q)\delta) - 1} \right) (1 - F(v^*(q))).$$

Note that for I' small (e.g. close to 0 like an exponential distribution that has constant hazard rate), and $F(v^*) < \frac{1}{2}$, this is positive.

To specify one such game, let values be distributed according to an exponential distribution with parameter λ. This distribution has $I(v) = \frac{1}{\lambda}$ and $I'(v) = 0$. Then the derivative of welfare w.r.t q in this game is:

$$\delta(1 - 2F(v^*(q))) = \delta(1 - 2(1 - e^{-\lambda v^*(q)})) = \delta(-1 + 2e^{-\lambda v^*(q)}).$$

This is positive whenever $e^{-\lambda v^*(q)} > \frac{1}{2}$, or equivalently, whenever $\lambda v^*(q) < \ln 2$. Since $v^*(q) \in [0, 1]$ for all q, any $\lambda < \ln 2$ would suffice to ensure that consumer surplus is increasing in q. □

Part III

Eliciting Data from Privacy-Aware Agents

Chapter 6

PRIVATE AND TRUTHFUL LINEAR REGRESSION

6.1 Introduction

Fitting a linear model is perhaps the most fundamental and basic learning task, with diverse applications from statistics to experimental sciences like medicine and sociology. In many settings, the data from which a model is to be learnt are not held by the analyst performing the regression task, but must be elicited from individuals. Such settings clearly include medical trials and census surveys, as well as mining online behavioral data, a practice currently happening at a massive scale.

If data are held by self-interested individuals, it is not enough to simply run a regression—the data holders may wish to influence the outcome of the computation, either because they could benefit directly from certain outcomes, or to mask their input due to privacy concerns. In this case, it is necessary to model the utility functions of the individuals and to design mechanisms that provide proper incentives. Ideally, such mechanisms should still allow for accurate computation of the underlying regression. A tradeoff then emerges between the accuracy of the computation and the budget required to compensate participants.

In this chapter, we focus on the problem posed by data holders who are concerned with their privacy. Our approach can easily be generalized to handle individuals manipulating the computation's outcome for other reasons, but for clarity we treat only privacy concerns. We consider a population of players, each holding private data, and an analyst who wishes to compute a linear model from their data. The analyst must design a mechanism (a computation he will do and payments he will give the players) that incentivizes the players to provide information that will allow for accurate computation, while minimizing the payments the analyst must make.

We use a model of players' costs for privacy based on the framework of differential privacy introduced in Chapter 2, and the privacy cost models presented in Section 3.4. Incentivizing most players to truthfully report their data to the analyst constrains our design to mechanisms that are differentially private. This immediately creates a number of challenges: to provide privacy, players must be rewarded based on *noisy aggregates* of other players reports, rather than the actual reports; some players may have such high costs for privacy that they cannot be incentivized to report truthfully; the incentive for all others to report truthfully must overcome the impact of the non-participating players and the intentional introduction of noise for privacy; and, the intentional introduction of noise degrades the accuracy of the final computation. In addition, differentially private computation of a linear model necessarily produces a biased estimation; existing approaches [48] to design mechanisms to elicit data from privacy-sensitive individuals do not generalize well to biased estimators. Overcoming these challenges, through appropriate design of the computation and payment scheme, is the main technical contribution of this chapter.

Our Results

We study the problem of eliciting data from privacy-aware individuals in the context of linear regression. We present a mechanism (Algorithm 4), which, under appropriate choice of parameters and fairly mild technical assumptions, satisfies the following properties: it is (a) *accurate* (Theorem 18), i.e., computes an estimator whose squared ℓ_2 distance to the true linear model goes to zero as the number of players increases, (b) *asymptotically truthful* (Theorem 17), in that players have no incentive to misreport their data, (c) *individually rational* (Theorem 19), as players receive positive utility and are incentivized to participate, and (d) it requires an *asymptotically small budget* (Theorem 20), as total payments to players go to zero as the number of players increases. Our technical assumptions are on how individuals experience privacy losses and on the distribution from which these losses are drawn. Accuracy of the computation is attained by establishing that the algorithm provides differential privacy (Theorem 16), and that it provides payments such that the vast majority of players are incentivized to participate and to report truthfully (Theorems 17 and 19). An informal statement appears in Theorem 15.

The fact that our total budget decreases in the number of individuals in the population is an effect of the approach we use to eliciting truthful participation, which is based on the peer prediction technology and the model of agents' costs for privacy (both presented in Section 6.2). A similar effect was seen by Ghosh et al. [48]. As they note, costs would no longer tend to zero if our model incorporated some fixed cost for interacting with each individual.

Related Work

Starting with Ghosh and Roth [47], a series of papers have studied data acquisition problems from agents that have privacy concerns. The vast majority of this work [42, 69, 80, 24] operates in a model where players cannot lie about their data.

Their only recourse is to withhold it or perhaps to lie about their costs for privacy. A related thread [47, 78, 101, 18] explores cost models based on the notion of differential privacy. This latter body of work is discussed in more detail in Section 3.4.

Our setting is closest to, and inspired by, Ghosh et al. [48], who bring the technology of peer prediction to bear on the problem of incentivizing truthful reporting in the presence of privacy concerns. The peer prediction approach of Miller, Resnick, and Zeckhauser [74] incentivizes truthful reporting (in the absence of privacy constraints) by rewarding players for reporting information that is predictive of the reports of other agents. This allows the analyst to leverage correlations between players' information. Ghosh et al. [48] adapt the peer prediction approach to overcome a number of challenges presented by privacy-sensitive individuals. The mechanism and analysis of Ghosh et al. [48] was for the simplest possible statistic: the sum of private binary types. In contrast, here we regress a linear model over player data, a significantly more sophisticated learning task. In particular, to attain accurate, privacy-preserving linear regression, we deal with biased private estimators, which interferes with our ability to incentivize truth-telling, and hence compute an accurate statistic.

Linear regression under strategic agents has been studied in a variety of different contexts. Dekel, Fischer, and Procaccia [32] consider an analyst that regresses a "consensus" model across data coming from multiple strategic agents; agents would like the consensus value to minimize a loss over their own data, and they show that, in this setting, empirical risk minimization is group strategyproof. A similar result, albeit in a more restricted setting, is established by Perote and Perote-Pena [83]. Regressing a linear model over data from strategic agents that can only manipulate their costs, but not their data, was studied by Horel, Ioannidis, and Muthukrishnan [60] and Cai, Daskalakis, and Papadimitriou [10], while Ioannidis and Loiseau [63] consider a setting without payments, in which agents receive a utility as a function of estimation accuracy. We depart from the above approaches by considering agents whose utilities depend on their loss of *privacy*, an aspect absent from the above works.

Finally, we note a growing body of work on differentially private empirical risk minimization. Our mechanism is based on the outcome perturbation algorithm of Chaudhuri, Monteleoni, and Sarwate [14]. Other algorithms from this literature—such as the localization algorithm of Bassily, Smith, and Thakurta [4] or objective perturbation of Chaudhuri, Monteleoni, and Sarwate [14]—could be used instead, and would likely yield even better accuracy guarantees. We chose the output perturbation mechanism because it provides an explicit characterization of the noise added to preserve privacy, which allows the analysis to better highlight the challenges of incorporating privacy into our setting.

6.2 Model and Preliminaries

In this section we present our model and technical preliminaries on regression, peer prediction, and the analyst's mechanism design objectives.

A Regression Setting

We consider a population where each player $i \in [n] \equiv \{1, \ldots, n\}$ is endowed data consisting of a vector $x_i \in \mathbb{R}^d$ (i.e., player i's *features*) and a variable $y_i \in \mathbb{R}$ (i.e., her *response* variable). We assume that responses are linearly related to the features. That is, there exists a $\theta \in \mathbb{R}^d$ such that

$$y_i = \theta^\top x_i + z_i, \quad \text{for all } i \in [n], \tag{6.1}$$

where z_i are zero-mean noise variables.

An analyst wishes to infer a linear model from the players' data; that is, he wishes to estimate θ, e.g., by performing linear regression on the players' data. However, players incur a privacy cost from revelation of their data and need to be properly incentivized to truthfully reveal it to the analyst. More specifically, we assume, as in Ioannidis and Loiseau [63], that player i can manipulate her responses y_i but not her features x_i. This is indeed the case when features are measured directly by the analyst (e.g., are observed during a physical examination or are measured in a lab) or are verifiable (e.g., features are extracted from a player's medical record or are listed on her ID). A player may misreport her response y_i, on the other hand, which is unverifiable; this would be the case if, e.g., y_i is the answer the player gives to a survey question about her preferences or habits.

We assume that players are strategic and may lie either to increase the payment they extract from the analyst or to mitigate any privacy violation they incur from the disclosure of their data. To address such strategic behavior, the analyst will design a mechanism $\mathcal{M} : (\mathbb{R}^d \times \mathbb{R})^n \to \mathbb{R}^d \times \mathbb{R}^n_+$ that takes as input all player data (namely, the features x_i and possibly perturbed responses $\hat{y}_i$ of all players), and outputs an estimate $\hat{\theta}$ and a set of non-negative payments $\{\pi_i\}_{i\in[n]}$ to the players. Informally, we seek mechanisms that allow for *accurate* estimation of θ while requiring only asymptotically *small budget*. In order to ensure accurate estimation of θ, we will require that our mechanism *incentivizes truthful participation* on the part of most players, which in turn will require that we provide an appropriate *privacy guarantee*. We ensure privacy by requiring the analyst to use a mechanism that satisfies differential privacy, as defined in Chapter 2. Clearly, all of the above also depend on the players' rational behavior and, in particular, their utilities; we formally present our model of player utilities in the next subsection.

Throughout our analysis, we assume that θ is drawn independently from a known distribution $\mathcal{F}$, the attribute vectors x_i are drawn independently from the uniform distribution on the d-dimensional unit ball,[1] and the noise terms z_i are drawn independently from a known distribution $\mathcal{G}$. Thus θ, $\{x_i\}_{i\in[n]}$, and $\{z_i\}_{i\in[n]}$ are independent random variables, while responses $\{y_i\}_{i\in[n]}$ are determined according to (6.1). As a result, responses are conditionally independent given θ.

We require some additional bounded support assumptions on these distributions. In short, these boundedness assumptions are needed to ensure the sensitivity of mechanism $\mathcal{M}$ is finite; it is also natural in practice that both features and responses take values in a bounded domain. More precisely, we assume that the distribution $\mathcal{F}$ has bounded support, such that $\|\theta\|_2^2 \leq B$ for some constant B; we also require the noise distribution $\mathcal{G}$ to have mean zero, finite variance σ^2, and bounded support: $\mathsf{supp}(\mathcal{G}) = [-M, M]$ for some constant M. These assumptions together imply that $|\theta^\top x_i| \leq B$ and $|y_i| \leq B + M$.

Mechanism Design Objectives: Privacy and Utility

The analyst must design a mechanism $\mathcal{M} : (\mathbb{R}^d \times \mathbb{R})^n \to \mathbb{R}^d \times \mathbb{R}_+^n$ that receives as input $(x_i, \hat{y}_i)$ from each player i, and outputs an estimate $\hat{\theta} \in \mathbb{R}^d$ and a non-negative payment π_i to each player. We seek mechanisms that satisfy the following properties: (a) truthful reporting is a Bayes Nash equilibrium, (b) players are ensured non-negative utilities from truthful reporting, (c) the estimator computed under truthful reporting is highly accurate, (d) the budget required from the analyst to run the mechanism is small, and (e) the mechanism is jointly differentially private.

To satisfy the first two conditions, we require $\mathcal{M}$ to be *Bayes Nash incentive compatible* (Definition 13) and *interim individually rational* (Definition 14), both defined formally in Section 3.1. Our *accuracy* notion is the *mean squared error* (with

[1] See Theorem 14 and its accompanying Remark for a discussion of generalizing beyond the uniform distribution.

respect to the ℓ_2 norm) between the estimate $\hat{\theta}$ and the true parameter θ.

Definition 25 (Accuracy)**.** *A regression is* η-accurate *if for all realizable parameters* θ*, it outputs an estimate* $\hat{\theta}$ *such that* $\mathbb{E}\left[\|\hat{\theta}-\theta\|_2^2\right] \leq \eta$*.*

We will also be concerned with the total amount spent by the analyst in the mechanism. The *budget* $\mathcal{B}$ of a mechanism is the sum of all payments made to players. That is, $\mathcal{B} = \sum_i \pi_i$.

Definition 26 (Asymptotically small budget)**.** *An* asymptotically small budget *is such that* $\mathcal{B} = \sum_{i=1}^{n} \pi_i(X, y) = o(1)$,*for all realizable* (X, y).

In our setting, joint differential privacy is the appropriate privacy notion because it makes sense to assume that the payment to a player is in some sense "private," in that it is shared neither publicly nor with other players. To that end, we assume that the estimate $\hat{\theta}$ computed by the mechanism $\mathcal{M}$ is a publicly observable output; in contrast, each payment π_i is observable *only by player* i. Hence, from the perspective of each player i, the mechanism output that is publicly released and that, in turn, might violate her privacy, is $(\hat{\theta}, \pi_{-i})$, where π_{-i} comprises all payments excluding player i's payment. The analyst will design a mechanism $\mathcal{M}$ that computes $\hat{\theta}$ under the constraint of differential privacy. The payment π_i to each player will be computed as function of only their input $(x_i, \hat{y}_i)$, and by the Billboard Lemma (Lemma 1), the mechanism $\mathcal{M}$ will be jointly differentially private.

The use of joint differential privacy is natural, but it is also necessary to incentivize truthfulness. Requiring that a player's payment π_i be ϵ-differentially private implies that a player's unilateral deviation changes the distribution of her payment only slightly. As discussed in Sections 3.2 and 3.3, under full differential privacy, a player's payment would remain roughly the same no matter what she reports, which cannot incentivize truthful reporting.

We adopt the modeling assumptions described in Section 3.4: we assume that each player i has a cost parameter $c_i \in \mathbb{R}_+$, which determines her sensitivity to the privacy violation incurred by the revelation of her data to her analyst, and a privacy cost function $f_i(c_i, \epsilon)$, which describes the cost she incurs when her data is used in an ϵ-jointly differentially private computation. We also assume that players have *quasilinear utilities*, so if player i receives payment π_i for her report, and experiences cost $f_i(c_i, \epsilon)$ from her privacy loss, her utility is $u_i = \pi_i - f_i(c_i, \epsilon)$.

We make the following assumption, motivated in Section 3.4, that privacy cost functions are upper bounded by a quadratic function of ϵ.

Assumption 4 ([18])**.** *The privacy cost function of each player i satisfies*

$$f_i(c_i, \epsilon) \le c_i \epsilon^2.$$

Throughout our analysis, we assume that the privacy cost parameters are also random variables, sampled from a distribution C. We allow c_i to depend on player i's data (x_i, y_i); however, we assume conditioned on (x_i, y_i), that c_i does not reveal any additional information about the costs or data of any other agents. Formally:

Assumption 6. *Given (x_i, y_i), (X_{-i}, y_{-i}, c_{-i}) is conditionally independent of c_i:*

$$\Pr[(X_{-i}, y_{-i}, c_{-i})|(x_i, y_i), c_i] = \Pr[(X_{-i}, y_{-i}, c_{-i})|(x_i, y_i), c_i'],$$

for all (X_{-i}, y_{-i}, c_{-i}), (x_i, y_i), c_i, c_i'.

We also make the following additional technical assumption on the tail of C.

Assumption 7. *The conditional marginal distribution satisfies*

$$\min_{x_i, y_i} \left(\Pr_{c_j \sim C | x_i, y_i}[c_j \le \tau] \right) \ge 1 - \tau^{-p},$$

for some constant $p > 1$.

Note that Assumption 7 implies that $\Pr_{c_i \sim C}[c_i \le \tau] \ge 1 - \tau^{-p}$.

Linear and Ridge Regression

In this section we review some basic properties of linear regression and ridge regression, the methods the analyst can employ to estimate the parameter vector $\theta \in \mathbb{R}^d$. Let $X = [x_i]_{i \in [n]} \in \mathbb{R}^{n \times d}$ denote the $n \times d$ matrix of features, and $y = [y_i]_{i \in [n]} \in \mathbb{R}^n$ the vector of responses. Estimating θ through *ridge regression* amounts to minimizing the following regularized quadratic loss function:

$$\mathcal{L}(\theta; X, y) = \sum_{i=1}^{n} \ell(\theta; x_i, y_i) = \sum_{i=1}^{n} (y_i - \theta^\top x_i)^2 + \gamma \|\theta\|_2^2 . \tag{6.2}$$

The ridge regression estimator can be written as:

$$\hat{\theta}^R = \arg\min_{\theta \in \mathbb{R}^d} \mathcal{L}(\theta; X, y) = (\gamma I + X^\top X)^{-1} X^\top y, \tag{6.3}$$

where I is the $d \times d$ identity matrix.

The parameter $\gamma > 0$, known as the regularization parameter, ensures that the loss function is *strongly convex* (Definition 27) and, in particular, that the minimizer $\hat{\theta}^R$ of (6.2) is unique.

Definition 27 (Strong Convexity)**.** *A function $f : \mathbb{R}^d \to \mathbb{R}$ is η-strongly convex if*

$$H\left(f(\theta)\right) - \eta I \text{ is positive semi-definite for all } \theta \in \mathbb{R}^d,$$

where $H(f(\theta))$ is the Hessian[2] *of f, and I is the $d \times d$ identity matrix.*

Strong convexity requires that the eigenvalues of the Hessian of f are bounded away from zero. Note that when f is a one-dimensional function ($d = 1$), strong convexity reduces to the requirement that the second derivative of f is strictly positive everywhere: $f''(\theta) \geq \eta > 0$ for all $\theta \in \mathbb{R}^d$. Lemma 8 shows that regularizing the quadratic loss $\mathcal{L}$ ensures that it is strongly convex. The proof is in the Appendix of Section 6.6.

Lemma 8. *$\mathcal{L}(\theta; X, y)$ is 2γ-strongly convex in θ.*

When $\gamma = 0$, the estimator is the standard *linear regression* estimator, which we denote by $\hat{\theta}^L = (X^\top X)^{-1} X^\top y$. The linear regression estimator is unbiased, i.e., it satisfies $\mathbb{E}[\hat{\theta}^L] = \theta$. The same is not true when $\gamma > 0$; the general ridge regression estimator $\hat{\theta}^R$ is *biased.*

Nevertheless, in practice $\hat{\theta}^R$ is preferable to $\hat{\theta}^L$ as it can achieve a desirable trade-off between *bias* and *variance*. To see this, consider the mean squared loss error of the estimation $\hat{\theta}^R$, namely, $\mathbb{E}[||\hat{\theta}^R - \theta||_2^2]$, which can be written as:

$$\begin{aligned}\mathbb{E}[||\hat{\theta}^R - \theta||_2^2] &= \mathbb{E}[||\hat{\theta}^R - \mathbb{E}[\hat{\theta}^R]||_2^2] + ||\mathbb{E}[\hat{\theta}^R] - \theta||_2^2 \\ &= \mathtt{trace}(\mathtt{Cov}(\hat{\theta}^R)) + ||\,\mathtt{bias}(\hat{\theta}^R)||_2^2,\end{aligned}$$

where $\mathtt{Cov}(\hat{\theta}^R) = \mathbb{E}[(\hat{\theta}^R - \mathbb{E}[\hat{\theta}^R])(\hat{\theta}^R - \mathbb{E}[\hat{\theta}^R])^\top]$ and $\mathtt{bias}(\hat{\theta}^R) = \mathbb{E}[\hat{\theta}^R] - \theta$ are the covariance and bias, respectively, of estimator $\hat{\theta}^R$. These can be computed in

[2] The *Hessian* H of function f is a $d \times d$ matrix of its partial second derivatives, where

$$H(f(\theta))_{jk} = \frac{\partial^2 f(\theta)}{\partial\theta_j \partial\theta_k}.$$

A $d \times d$ matrix A is *positive semi-definite* (PSD) if for all $v \in \mathbb{R}^d$, $v^\top A v \geq 0$.

closed form as:

$$\texttt{Cov}(\hat{\theta}^R) = \sigma^2(\gamma I + X^\top X)^{-1}X^\top X(\gamma I + X^\top X)^{-1}, \tag{6.4}$$

$$\texttt{bias}(\hat{\theta}^R) = -\gamma(\gamma I + X^\top X)^{-1}\theta, \tag{6.5}$$

where σ^2 is the variance of the noise variables z_i in Equation (6.1). It is easy to see that decreasing γ decreases the bias, but may significantly increase the variance. For example in the case where $\text{rank}(X) < d$, the matrix $X^\top X$ is not invertible, and the trace of the covariance tends to infinity as γ tends to zero.

Whether $\texttt{trace}(\texttt{Cov}(\hat{\theta}^R))$ is large and, therefore, whether regularizing the square loss is necessary, depends on largest eigenvalue (i.e., the *spectral norm*) of $(X^\top X)^{-1}$. Although this can be infinite for arbitrary X, by the law of large numbers, we expect that if we sample the features x_i independently from an isotropic distribution, then $\frac{1}{n}(X^\top X)$ should converge to the covariance of this distribution (namely $\Sigma = cI$ for some constant c). As such, for large n both the largest and smallest eigenvalues of $X^\top X$ should be of the order of n, leading to an estimation of ever decreasing variance even when $\gamma = 0$. The following theorem, which follows as a corollary of a result by Vershynin [98], formalizes this notion, providing bounds on both the largest and smallest eigenvalue of $X^\top X$ and $\gamma I + X^\top X$. The proof is deferred to the Appendix in Section 6.6.

Theorem 14. *Let $\xi \in (0,1)$, and $t \geq 1$. Let $\|\cdot\|$ denote the spectral norm. If $\{x_i\}_{i\in[n]}$ are i.i.d. and sampled uniformly from the unit ball, then with probability at least $1 - d^{-t^2}$, when $n \geq C(\frac{t}{\xi})^2(d+2)\log d$, for some absolute constant C, then,*

$$\left\|X^\top X\right\| \leq (1+\xi)\frac{1}{d+2}n, \textit{ and } \left\|(X^\top X)^{-1}\right\| \leq \frac{1}{(1-\xi)\frac{1}{d+2}n}, \textit{ and}$$

$$\left\|\gamma I + X^\top X\right\| \leq \gamma + (1+\xi)\frac{1}{d+2}n, \textit{ and } \left\|(\gamma I + X^\top X)^{-1}\right\| \leq \frac{1}{\gamma + (1-\xi)\frac{1}{d+2}n}.$$

Remark A generalization of Theorem 14 holds for $\{x_i\}_{i\in[n]}$ sampled from any distribution with a covariance Σ whose smallest eigenvalue is bounded away from zero (see Vershynin [98]). We restrict our attention to the unit ball for simplicity and concreteness.

Peer Prediction and the Brier Scoring Rule

Peer prediction [74] is a useful method of inducing truthful reporting among players that hold data generated by the same statistical model. In short, each player reports her data to an analyst and is paid based upon how well her report predicts the report of other players; tying each player's payment to how closely it predicts peer reports is precisely what induces truthfulness.

We employ peer prediction to elicit data from privacy-aware players through the use of *strictly proper scoring rules*. A scoring rule takes as input a parameter value—either directly observed or computed from the reports of other players—and a player's reported belief about the parameter value, and outputs a payment. A strictly proper scoring rule is one in which the expected payment is uniquely maximized by truthful reporting. In particular, we will use the *Brier scoring rule* [9], which is known to be strictly proper.

Definition 28 (Brier scoring rule [9])**.** *The* Brier scoring rule *is the function B : $\mathbb{R}^2 \to \mathbb{R}$, defined as:*

$$B(p,q) = 1 - 2(p - 2pq + q^2).$$

The Brier scoring rule can used as a payment scheme in peer prediction mechanisms to elicit reports q from players about the value of a parameter p. Note that when the Brier scoring rule is used for the prediction of a binary event $\omega \in \{0,1\}$, the function reduces to $B(\omega,q) = 1-2(q-\omega)^2$, and a player's report q can be interpreted about her prediction of the probability p with which the event occurs.

The mechanisms in this chapter (Algorithms 3 and 4) use payment rule $B_{a,b}(p,q)$, which is a parametrized rescaling of the scoring rule $B(p,q)$, defined as follows:

$$B_{a,b}(p,q) = a - b\left(p - 2pq + q^2\right).$$

Any positive-affine transformation of a strictly proper scoring rule remains strictly proper [7]. The rescaled Brier scoring rule satisfies this criterion as $B_{a,b}(p,q) = a' + b'B(p,q)$ where $a' = a - b/2$ and $b' = b/2$. Thus $B_{a,b}(p,q)$ is a strictly proper scoring rule for $b > 0$, and is uniquely maximized by reporting the true belief $q = p$.

In our setting, the analyst asks each player i to report $(x_i, \hat{y}_i)$ and then computes an estimate $\hat{\theta}$ of the true parameter θ. The analyst performs a Bayesian update on behalf of each player based on her reported data, and pays the player based upon how closely the expectation of the posterior belief matched $\hat{\theta}$ according to the scoring rule $B_{a,b}(p,q)$. The parameter p being estimated by each player i is $x_i^\top \hat{\theta}$, which is the inner product of player i's features x_i with an estimated parameter $\hat{\theta}$. Her input belief q will be $x_i^\top \mathbb{E}[\theta | x_i, \hat{y}_i]$, which is the inner product of her features

Algorithm 3 Truthful Regression Mechanism(a, b)

```
Solicit reports $X \in (\mathbb{R}^d)^n$ and $\hat{y} \in \mathbb{R}^n$
Analyst computes $\hat{\theta}^L = (X^\top X)^{-1} X^\top \hat{y}$ and $\hat{\theta}^L_{-i} = (X_{-i}^\top X_{-i})^{-1} X_{-i}^\top \hat{y}_{-i}$ for each $i \in [n]$
Output estimator $\hat{\theta}^L$
Pay each player $i$, $\pi_i = B_{a,b}(x_i^\top \hat{\theta}^L_{-i}, x_i^\top \mathbb{E}[\theta | x_i, \hat{y}_i])$
```

x_i with her posterior beliefs about θ, given her *reported* data x_i and $\hat{y}_i$. Players are also able to opt-out of the mechanism by reporting $q = \perp$; we define $B_{a,b}(p, \perp) = 0$. Since the x_i are verifiable by the analyst, a misreport of x_i will be interpreted as a report of $\hat{y}_i = \perp$.

6.3 Truthful Regression without Privacy Constraints

To illustrate the ideas we use in the rest of the chapter we present in this section a mechanism which incentivizes truthful reporting in the absence of privacy concerns. If the players do not have privacy concerns (i.e., $c_i = 0$ for all $i \in [n]$), the analyst can simply collect data, estimate θ using linear regression, and compensate players using the following rescaled Brier scoring rule presented in Section 6.2:

$$B_{a,b}(p, q) = a - b\left(p - 2pq + q^2\right).$$

The Truthful Regression Mechanism is formally presented in Algorithm 3. In the spirit of peer prediction, a player's payment depends on how well her reported $\hat{y}_i$ agrees with the predicted value of y_i, as constructed by the estimate $\hat{\theta}^L_{-i}$ of θ produced by all her peers. We now show that truthful reporting is a Bayes Nash equilibrium.

Lemma 9 (Truthfulness)**.** *For all $a, b > 0$, truthful reporting is a Bayes Nash equilibrium under Algorithm 3.*

Proof. Recall that conditioned on x_i, y_i, the distribution of X_{-i}, y_{-i} is independent of c_i. Hence, assuming all other players are truthful, player i's expected payment conditioned on her data (x_i, y_i) and her cost c_i, for reporting $\hat{y}_i$ is

$$\begin{aligned}\mathbb{E}[\pi_i | x_i, y_i, c_i] &= \mathbb{E}\left[B_{a,b}(x_i^\top \hat{\theta}^L_{-i}, x_i^\top \mathbb{E}[\theta | x_i, \hat{y}_i]) | x_i, y_i\right] \\ &= B_{a,b}\left(x_i^\top \mathbb{E}[\hat{\theta}^L_{-i} | x_i, y_i], x_i^\top \mathbb{E}[\theta | x_i, \hat{y}_i]\right).\end{aligned}$$

The second equality is due to the linearity of $B_{a,b}$ in its first argument, as well as the linearity of the inner product. Note that $B_{a,b}$ is uniquely maximized by reporting

$\hat{y}_i$ such that $\mathbb{E}[\theta|x_i, \hat{y}_i]^\top x_i = \mathbb{E}[\hat{\theta}^L_{-i}|x_i, y_i]^\top x_i$. Since $\hat{\theta}^L$ is an unbiased estimator of θ, then $\mathbb{E}[\hat{\theta}^L_{-i}|x_i, y_i] = \mathbb{E}[\theta|x_i, y_i]$. Thus the optimal misreport is $\hat{y}_i$ such that $\mathbb{E}[\theta|x_i, \hat{y}_i]^\top x_i = \mathbb{E}[\theta|x_i, y_i]^\top x_i$, so truthful reporting is a Bayes Nash equilibrium. □

We note that truthfulness is essentially a consequence of (1) the fact that $B_{a,b}$ is a strictly proper scoring rule (as it is positive-affine in its first argument and strictly concave in its second argument), and (2) most importantly, the fact that $\hat{\theta}^L_{-i}$ is an unbiased estimator of θ. Moreover, as in the case of the simple peer prediction setting presented in Section 6.2, truthfulness persists even if $\hat{\theta}^L_{-i}$ in Algorithm 3 is replaced by a linear regression estimator constructed over responses restricted to an arbitrary set $S \subseteq [n] \setminus i$.

Truthful reports enable accurate computation of the estimator with high probability, with accuracy parameter $\eta = O(\frac{1}{n})$.

Lemma 10 (Accuracy). *Under truthful reporting, with probability at least* $1 - d^{-t^2}$ *and when* $n \geq C(\frac{t}{\xi})^2(d+2)\log d$*, the accuracy the estimator* $\hat{\theta}^L$ *in Algorithm 3 is* $\mathbb{E}\left[\left\|\hat{\theta}^L - \theta\right\|_2^2\right] \leq \frac{\sigma^2}{(1-\xi)\frac{1}{d+2}n}$.

Proof. Note that $\mathbb{E}\left[\left\|\hat{\theta}^L - \theta\right\|_2^2\right] = \mathtt{trace}(\mathtt{Cov}(\hat{\theta}^L)) \overset{(6.4)}{=} \sigma^2\,\mathtt{trace}\left((X^\top X)^{-1}\right)$. For i.i.d. features x_i, the spectrum of matrix $X^\top X$ can be asymptotically characterized Theorem 14, and the lemma follows. □

Remark Note that individual rationality and a small budget can be trivially attained in the absence of privacy costs. To ensure individual rationality of Algorithm 3, payments π_i must be non-negative, but can be made arbitrarily small. Thus payments can be scaled down to reduce the analyst's total budget. For example, setting $a = b(B + 2B(B+M) + (B+M)^2 - 1)$ and $b = \frac{1}{n^2}$ ensures $\pi_i \geq 0$ for all players i, and the total required budget is $\frac{1}{n}(2B + 4B(B+M) + (B+M)^2) = O(\frac{1}{n})$.

6.4 Truthful Regression with Privacy Constraints

As we saw in the previous section, in the absence of privacy concerns, it is possible to devise payments that incentivize truthful reporting. These payments compensate players based on how well their report agrees with a response predicted by $\hat{\theta}^L$ estimated using other player's reports.

Players whose utilities depend on privacy raise several challenges. Recall that the parameters estimated by the analyst, and the payments made to players, need to satisfy joint differential privacy, and hence any estimate of θ revealed publicly by the analyst or used in a payment must be ϵ-differentially private. Unfortunately, the sensitivity of the linear regression estimator $\hat{\theta}^L$ to changes in the input data is, in general, unbounded. As a result, it is not possible to construct a non-trivial differentially private version of $\hat{\theta}^L$ by, e.g., adding noise to its output.

In contrast, differentially private versions of regularized estimators like the ridge regression estimator $\hat{\theta}^R$ can be constructed. Recent techniques have been developed for precisely this purpose, not only for ridge regression but for the broader class of learning through (convex) empirical risk minimization [14, 4]. In short, the techniques of Chaudhuri, Monteleoni, and Sarwate [14] and Bassily, Smith, and Thakurta [4] succeed precisely because, for $\gamma > 0$, the regularized loss of Equation (6.2) is *strongly convex* as shown in Lemma 8. This implies that the sensitivity of $\hat{\theta}^R$ is bounded, and a differentially private version of $\hat{\theta}^R$ can be constructed by adding noise of appropriate variance or though alternative techniques such as objective perturbation.

The above results suggest that a possible approach to constructing a truthful, accurate mechanism in the presence of privacy-conscious players is to modify Algorithm 3 by replacing $\hat{\theta}^L$ with a ridge regression estimator $\hat{\theta}^R$, both with respect to the estimate released globally and to any estimates used in computing payments. Unfortunately, such an approach breaks the mechanism's truthfulness guarantee because $\hat{\theta}^R$ is a biased estimator. The linear regression estimator $\hat{\theta}^L$ ensured that the scoring rule $B_{a,b}$ was maximized precisely when players reported their response variable truthfully. However, in the presence of a biased estimator, it can easily be seen that the optimal report of player i deviates from truthful reporting by a quantity proportional to the expected bias.

We address this issue for large n using again the concentration result of Theorem 14. This ensures that for large n, the spectrum of $X^\top X$ should grow roughly linearly with n, with high probability. By Equations (6.4), this implies that as long as γ grows more slowly than n, the bias term of $\hat{\theta}^R$ converges to zero with high probability. Together, these statements ensure that for an appropriate choice of γ, we attain approximate truthfulness for large n, while also ensuring that the output of our mechanism remains differentially private for all n. We formalize this intuition in the next section by presenting a mechanism based on ridge regression, and proving that it indeed attains approximate truthfulness for large n, while also remaining jointly differentially private.

Private Regression Mechanism

Our Private Regression Mechanism is presented in Algorithm 4, which is a differentially private varaint of the Truthful Regression Mechanism in Algorithm 3. We incorporate into our mechanism the Output Perturbation algorithm from Chaudhuri, Monteleoni, and Sarwate [14], which first computes the ridge regression estimator and then adds noise to the output. This approach is used to ensure that the mechanism's output satisfies joint differential privacy.

The noise vector v is drawn according to the following distribution P_L, which is a high-dimensional Laplace distribution with parameter $\frac{4B+2M}{\gamma\epsilon}$:

$$P_L(v) \propto \exp\left(\frac{-\gamma\epsilon}{4B+2M}\, \|v\|_2\right).$$

Algorithm 4 Private Regression Mechanism(γ, ϵ, a, b)

```
  Solicit reports $X \in \left(\mathbb{R}^d\right)^n$ and $\hat{y} \in \mathbb{R}^n$
  Randomly partition players into two groups, with respective data pairs $(X_0, \hat{y}_0)$ and $(X_1, \hat{y}_1)$
  Analyst computes $\hat{\theta}^R = (\gamma I + X^\top X)^{-1} X^\top \hat{y}$ and $\hat{\theta}^R_j = (\gamma I + X_j^\top X_j)^{-1} X_j^\top \hat{y}_j$ for $j = 0, 1$
  Independently draw $v, v_0, v_1 \in \mathbb{R}^d$ according to distribution $P_L$
  Compute estimators $\hat{\theta}^P = \hat{\theta}^R + v$, $\hat{\theta}^P_0 = \hat{\theta}^R_0 + v_0$, and $\hat{\theta}^P_1 = \hat{\theta}^R_1 + v_1$
  Output estimator $\hat{\theta}^P$
  Pay each player $i$ in group $j$, $\pi_i = B_{a,b}((\hat{\theta}^P_{1-j})^\top x_i, \mathbb{E}[\theta|x_i, \hat{y}_i]^\top x_i)$ for $j = 0, 1$
```

Here we state an informal version of our main result. The formal version of this result is stated in Corollary 2, which aggregates and instantiates Theorems 16, 17, 18, 19, and 20.

Theorem 15 (Main result (Informal))**.** *Under Assumptions 4, 6, and 7, there exist ways to set γ, ϵ, a, and b in Algorithm 4 to ensure that with high probability:*

1. *the output of Algorithm 4 is $o(\frac{1}{\sqrt{n}})$-jointly differentially private,*
2. *it is an $o\left(\frac{1}{n}\right)$-approximate Bayes Nash equilibrium for a $(1 - o(1))$-fraction of players to truthfully report their data,*
3. *the computed estimator $\hat{\theta}^P$ is $o(1)$-accurate,*

4. *it is individually rational for a* $(1 - o(1))$*-fraction of players to participate in the mechanism, and*

5. *the required budget from the analyst is* $o(1)$.

6.5 Analysis of Algorithm 4

In this section, we flesh out the claims made in Theorem 15. All proofs are deferred to Appendix 6.6.

Theorem 16 (Privacy)**.** *The mechanism in Algorithm 4 is* 2ϵ*-jointly differentially private.*

Proof idea We first show that the estimators $\hat{\theta}^P, \hat{\theta}^P_0, \hat{\theta}^P_1$ computed by Algorithm 4 together satisfy 2ϵ-differential privacy, by bounding the maximum amount that any player's report can affect the estimators. We then use the Billboard Lemma (Lemma 1) to show that the estimators, together with the vector of payments, satisfy 2ϵ-joint differential privacy.

Once we have established the privacy guarantee, we can build on this to get truthful participation and hence accuracy. To do so, we first show that a symmetric threshold strategy equilibrium exists, in which all agents with cost parameter c_i below some threshold τ should participate and truthfully report their y_i. We define $\tau_{\alpha,\beta}$ to be the cost threshold such that (1) with probability $1 - \beta$ (with respect to the prior from which costs are drawn), at least a $(1 - \alpha)$-fraction of players have cost parameter $c_i \leq \tau_{\alpha,\beta}$, and (2) conditioned on her own data, each player i believes that with probability $1 - \alpha$, any other player j will have cost parameter $c_j \leq \tau_{\alpha,\beta}$.

Definition 29 (Threshold $\tau_{\alpha,\beta}$)**.** *Fix a marginal cost distribution* $\mathcal{C}$ *on* $\{c_i\}$*, and let*

$$\tau^1_{\alpha,\beta} = \inf_{\tau} \left(\Pr_{c\sim\mathcal{C}}\left[|\{i : c_i \leq \tau\}| \geq (1-\alpha)n\right] \geq 1-\beta\right),$$

$$\tau^2_{\alpha} = \inf_{\tau}\left(\min_{x_i,y_i}\left(\Pr_{c_j\sim\mathcal{C}|x_i,y_i}[c_j \leq \tau]\right) \geq 1-\alpha\right).$$

Define $\tau_{\alpha,\beta}$ *to be the larger of these thresholds:* $\tau_{\alpha,\beta} = \max\{\tau^1_{\alpha,\beta}, \tau^2_{\alpha}\}$.

We also define the threshold strategy σ_τ, in which a player reports truthfully if her cost c_i is below τ, and is allowed to misreport arbitrarily if her cost is above τ.

Definition 30 (Threshold strategy)**.** *Define the threshold strategy σ_τ as follows:*

$$\sigma_\tau(x_i, y_i, c_i) = \begin{cases} \textit{Report } \hat{y}_i = y_i & \textit{if } c_i \le \tau, \\ \textit{Report arbitrary } \hat{y}_i & \textit{otherwise.} \end{cases}$$

We show that $\sigma_{\tau_{\alpha,\beta}}$ forms a symmetric threshold strategy equilibrium in the Private Regression Mechanism of Algorithm 4.

Theorem 17 (Truthfulness)**.** *Fix a participation goal $1-\alpha$, a privacy parameter ϵ, a desired confidence parameter β, $\xi \in (0,1)$, and $t \ge 1$. Then under Assumptions 4 and 6, with probability $1 - d^{t^2}$ and when $n \ge C(\frac{t}{\xi})^2(d+2)\log d$, the symmetric threshold strategy $\sigma_{\tau_{\alpha,\beta}}$ is an η-approximate Bayes-Nash equilibrium in Algorithm 4 for*

$$\eta = b\left(\frac{\alpha n}{\gamma}(4B + 2M) + \frac{\gamma B}{\gamma + (1-\xi)\frac{1}{d+2}n}\right)^2 + \tau_{\alpha,\beta}\epsilon^2.$$

Proof idea There are three primary sources of error which cause the estimator $\hat{\theta}^P$ to differ from a player's posterior on θ. First, ridge regression is a biased estimation technique; second, Algorithm 4 adds noise to preserve privacy; third, players with cost parameter c_i above threshold $\tau_{\alpha,\beta}$ are allowed to misreport their data. We show how to control the effects of these three sources of error, so that $\hat{\theta}^P$ is "not too far" from a player's posterior on θ. Finally, we use strong convexity of the payment rule to show that any player's payment from misreporting is at most η greater than from truthful reporting.

Theorem 18 (Accuracy)**.** *Fix a participation goal $1-\alpha$, a privacy parameter ϵ, a desired confidence parameter β, $\xi \in (0,1)$, and $t \ge 1$. Then under the symmetric threshold strategy $\sigma_{\tau_{\alpha,\beta}}$, Algorithm 4 will output an estimator $\hat{\theta}^P$ such that with probability at least $1 - \beta - d^{-t^2}$, and when $n \ge C(\frac{t}{\xi})^2(d+2)\log d$,*

$$\mathbb{E}[||\hat{\theta}^P - \theta||_2^2] = O\left(\left(\frac{\alpha n}{\gamma} + \frac{1}{\gamma\epsilon}\right)^2 + \left(\frac{\gamma}{n}\right)^2 + \left(\frac{1}{n}\right)^2 + \frac{\alpha n}{\gamma} + \frac{1}{\gamma\epsilon}\right).$$

Proof idea As in Theorem 17, we control the three sources of error in the estimator $\hat{\theta}^P$ — the bias of ridge regression, the noise added to preserve privacy, and the error due to some players misreporting their data — this time measuring distance with respect to the expected ℓ_2 norm difference.

We next see that players whose cost parameters are below the threshold $\tau_{\alpha,\beta}$ are incentivized to participate.

Theorem 19 (Individual Rationality)**.** *Under Assumption 4, the mechanism in Algorithm 4 is individually rational for all players with cost parameters $c_i \leq \tau_{\alpha,\beta}$ as long as,*

$$a \geq \left(\frac{\alpha n}{\gamma}(4B + 2M) + \frac{\gamma B}{\gamma + (1 - \xi)\frac{1}{d+2}n} + B \right)(b + 2bB) + bB^2 + \tau_{\alpha,\beta}\epsilon^2,$$

regardless of the reports from players with cost coefficients above $\tau_{\alpha,\beta}$.

Proof idea A player's utility from participating in the mechanism is her payment minus her privacy cost. The parameter a in the payment rule is a constant offset that shifts each player's payment. We lower bound the minimum payment from Algorithm 4 and upper bound the privacy cost of any player with cost coefficient below threshold $\tau_{\alpha,\beta}$. If a is larger than the difference between these two terms, then any player with cost coefficient below threshold will receive non-negative utility.

Finally, we analyze the total cost to the analyst for running the mechanism.

Theorem 20 (Budget)**.** *The total budget required by the analyst to run Algorithm 4 when players utilize threshold equilibrium strategy $\sigma_{\tau_{\alpha,\beta}}$ is*

$$\mathcal{B} \leq n \left[a + \left(\frac{\alpha n}{\gamma}(4B + 2M) + \frac{\gamma B}{\gamma + (1 - \xi)\frac{1}{d+2}n} + B \right)(b + 2bB) \right].$$

Proof idea The analyst's budget is the sum of all payments made to players in the mechanism. We upper bound the maximum payment to any player, and the total budget required is at most n times this maximum payment.

Formal Statement of Main Result

In this section, we present our main result, Corollary 2, which instantiates Theorems 16, 17, 18, 19, and 20 with a setting of all parameters to get the bounds promised in Theorem 15. Before stating our main result, we first require the following lemma which asymptotically bounds $\tau_{\alpha,\beta}$ for an arbitrary bounded distribution. We use this to control the asymptotic behavior of $\tau_{\alpha,\beta}$ under Assumption 7.

Lemma 11. *For a cost distribution C with conditional marginal CDF lower bounded by some function F:*

$$\min_{x_i, y_i} \left(\Pr_{c_j \sim C | x_i, y_i}[c_j \leq \tau] \right) \geq F(\tau),$$

then

$$\tau_{\alpha,\beta} \leq \max\{F^{-1}(1 - \alpha\beta), F^{-1}(1 - \alpha)\}.$$

We note that under Assumption 7, Lemma 11 implies that

$$\tau_{\alpha,\beta} \leq \max\{(\alpha\beta)^{-1/p}, (\alpha)^{-1/p}\}.$$

Using this fact, we can state a formal version of our main result.

Corollary 2 (Main result (Formal))**.** *Choose $\delta \in (0, \frac{p}{2+2p})$. Then under Assumptions 4, 6, and 7, setting $\gamma = n^{1-\frac{\delta}{2}}$, $\epsilon = n^{-1+\delta}$, $a = (6B + 2M)(1 + B)^2 n^{-\frac{3}{2}} + n^{-\frac{3}{2}+\delta}$, and $b = n^{-\frac{3}{2}}$ in Algorithm 4, and taking $\alpha = n^{-\delta}$, $\beta = n^{-\frac{p}{2}+\delta(1+p)}$, $\xi = 1/2$, and $t = \sqrt{\frac{n}{4C(d+2)\log d}}$, ensures that with probability $1 - d^{\Theta(-n)} - n^{-\frac{p}{2}+\delta(1+p)}$:*

1. *the output of Algorithm 4 is $O\left(n^{-1+\delta}\right)$-jointly differentially private,*
2. *it is an $O\left(n^{-\frac{3}{2}+\delta}\right)$-approximate Bayes Nash equilibrium for a $1 - O\left(n^{-\delta}\right)$ fraction of players to truthfully report their data,*
3. *the computed estimate $\hat{\theta}^P$ is $O\left(n^{-\delta}\right)$-accurate,*
4. *it is individually rational for a $1 - O\left(n^{-\delta}\right)$ fraction of players to participate in the mechanism, and*
5. *the required budget from the analyst is $O\left(n^{-\frac{1}{2}+\delta}\right)$.*

This follows from instantiating Theorems 16, 17, 18, 19, and 20 with the specified parameters. Note that the choice of δ controls the trade-off between approximation factors for the desired properties.

Remark Note that different settings of parameters can be used to yield a different trade-off between approximation factors in the above result. For example, if the analyst is willing to supply a higher budget (say constant or increasing with n), he could improve on the accuracy guarantee.

6.6 Appendix: Omitted Proofs

Proof of Lemma 8 (Strong Convexity)

Lemma 8. *$\mathcal{L}(\theta; X, y)$ is 2γ-strongly convex in θ.*

Proof. We first compute the Hessian of $\mathcal{L}(\theta; X, y)$. For notational ease, we will suppress the dependence of $\mathcal{L}$ on X and y, and denote the loss function as $\mathcal{L}(\theta)$. We will use x_{ij} to denote the j-th coordinate of x_i, and θ_j to denote the j-th coordinate of θ.

$$\begin{aligned}
\frac{\partial \mathcal{L}(\theta)}{\partial \theta_j} &= \sum_{i=1}^{n} \left[-2y_i x_{ij} + 2(\theta^\top x_i)x_{ij}\right] + 2\gamma\theta_j \\
\frac{\partial \mathcal{L}(\theta)}{\partial \theta_j \partial \theta_k} &= \sum_{i=1}^{n} \left[2(x_{ik})x_{ij}\right] \text{ for } j \neq k \\
\frac{\partial \mathcal{L}(\theta)}{\partial \theta_j^2} &= \sum_{i=1}^{n} \left[2(x_{ij})^2\right] + 2\gamma
\end{aligned}$$

The Hessian of $\mathcal{L}$ is

$$H(\mathcal{L}(\theta)) = \sum_{i=1}^{n} x_i x_i^\top + 2\gamma I,$$

where I is the identity matrix. Thus,

$$H(\mathcal{L}(\theta)) - 2\gamma I = \sum_{i=1}^{n} x_i x_i^\top,$$

which is positive semi-definite. To see this, let v be an arbitrary vector in $\mathbb{R}^d$. Then for each i, $v(x_i x_i^\top)v^\top = (vx_i)^2 \geq 0$. The sum of PSD matrices is also PSD, so $\mathcal{L}(\theta)$ is 2γ-strongly convex. □

Proof of Theorem 14 (Concentration of Spectral Norms)

Theorem 14. *Let $\xi \in (0,1)$, and $t \geq 1$. Let $\|\cdot\|$ denote the spectral norm. If $\{x_i\}_{i\in[n]}$ are i.i.d. and sampled uniformly from the unit ball, then with probability at least $1 - d^{-t^2}$, when $n \geq C(\frac{t}{\xi})^2(d+2)\log d$, for some absolute constant C, then,*

$$\left\|X^\top X\right\| \leq (1+\xi)\frac{1}{d+2}n, \textit{ and } \left\|(X^\top X)^{-1}\right\| \leq \frac{1}{(1-\xi)\frac{1}{d+2}n}, \textit{ and}$$

$$\left\|\gamma I + X^\top X\right\| \leq \gamma + (1+\xi)\frac{1}{d+2}n, \textit{ and } \left\|(\gamma I + X^\top X)^{-1}\right\| \leq \frac{1}{\gamma + (1-\xi)\frac{1}{d+2}n}.$$

Proof of Theorem 14. We will first require Lemma 12, which characterizes the covariance matrix of the distribution on X.

Lemma 12. *The covariance matrix of x is $\Sigma = \frac{1}{d+2}I$.*

Proof of Lemma 12. Let $z_1, \ldots, z_d \sim N(0,1)$, and let $u \sim U[0,1]$, all drawn independently. Define, $r = \sqrt{z_1^2 + \cdots + z_d^2}$ and $Z = (u^{1/d}\frac{z_1}{r}, \ldots, u^{1/d}\frac{z_d}{r})$. Then Z

describes a uniform distribution over the d-dimensional unit ball [66]. Recall that this is the same distribution from which the x_i are drawn. By the symmetry of the uniform distribution, $\mathbb{E}[Z] = \vec{0}$, and $Cov(Z)$ must be some scalar times the Identity matrix. Then to compute the covariance matrix of Z, it will suffice to compute the variance of some coordinate Z_i of Z. Since each coordinate of Z has mean 0, then $Var(Z_i) = \mathbb{E}[Z_i^2] + \mathbb{E}[Z_i]^2 = \mathbb{E}[Z_i^2]$.

$$\begin{aligned}
\sum_{i=1}^{d} \mathbb{E}[Z_i^2] &= \mathbb{E}\left[\sum_{i=1}^{d} Z_i^2\right] \\
&= \mathbb{E}\left[\sum_{i=1}^{d} \left(u^{1/d}\frac{z_i}{r}\right)^2\right] \\
&= \mathbb{E}[u^{2/d}]\mathbb{E}\left[(\frac{1}{r})^2 \sum_{i=1}^{d} z_i^2\right] \\
&= \mathbb{E}[u^{2/d}] \\
&= \frac{d}{d+2}
\end{aligned}$$

By symmetry of coordinates, $\mathbb{E}[Z_i^2] = \mathbb{E}[Z_j^2]$ for all i, j. Then $\mathbb{E}[Z_i^2] = \frac{1}{d+2}$, and the covariance matrix of Z (and of the x_i since both variables have the same distribution) is $\Sigma = \frac{1}{d+2} I$. □

From Corollary 5.52 in [98] and the calculation of covariance in Lemma 12, for any $\xi \in (0, 1)$ and $t \geq 1$, with probability at least $1 - d^{-t^2}$,

$$\left\| \frac{1}{n} X^\top X - \frac{1}{d+2} I \right\| \leq \xi \frac{1}{d+2}, \tag{6.6}$$

when $n \geq C(\frac{t}{\xi})^2 (d+2) \log d$, for some absolute constant C. We assume for the remainder of the proof that inequality (6.6) holds, which is the case except with probability at most d^{-t^2}, as long as n is sufficiently large. Then

$$\left\| X^\top X - \frac{1}{d+2} nI \right\| \leq \xi \frac{1}{d+2} n.$$

Let $\lambda_{\max}(A)$ and $\lambda_{\min}(A)$ denote respectively the maximum and minimum eigenvalues of a matrix A. By definition, $\lambda_{\max}(A) = \|A\|$.

Assume towards a contradiction that $\lambda_{\max}(X^\top X) = (1+\xi)\frac{1}{d+2}n + \delta$ for $\delta > 0$.

$$\begin{aligned} \xi \frac{1}{d+2} n &\geq \left\| X^\top X - \frac{1}{d+2} nI \right\| \\ &= \left\| X^\top X \right\| - \frac{1}{d+2} n \\ &= \lambda_{\max}(X^\top X) - \frac{1}{d+2} n \\ &= (1+\xi)\frac{1}{d+2} n + \delta - \frac{1}{d+2} n \\ &= \xi \frac{1}{d+2} n + \delta \end{aligned}$$

This implies $\delta \leq 0$, which is a contradiction. Thus $\lambda_{\max}(X^\top X) = \|X^\top X\| \leq (1+\xi)\frac{1}{d+2}n$.

Similarly, assume that $\lambda_{\min}(X^\top X) = (1-\xi)\frac{1}{d+2}n - \delta$ for some $\delta > 0$. Since all eigenvalues are positive, it must be the case that $\lambda_{\min}(X^\top X) \geq 0$.

$$\begin{aligned} 0 &\geq \lambda_{\min}(X^\top X - \frac{1}{d+2} nI) \\ &= \lambda_{\min}(X^\top X) - \frac{1}{d+2} n \\ &= (1-\xi)\frac{1}{d+2} n - \delta - \frac{1}{d+2} n \\ &= -\xi \frac{1}{d+2} n - \delta \end{aligned}$$

This is also a contradiction, so $\lambda_{\min}(X^\top X) \geq (1-\xi)\frac{1}{d+2}n$. For any matrix A, $\lambda_{\max}(A^{-1}) = \frac{1}{\lambda_{\min}(A)}$. Thus,

$$\begin{aligned} \lambda_{\min}(X^\top X) &= \frac{1}{\lambda_{\max}\left((X^\top X)^{-1}\right)} \\ &= \frac{1}{\|(X^\top X)^{-1}\|} \\ &\geq (1-\xi)\frac{1}{d+2} n \\ \implies \quad \|(X^\top X)^{-1}\| &\leq (1-\xi)\frac{1}{d+2} n. \end{aligned}$$

Using the fact that λ is an eigenvalue of a matrix A if and only if $(\lambda + c)$ is an eigenvalue of $(A + cI)$, we have the following inequalities to complete the proof:

$$\left\| \gamma I + X^\top X \right\| = \lambda_{\max}(\gamma I + X^\top X) \leq \gamma + (1+\xi)\frac{1}{d+2} n$$

$$\left\| (\gamma I + X^\top X)^{-1} \right\| = \frac{1}{\lambda_{\min}(\gamma I + X^\top X)} \leq \frac{1}{\gamma + (1-\xi)\frac{1}{d+2} n}.$$

□

Proof of Theorem 16 (Privacy)

We will now prove that the estimator $\hat{\theta}^P$ and the vector of payments π of the mechanism in Algorithm 4 is 2ϵ-jointly differentially private. First, we need the following lemma to bound the sensitivity of $\hat{\theta}^P$.

Recall from Section 2.3 that the *sensitivity* (Definition 4) of a function is the maximum change in the output when a single player changes her data. Since our mechanism produces a vector-valued output, we measure this change from with respect to the ℓ_2 norm.

The following lemma follows from Chaudhuri, Monteleoni, and Sarwate [14]; a proof is provided for completeness.

Lemma 13. *The sensitivity of $\hat{\theta}^R$ is $\frac{1}{\gamma}(4B + 2M)$.*

Proof. Let (X, y) and (X', y') be two arbitrary neighboring databases that differ only in the i-th entry. Let $\hat{\theta}^R$ and $(\hat{\theta}^R)'$ respectively denote the ridge regression estimators computed on (X, y) and (X', y'). Define $g(\theta)$ to be the change in loss when θ is used as an estimator for (X', y') and (X, y). Then,

$$\begin{aligned} g(\theta) &= \mathcal{L}(\theta; X', y') - \mathcal{L}(\theta; X, y) \\ &= \left(\theta^\top x_i - y_i\right)^2 - \left(\theta^\top x_i' - y_i'\right)^2. \end{aligned}$$

Lemma 7 of Chaudhuri, Monteleoni, and Sarwate [14] says that if $\mathcal{L}(\theta; X, y)$ and $\mathcal{L}(\theta; X', y')$ are both Γ-strongly convex, then $\left\|\hat{\theta}^R - (\hat{\theta}^R)'\right\|_2$ is bounded above by $\frac{1}{\Gamma} \cdot \max_\theta \|\nabla g(\theta)\|_2$. By Lemma 8, both $\mathcal{L}(\theta; X, y)$ and $\mathcal{L}(\theta; X', y')$ are 2γ-strongly convex, so $\left\|\hat{\theta}^R - (\hat{\theta}^R)'\right\|_2 \leq \frac{1}{2\gamma} \cdot \max_\theta \|\nabla g(\theta)\|_2$. We now bound $\|\nabla g(\theta)\|_2$ for an arbitrary θ:

$$\begin{aligned} \|\nabla g(\theta)\|_2 &= 2\left\|(\theta^\top x_i - y_i)x_i - (\theta^\top x_i' - y_i')x_i'\right\|_2 \\ &\leq 4\left|\theta^\top x_i - y_i\right| \|x_i\|_2 \\ &\leq 4\left(\left|\theta^\top x_i\right| + |y_i|\right) \\ &\leq 4(2B + M). \end{aligned}$$

Since this bound holds for all θ, it must be the case that $\max_\theta \|\nabla g(\theta)\|_2 \leq 4(2B + M)$ as well. Then by Lemma 7 of Chaudhuri, Monteleoni, and Sarwate [14],

$$\left\|\hat{\theta}^R - (\hat{\theta}^R)'\right\|_2 \leq \frac{4}{2\gamma}(2B + M) = \frac{1}{\gamma}(4B + 2M).$$

Since (X, y) and (X', y') were two arbitrary neighboring databases, this bounds the sensitivity of the computation. Thus changing the input of one player can change the ridge regression estimator (with respect to the ℓ_2 norm) by at most $\frac{1}{\gamma}(4B + 2M)$. □

We now prove that the output of Algorithm 4 satisfies 2ϵ-joint differential privacy.

Theorem 16 (Privacy)**.** *The mechanism in Algorithm 4 is 2ϵ-jointly differentially private.*

Proof. We begin by showing that the estimator $\hat{\theta}^P$ output by Algorithm 4 is ϵ-differentially private.

Let h denote the PDF of $\hat{\theta}^P$ output by Algorithm 4, and ν denote the PDF of the noise vector v. Let (X, y) and (X', y') be any two databases that differ only in the i-th entry, and let $\hat{\theta}^R$ and $(\hat{\theta}^R)'$ respectively denote the ridge regression estimators computed on these two databases.

The output estimator $\hat{\theta}^P$ is the sum of the ridge regression estimator $\hat{\theta}^R$, and the noise vector v; the only randomness in the choice of $\hat{\theta}^P$ is the noise vector, because $\hat{\theta}^R$ is computed deterministically on the data. Thus the probability that Algorithm 4 outputs a particular $\hat{\theta}^P$ is equal to the probability that the noise vector is exactly the difference between $\hat{\theta}^P$ and $\hat{\theta}^R$. Fixing an arbitrary $\hat{\theta}^P$, let $\hat{v} = \hat{\theta}^P - \hat{\theta}^R$ and $\hat{v}' = \hat{\theta}^P - (\hat{\theta}^R)'$. Then,

$$\frac{h(\hat{\theta}^P|(X, y))}{h(\hat{\theta}^P|(X', y'))} = \frac{\nu(\hat{v})}{\nu(\hat{v}')} = \exp\left(\frac{\gamma\epsilon}{8B + 4M}\left(\|\hat{v}'\|_2 - \|\hat{v}\|_2\right)\right). \tag{6.7}$$

By definition, $\hat{\theta}^P = \hat{\theta}^R + \hat{v} = (\hat{\theta}^R)' + \hat{v}'$. Rearranging terms gives $\hat{\theta}^R - (\hat{\theta}^R)' = \hat{v}' - \hat{v}$. By Lemma 13 and the triangle inequality,

$$\|\hat{v}'\|_2 - \|\hat{v}\|_2 \leq \|\hat{v}' - \hat{v}\|_2 = \left\|\hat{\theta}^R - (\hat{\theta}^R)'\right\|_2 \leq \frac{1}{\gamma}(4B + 2M).$$

Plugging this into Equation (6.7) gives the desired inequality,

$$\frac{h(\hat{\theta}^P|(X, y))}{h(\hat{\theta}^P|(X', y'))} \leq \exp\left(\frac{\gamma\epsilon}{4B + 2M}\frac{1}{\gamma}(4B + 2M)\right) = \exp(\epsilon).$$

Next, we show that the output $(\hat{\theta}^P, \hat{\theta}_0^P, \hat{\theta}_1^P, \{\pi_i\}_{i\in[n]})$ of the mechanism satisfies joint differential privacy using the Billboard Lemma. The estimators $\hat{\theta}_0^P$ and $\hat{\theta}_1^P$ are computed in the same way as $\hat{\theta}^P$, so $\hat{\theta}_0^P$ and $\hat{\theta}_1^P$ each satisfy ϵ-differential privacy. Since

$\hat{\theta}^P_0$ and $\hat{\theta}^P_1$ are computed on disjoint subsets of the data, then by Theorem 4 of McSherry [72], together they satisfy ϵ-differential privacy. The estimator a player should use to compute her payments depends only on the partition of players, which is independent of the data because it is chosen uniformly at random. Then by Basic Composition (Theorem 1), the estimators $(\hat{\theta}^P, \hat{\theta}^P_0, \hat{\theta}^P_1)$ together satisfy 2ϵ-differential privacy.

Each player's payment π_i is a function of only her private information — her report $(x_i, \hat{y}_i)$ and her group in the partition of players — and the 2ϵ-differentially private collection of estimators $(\hat{\theta}^P, \hat{\theta}^P_0, \hat{\theta}^P_1)$. By the Billboard Lemma (Lemma 1), the output $(\hat{\theta}^P, \hat{\theta}^P_0, \hat{\theta}^P_1, \{\pi_i\}_{i \in [n]})$ of Algorithm 4 satisfies 2ϵ-joint differential privacy. □

Proof of Theorem 17 (Truthfulness)

In order to show that $\sigma_{\tau_{\alpha,\beta}}$ is an approximate Bayes Nash equilibrium, we require the following three lemmas. Lemma 14 bounds the expected number of players who will misreport under the strategy profile $\sigma_{\tau_{\alpha,\beta}}$. Lemma 15 bounds the norm of the expected difference of two estimators output by Algorithm 4 run on different datasets, as a function of the number of players whose data differs between the two datasets. Lemma 16 bounds the first two moments of the noise vector that is added to preserve privacy.

Lemma 14. *Under symmetric strategy profile $\sigma_{\tau_{\alpha,\beta}}$, each player expects that at most an α-fraction of other players will misreport, given Assumption 6.*

Proof. Let S_{-i} denote the set of players other than i who truthfully report under strategy $\sigma_{\tau_{\alpha,\beta}}$. From the perspective of player i, the cost coefficients of all other players are drawn independently from the posterior marginal distribution $C|_{x_i,y_i}$. By the definition of $\tau_{\alpha,\beta}$, player i believes that each other player truthfully reports independently with probability at least $1-\alpha$. Thus $\mathbb{E}[|S_{-i}| \; |x_i, y_i] \geq (1-\alpha)(n-1)$. □

Lemma 15. *Let $\hat{\theta}^R$ and $(\hat{\theta}^R)'$ be the ridge regression estimators computed on two fixed databases that differ on the input of at most k players. Then*

$$\left\|\hat{\theta}^R - (\hat{\theta}^R)'\right\|_2 \leq \frac{k}{\gamma}(4B + 2M).$$

Proof. Since the two databases differ on the reports of at most k players, we can define a sequence of databases $D_0, \ldots, D_k$, that each differ from the previous database in the input of at most one player, and D_0 is the input that generated $\hat{\theta}^R$, and D_k is the input that generated $(\hat{\theta}^R)'$. Consider running Algorithm 4 on each database D_j in the sequence. For each D_j, let $\hat{\theta}^R_j$ be the ridge regression estimator computed on D_j. Note that $\hat{\theta}^R_0 = \hat{\theta}^R$ and $\hat{\theta}^R_k = (\hat{\theta}^R)'$.

$$\begin{aligned}
\left\|\hat{\theta}^R - (\hat{\theta}^R)'\right\|_2 &= \left\|\hat{\theta}^R_0 - \hat{\theta}^R_k\right\|_2 \\
&= \left\|\hat{\theta}^R_0 - \hat{\theta}^R_1 + \hat{\theta}^R_1 - \ldots - \hat{\theta}^R_{k-1} + \hat{\theta}^R_{k-1} - \hat{\theta}^R_k\right\|_2 \\
&\leq \left\|\hat{\theta}^R_0 - \hat{\theta}^R_1\right\|_2 + \left\|\hat{\theta}^R_1 - \hat{\theta}^R_2\right\|_2 + \ldots + \left\|\hat{\theta}^R_{k-1} - \hat{\theta}^R_k\right\|_2 \\
&\leq k \cdot \max_j \left\|\hat{\theta}^R_j - \hat{\theta}^R_{j+1}\right\|_2
\end{aligned}$$

For each j, $\hat{\theta}^R_j$ and $\hat{\theta}^R_{j+1}$ are the ridge regression estimators computed on databases that differ in the data of at most a single player. That means either the databases are the same, so $\hat{\theta}^R_j = \hat{\theta}^R_{j+1}$ and their normed difference is 0, or they differ in the report of exactly one player. In the latter case, Lemma 13 bounds $\|\hat{\theta}^R_j - \hat{\theta}^R_{j+1}\|_2$ above by $\frac{1}{\gamma}(4B + 2M)$ for each j, including the j which maximizes the normed difference.

Combining this fact with the above inequalities gives,

$$\left\|\hat{\theta}^R - (\hat{\theta}^R)'\right\|_2 \leq \frac{k}{\gamma}(4B + 2M).$$

□

Lemma 16. *The noise vector v sampled according to P_L in Algorithm 4 satisfies:* $\mathbb{E}[v] = \vec{0}$ *and* $\mathbb{E}[\|v\|_2^2] = 2\left(\frac{4B+2M}{\gamma\epsilon}\right)^2$ *and* $\mathbb{E}[\|v\|_2] = \frac{4B+2M}{\gamma\epsilon}$.

Proof. For every $\bar{v} \in \mathbb{R}^d$, there exists $-\bar{v} \in \mathbb{R}^d$ that is drawn with the same probability, because $\|\bar{v}\|_2 = \|-\bar{v}\|_2$. Thus,

$$\mathbb{E}[v] = \int_{\bar{v}} \bar{v} \Pr(v = \bar{v}) d\bar{v} = \frac{1}{2}\int_{\bar{v}} (\bar{v} + -\bar{v}) \Pr(v = \bar{v}) d\bar{v} = \vec{0}.$$

The distribution of v is a high dimensional Laplacian with parameter $\frac{4B+2M}{\gamma\epsilon}$ and mean zero. It follows immediately that $\mathbb{E}[\|v\|_2^2] = 2\left(\frac{4B+2M}{\gamma\epsilon}\right)^2$ and $\mathbb{E}[\|v\|_2] = \frac{4B+2M}{\gamma\epsilon}$. □

We now prove that symmetric threshold strategy $\sigma_{\tau_{\alpha,\beta}}$ is an approximate Bayes Nash equilibrium of the Private Regression Mechanism in Algorithm 4.

Theorem 17 (Truthfulness)**.** *Fix a participation goal $1-\alpha$, a privacy parameter ϵ, a desired confidence parameter β, $\xi \in (0,1)$, and $t \geq 1$. Then under Assumptions 4 and 6, with probability $1 - d^{t^2}$ and when $n \geq C(\frac{t}{\xi})^2(d+2)\log d$, the symmetric threshold strategy $\sigma_{\tau_{\alpha,\beta}}$ is an η-approximate Bayes-Nash equilibrium in Algorithm 4 for*

$$\eta = b\left(\frac{\alpha n}{\gamma}(4B+2M) + \frac{\gamma B}{\gamma + (1-\xi)\frac{1}{d+2}n}\right)^2 + \tau_{\alpha,\beta}\epsilon^2.$$

Proof. Suppose all players other than i are following strategy $\sigma_{\tau_{\alpha,\beta}}$. Let player i be in group $1-j$, so she is paid according to the estimator computed on the data of group j. Let $\hat{\theta}^P_j$ be the estimator output by Algorithm 4 on the reported data of group j under this strategy, and let $(\hat{\theta}^R_j)'$ be the ridge regression estimator computed within Algorithm 4 when all players in group j follow strategy $\sigma_{\tau_{\alpha,\beta}}$. Let $\hat{\theta}^R_j$ be the ridge regression estimator that would have been computed within Algorithm 4 if all players in group j had reported truthfully. For ease of notation, we will suppress the subscripts on the estimators for the remainder of the proof.

We will show that $\sigma_{\tau_{\alpha,\beta}}$ is an approximate Bayes Nash equilibrium by bounding player i's incentive to deviate. We assume that $c_i \leq \tau_{\alpha,\beta}$ (otherwise there is nothing to show because player i would be allowed to submit an arbitrary report under $\sigma_{\tau_{\alpha,\beta}}$). We first compute the maximum amount that player i can increase her payment by misreporting to Algorithm 4. Consider the change in expected payment to player i from a fixed (deterministic) misreport, $\hat{y}_i \neq y_i$.

$$\begin{aligned}&\mathbb{E}[B_{a,b}((\hat{\theta}^P)^\top x_i, \mathbb{E}[\theta|x_i,\hat{y}_i]^\top x_i)|x_i,y_i] - \mathbb{E}[B_{a,b}((\hat{\theta}^P)^\top x_i, \mathbb{E}[\theta|x_i,y_i]^\top x_i)|x_i,y_i]\\&\quad = B_{a,b}(\mathbb{E}[\hat{\theta}^P|x_i,y_i]^\top x_i, \mathbb{E}[\theta|x_i,\hat{y}_i]^\top x_i) - B_{a,b}(\mathbb{E}[\hat{\theta}^P|x_i,y_i]^\top x_i, \mathbb{E}[\theta|x_i,y_i]^\top x_i)\end{aligned}$$

The rule $B_{a,b}$ is a proper scoring rule, so it is uniquely maximized when its two arguments are equal. Thus any misreport of player i cannot yield payment greater than $B_{a,b}(\mathbb{E}[\hat{\theta}^P|x_i,y_i]^\top x_i, \mathbb{E}[\hat{\theta}^P|x_i,y_i]^\top x_i)$, so the expression of interest is bounded above by the following:

$$
\begin{aligned}
&B_{a,b}(\mathbb{E}[\hat{\theta}^P|x_i,y_i]^\top x_i, \mathbb{E}[\hat{\theta}^P|x_i,y_i]^\top x_i) - B_{a,b}(\mathbb{E}[\hat{\theta}^P|x_i,y_i]^\top x_i, \mathbb{E}[\theta|x_i,y_i]^\top x_i) \\
&= a - b\left(\mathbb{E}[\hat{\theta}^P|x_i,y_i]^\top x_i - 2(\mathbb{E}[\hat{\theta}^P|x_i,y_i]^\top x_i)^2 + (\mathbb{E}[\hat{\theta}^P|x_i,y_i]^\top x_i)^2\right) \\
&\quad - a + b\left(\mathbb{E}[\hat{\theta}^P|x_i,y_i]^\top x_i - 2(\mathbb{E}[\hat{\theta}^P|x_i,y_i]^\top x_i)(\mathbb{E}[\theta|x_i,y_i]^\top x_i) + (\mathbb{E}[\theta|x_i,y_i]^\top x_i)^2\right) \\
&= b\left((\mathbb{E}[\hat{\theta}^P|x_i,y_i]^\top x_i)^2 - 2(\mathbb{E}[\hat{\theta}^P|x_i,y_i]^\top x_i)(\mathbb{E}[\theta|x_i,y_i]^\top x_i) + (\mathbb{E}[\theta|x_i,y_i]^\top x_i)^2\right) \\
&= b\left(\mathbb{E}[\hat{\theta}^P|x_i,y_i]^\top x_i - \mathbb{E}[\theta|x_i,y_i]^\top x_i\right)^2 \\
&= b\left(\mathbb{E}[\hat{\theta}^P - \theta|x_i,y_i]^\top x_i\right)^2 \\
&\leq b(||\mathbb{E}[\hat{\theta}^P - \theta|x_i,y_i]||_2^2 ||x_i||_2^2) \\
&\leq b||\mathbb{E}[\hat{\theta}^P - \theta|x_i,y_i]||_2^2.
\end{aligned}
$$

We continue by bounding the term $||\mathbb{E}[\hat{\theta}^P - \theta|x_i,y_i]||_2$.

$$
\begin{aligned}
||\mathbb{E}[\hat{\theta}^P - \theta|x_i,y_i]||_2 &= ||\mathbb{E}[\hat{\theta}^P - \hat{\theta}^R + \hat{\theta}^R - \theta|x_i,y_i]||_2 \\
&= ||\mathbb{E}[(\hat{\theta}^R)' + v - \hat{\theta}^R + \hat{\theta}^R - \theta|x_i,y_i]||_2 \\
&= ||\mathbb{E}[v|x_i,y_i] + \mathbb{E}[(\hat{\theta}^R)' - \hat{\theta}^R|x_i,y_i] + \mathbb{E}[\hat{\theta}^R - \theta|x_i,y_i]||_2 \\
&\leq ||\mathbb{E}[v|x_i,y_i]||_2 + ||\mathbb{E}[(\hat{\theta}^R)' - \hat{\theta}^R|x_i,y_i]||_2 + ||\mathbb{E}[\hat{\theta}^R - \theta|x_i,y_i]||_2
\end{aligned}
$$

We again bound each term separately. In the first term, the noise vector is drawn independently of the data, so $\mathbb{E}[v|x_i,y_i] = \mathbb{E}[v]$, which equals $\vec{0}$ by Lemma 16. Thus $||\mathbb{E}[v|x_i,y_i]||_2 = 0$.

Jensen's inequality bounds the second term above by $\mathbb{E}[||(\hat{\theta}^R)' - \hat{\theta}^R||_2|x_i,y_i]$. The random variables $(\hat{\theta}^R)'$ and $\hat{\theta}^R$ are the ridge regression estimators of two (random) databases that differ only on the data of players who misreported under threshold strategy $\sigma_{\tau_{\alpha,\beta}}$. By Lemma 14, player i believes that at most αn players will misreport their $\hat{y}_j$,[3] so for all pairs of databases over which the expectation is taken, $(\hat{\theta}^R)'$ and $\hat{\theta}^R$ differ in the input of at most αn players. By Lemma 15, their normed difference is bounded above by $\frac{\alpha n}{\gamma}(4B + 2M)$. Since this bound applied to every term over which the expectation is taken, it also bounds the expectation.

For the third term, $\mathbb{E}[\hat{\theta}^R - \theta|x_i,y_i] = \texttt{bias}(\hat{\theta}^R|x_i,y_i)$. Recall that $\hat{\theta}^R$ is actually $\hat{\theta}^R_j$, which is computed independently of player i's data, but is still correlated with (x_i,y_i) through the common parameter θ. However, conditioned on the true θ,

[3] Lemma 14 promises that at most $\alpha(n-1)$ players will misreport. We use the weaker bound of αn for simplicity.

the bias of $\hat{\theta}^R$ is independent of player i's data. That is, $\texttt{bias}(\hat{\theta}^R|x_i, y_i, \theta) = \texttt{bias}(\hat{\theta}^R|\theta)$. We now expand the third term using nested expectations.

$$\begin{aligned}
\mathbb{E}_{X,z,\theta}\left[\hat{\theta}^R - \theta|x_i, y_i\right] &= \mathbb{E}_\theta\left[\mathbb{E}_{X,z}[\hat{\theta}^R - \theta|x_i, y_i, \theta]\right] \\
&= \mathbb{E}_\theta\left[\texttt{bias}(\hat{\theta}^R|x_i, y_i, \theta)\right] \\
&= \mathbb{E}_\theta\left[\texttt{bias}(\hat{\theta}^R|\theta)\right] \\
&= \texttt{bias}(\hat{\theta}^R) \\
&= -\gamma(\gamma I + X^\top X)^{-1}\theta
\end{aligned}$$

Then by Theorem 14, when $n \geq C(\frac{t}{\xi})^2(d+2)\log d$, the following holds with probability at least $1 - d^{-t^2}$.

$$\begin{aligned}
||\mathbb{E}[\hat{\theta}^R - \theta|x_i, y_i]||_2 &= ||-\gamma(\gamma I + X^\top X)^{-1}\theta||_2 \\
&\leq \gamma||(\gamma I + X^\top X)^{-1}||_2||\theta||_2 \\
&\leq \gamma\left(\frac{1}{\gamma + (1-\xi)\frac{1}{d+2}n}\right)B \\
&= \frac{\gamma B}{\gamma + (1-\xi)\frac{1}{d+2}n}
\end{aligned}$$

We will assume the above is true for the remainder of the proof, which will be the case except with probability at most d^{-t^2}. Thus with probability at least $1 - d^{-t^2}$, and when n is sufficiently large, the increase in payment from misreporting is bounded above by

$$b||\mathbb{E}[\hat{\theta}^P - \theta|x_i, y_i]||_2^2 \leq b\left(\frac{\alpha n}{\gamma}(4B + 2M) + \frac{\gamma B}{\gamma + (1-\xi)\frac{1}{d+2}n}\right)^2.$$

In addition to an increased payment, a player may also experience decreased privacy costs from misreporting. By Assumption 4, this decrease in privacy costs is bounded above by $c_i\epsilon^2$. We have assumed $c_i \leq \tau_{\alpha,\beta}$ (otherwise player i is allowed to misreport arbitrarily under $\sigma_{\tau_{\alpha,\beta}}$, and there is nothing to show). Then the decrease in privacy costs for player i is bounded above by $\tau_{\alpha,\beta}\epsilon^2$.

Therefore player i's total incentive to deviate is bounded above by η, and the symmetric threshold strategy $\sigma_{\tau_{\alpha,\beta}}$ forms an η-approximate Bayes Nash equilibrium for

$$\eta = b\left(\frac{\alpha n}{\gamma}(4B + 2M) + \frac{\gamma B}{\gamma + (1-\xi)\frac{1}{d+2}n}\right)^2 + \tau_{\alpha,\beta}\epsilon^2.$$

□

Proof of Theorem 18 (Accuracy)

In this section, we prove that the estimator $\hat{\theta}^P$ output by Algorithm 4 has high accuracy. We first require the following lemma, which uses the concentration inequalities of Theorem 14 to give high probability bounds on the distance from the ridge regression estimator to the true parameter θ.

Lemma 17. *Let $\hat{\theta}^R$ be the ridge regression estimator computed on a given database (X, y). Then with probability at least $1 - d^{-t^2}$, as long as $n \geq C(\frac{t}{\xi})^2(d+2)\log d$*

$$\mathbb{E}[\|\hat{\theta}^R - \theta\|_2^2] \leq \left(\frac{\gamma B}{\gamma + (1-\xi)\frac{1}{d+2}n}\right)^2 + \sigma^4\left(\frac{(1+\xi)\frac{1}{d+2}n}{(\gamma + (1-\xi)\frac{1}{d+2}n)^2}\right)^2$$

and

$$\mathbb{E}[\|\hat{\theta}^R - \theta\|_2] \leq \frac{\gamma B + Mn}{\gamma + (1-\xi)\frac{1}{d+2}n}.$$

Proof. Recall from Section 6.2 that

$$\mathbb{E}[\|\hat{\theta}^R - \theta\|_2^2] = \|\,\mathtt{bias}(\hat{\theta}^R)\|_2^2 + \mathtt{trace}(\mathtt{Cov}(\hat{\theta}^R)),$$

and

$$\begin{aligned}\mathbb{E}[\|\hat{\theta}^R - \theta\|_2] &= \mathbb{E}[\|\hat{\theta}^R - \mathbb{E}[\hat{\theta}^R] + \mathbb{E}[\hat{\theta}^R] - \theta\|_2] \\ &\leq \mathbb{E}[\|\hat{\theta}^R - \mathbb{E}[\hat{\theta}^R]\|_2] + \mathbb{E}[\|\mathbb{E}[\hat{\theta}^R] - \theta\|_2] \\ &= \mathbb{E}[\|\hat{\theta}^R - \mathbb{E}[\hat{\theta}^R]\|_2] + \mathbb{E}[\|\,\mathtt{bias}(\hat{\theta}^R)\|_2].\end{aligned}$$

We now expand the three remaining terms: $\|\,\mathtt{bias}(\hat{\theta}^R)\|_2$ and $\mathtt{trace}(\mathtt{Cov}(\hat{\theta}^R))$ and $\mathbb{E}[\|\hat{\theta}^R - \mathbb{E}[\hat{\theta}^R]\|_2]$. For the remainder of the proof, we will assume the concentration inequalities in Theorem 14 hold, which will be the case, except with probability at most d^{-t^2}, as long as $n \geq C(\frac{t}{\xi})^2(d+2)\log d$.

$$\begin{aligned}\|\,\mathtt{bias}(\hat{\theta}^R)\|_2 &= \|-\gamma(\gamma I + X^\top X)^{-1}\theta\|_2 \\ &\leq \gamma\|\theta\|_2\|(\gamma I + X^\top X)^{-1}\|_2 \\ &\leq \gamma B\|(\gamma I + X^\top X)^{-1}\|_2 \\ &\leq \frac{\gamma B}{\gamma + (1-\xi)\frac{1}{d+2}n}\end{aligned}$$

$$\begin{aligned}
\texttt{trace}(\texttt{Cov}(\hat{\theta}^R)) &= \|\,\texttt{Cov}(\hat{\theta}^R)\|_2^2 \\
&= \|\sigma^2(\gamma I + X^\top X)^{-1}X^\top X(\gamma I + X^\top X)^{-1}\|_2^2 \\
&\leq \sigma^4\|(\gamma I + X^\top X)^{-1}\|_2^2\|X^\top X\|_2^2\|(\gamma I + X^\top X)^{-1}\|_2^2 \\
&\leq \sigma^4\left(\frac{1}{\gamma + (1-\xi)\frac{1}{d+2}n}\right)^2\left((1+\xi)\frac{1}{d+2}n\right)^2\left(\frac{1}{\gamma + (1-\xi)\frac{1}{d+2}n}\right)^2 \\
&\leq \sigma^4\left(\frac{(1+\xi)\frac{1}{d+2}n}{\left(\gamma + (1-\xi)\frac{1}{d+2}n\right)^2}\right)^2
\end{aligned}$$

$$\begin{aligned}
\mathbb{E}[\|\hat{\theta}^R - \mathbb{E}[\hat{\theta}^R]\|_2] &= \mathbb{E}[\|\hat{\theta}^R - (\theta + \texttt{bias}(\hat{\theta}^R))\|_2] \\
&= \mathbb{E}[\|(\gamma I + X^\top X)^{-1}X^\top y - \theta + (\gamma I + X^\top X)^{-1}\gamma I\theta\|_2] \\
&= \mathbb{E}[\|(\gamma I + X^\top X)^{-1}X^\top(X\theta + z) - \theta + (\gamma I + X^\top X)^{-1}\gamma I\theta\|_2] \\
&= \mathbb{E}[\|(\gamma I + X^\top X)^{-1}(X^\top X + \gamma I)\theta - \theta + (\gamma I + X^\top X)^{-1}X^\top z\|_2] \\
&= \mathbb{E}[\|\theta - \theta + (\gamma I + X^\top X)^{-1}X^\top z\|_2] \\
&= \mathbb{E}[\|(\gamma I + X^\top X)^{-1}X^\top z\|_2] \\
&\leq \mathbb{E}[\|(\gamma I + X^\top X)^{-1}\|_2\|X^\top z\|_2] \\
&\leq \mathbb{E}[\|(\gamma I + X^\top X)^{-1}\|_2 Mn] \\
&\leq \frac{Mn}{\gamma + (1-\xi)\frac{1}{d+2}n}
\end{aligned}$$

Using these bounds, we see:

$$\mathbb{E}[\|\hat{\theta}^R - \theta\|_2^2] \leq \left(\frac{\gamma B}{\gamma + (1-\xi)\frac{1}{d+2}n}\right)^2 + \sigma^4\left(\frac{(1+\xi)\frac{1}{d+2}n}{(\gamma + (1-\xi)\frac{1}{d+2}n)^2}\right)^2,$$

and

$$\begin{aligned}
\mathbb{E}[\|\hat{\theta}^R - \theta\|_2] &\leq \frac{\gamma B}{\gamma + (1-\xi)\frac{1}{d+2}n} + \frac{Mn}{\gamma + (1-\xi)\frac{1}{d+2}n} \\
&= \frac{\gamma B + Mn}{\gamma + (1-\xi)\frac{1}{d+2}n}.
\end{aligned}$$

□

We now prove the accuracy guarantee for the estimator $\hat{\theta}^P$ output by Algorithm 4.

Theorem 18 (Accuracy)**.** *Fix a participation goal* $1-\alpha$*, a privacy parameter* ϵ*, a desired confidence parameter* β*,* $\xi \in (0,1)$*, and* $t \geq 1$*. Then under the symmetric threshold strategy* $\sigma_{\tau_{\alpha,\beta}}$*, Algorithm 4 will output an estimator* $\hat{\theta}^P$ *such that with probability at least* $1-\beta-d^{-t^2}$*, and when* $n \geq C(\frac{t}{\xi})^2(d+2)\log d$*,*

$$\mathbb{E}[\|\hat{\theta}^P-\theta\|_2^2] = O\left(\left(\frac{\alpha n}{\gamma}+\frac{1}{\gamma\epsilon}\right)^2+\left(\frac{\gamma}{n}\right)^2+\left(\frac{1}{n}\right)^2+\frac{\alpha n}{\gamma}+\frac{1}{\gamma\epsilon}\right).$$

Proof. Let the data held by players be (X,y), and let $\hat{y}$ be the reports of players under the threshold strategy $\sigma_{\tau_{\alpha,\beta}}$. As in Theorem 17, let $\hat{\theta}^P$ be the estimator output by Algorithm 4 on the reported data under this strategy, and let $(\hat{\theta}^R)'$ be the ridge regression estimator computed Algorithm 4 when all players follow strategy $\sigma_{\tau_{\alpha,\beta}}$. Let $\hat{\theta}^R$ be the ridge regression estimator that would have been computed within Algorithm 4 if all players had reported truthfully. Recall that v is the noise vector added in Algorithm 4.

$$\begin{aligned}
\mathbb{E}[\|\hat{\theta}^P-\theta\|_2^2] &= \mathbb{E}[\|\hat{\theta}^P-\hat{\theta}^R+\hat{\theta}^R-\theta\|_2^2] \\
&= \mathbb{E}\left[\|\hat{\theta}^P-\hat{\theta}^R\|_2^2+\|\hat{\theta}^R-\theta\|_2^2+2\left\langle\hat{\theta}^P-\hat{\theta}^R,\hat{\theta}^R-\theta\right\rangle\right] \\
&\leq \mathbb{E}[\|\hat{\theta}^P-\hat{\theta}^R\|_2^2]+\mathbb{E}[\|\hat{\theta}^R-\theta\|_2^2]+2\mathbb{E}[\|\hat{\theta}^P-\hat{\theta}^R\|_2\|\hat{\theta}^R-\theta\|_2]
\end{aligned}$$

We start by bounding the first term. Recall that the estimator $\hat{\theta}^P$ is equal to the ridge regression estimator on the *reported* data, plus the noise vector v added by Algorithm 4.

$$\begin{aligned}
\mathbb{E}[\|\hat{\theta}^P-\hat{\theta}^R\|_2^2] &= \mathbb{E}[\|(\hat{\theta}^R)'+v-\hat{\theta}^R\|_2^2] \\
&= \mathbb{E}[\|(\hat{\theta}^R)'-\hat{\theta}^R\|_2^2]+\mathbb{E}[\|v\|_2^2]+2\mathbb{E}[\langle(\hat{\theta}^R)'-\hat{\theta}^R,v\rangle] \\
&= \mathbb{E}[\|(\hat{\theta}^R)'-\hat{\theta}^R\|_2^2]+\mathbb{E}[\|v\|_2^2]+2\langle\mathbb{E}[(\hat{\theta}^R)'-\hat{\theta}^R],\mathbb{E}[v]\rangle \\
&= \mathbb{E}[\|(\hat{\theta}^R)'-\hat{\theta}^R\|_2^2]+2\left(\frac{4B+2M}{\gamma\epsilon}\right)^2 \quad \text{(by Lemma 16)}
\end{aligned}$$

The estimators $(\hat{\theta}^R)'$ and $\hat{\theta}^R$ are the ridge regression estimators of two (random) databases that differ only on the data of players who misreported under threshold strategy $\sigma_{\tau_{\alpha,\beta}}$. The definition of $\tau_{\alpha,\beta}$ ensures us that with probability $1-\beta$, at most αn players will misreport their $\hat{y}_j$. For the remainder of the proof, we will assume that at most αn players misreported to the mechanism, which will be the case except with probability β.

Thus for all pairs of databases over which the expectation is taken, $(\hat{\theta}^R)'$ and $\hat{\theta}^R$ differ in the input of at most αn players, and by Lemma 15, their normed difference is bounded above by $\left(\frac{\alpha n}{\gamma}(4B + 2M)\right)^2$. Since this bound applies to every term over which the expectation is taken, it also bounds the expectation.

Thus the first term satisfies the following bound:

$$\mathbb{E}[||\hat{\theta}^P - \theta||_2^2] \leq \left(\frac{\alpha n}{\gamma}(4B + 2M)\right)^2 + 2\left(\frac{4B + 2M}{\gamma \epsilon}\right)^2.$$

By Lemma 17, with probability at least $1 - d^{-t^2}$, when $n \geq C(\frac{t}{\xi})^2(d+2)\log d$, the second term is bounded above by

$$\mathbb{E}[||\hat{\theta}^R - \theta||_2^2] \leq \left(\frac{\gamma B}{\gamma + (1-\xi)\frac{1}{d+2}n}\right)^2 + \sigma^4\left(\frac{(1+\xi)\frac{1}{d+2}n}{(\gamma + (1-\xi)\frac{1}{d+2}n)^2}\right)^2.$$

We will also assume for the remainder of the proof that the above bound holds, which will be the case except with probability at most d^{-t^2}.

We now bound the third term.

$$\begin{aligned}
2\mathbb{E}[||\hat{\theta}^P - \hat{\theta}^R||_2||\hat{\theta}^R - \theta||_2] &= 2\mathbb{E}[||(\hat{\theta}^R)' + v - \hat{\theta}^R||_2||\hat{\theta}^R - \theta||_2] \\
&\leq 2\mathbb{E}[\left(||(\hat{\theta}^R)' - \hat{\theta}^R||_2 + ||v||_2\right)||\hat{\theta}^R - \theta||_2] \\
&= 2\mathbb{E}[||(\hat{\theta}^R)' - \hat{\theta}^R||_2||\hat{\theta}^R - \theta||_2] + 2\mathbb{E}[||v||_2||\hat{\theta}^R - \theta||_2] \\
&= 2\mathbb{E}[||(\hat{\theta}^R)' - \hat{\theta}^R||_2||\hat{\theta}^R - \theta||_2] + 2\mathbb{E}[||v||_2]\mathbb{E}[||\hat{\theta}^R - \theta||_2] \text{ (by independence)} \\
&= 2\mathbb{E}[||(\hat{\theta}^R)' - \hat{\theta}^R||_2||\hat{\theta}^R - \theta||_2] + 2\left(\frac{4B + 2M}{\gamma \epsilon}\right)\mathbb{E}[||\hat{\theta}^R - \theta||_2] \text{ (by Lemma 16)}
\end{aligned}$$

We have assumed at most αn players misreported (which will occur with probability at least $1 - \beta$), so for all pairs of databases over which the expectation in the first term is taken, Lemma 15 bounds $||(\hat{\theta}^R)' - \hat{\theta}^R||$ above by $\frac{\alpha n}{\gamma}(4B + 2M)$. Thus we continue bonding the third term:

$$
\begin{aligned}
2\mathbb{E}[\|(\hat{\theta}^R)' - \hat{\theta}^R\|_2\|\hat{\theta}^R - \theta\|_2] + 2\left(\frac{4B+2M}{\gamma\epsilon}\right)\mathbb{E}[\|\hat{\theta}^R - \theta\|_2]& \\
\leq 2\mathbb{E}\left[\left(\frac{\alpha n}{\gamma}(4B+2M)\right)\|\hat{\theta}^R - \theta\|_2\right] &+ 2\frac{4B+2M}{\gamma\epsilon}\mathbb{E}\left[\|\hat{\theta}^R - \theta\|_2\right] \\
= 2\left(\frac{\alpha n}{\gamma}(4B+2M)\right)\mathbb{E}\left[\|\hat{\theta}^R - \theta\|_2\right] &+ 2\frac{4B+2M}{\gamma\epsilon}\mathbb{E}\left[\|\hat{\theta}^R - \theta\|_2\right] \\
= 2\left(\frac{\alpha n}{\gamma}(4B+2M) + \frac{4B+2M}{\gamma\epsilon}\right)\mathbb{E}\left[\|\hat{\theta}^R - \theta\|_2\right]& \\
\leq 2\left(\frac{\alpha n}{\gamma}(4B+2M) + \frac{4B+2M}{\gamma\epsilon}\right)\frac{\gamma B + Mn}{\gamma + (1-\xi)\frac{1}{d+2}n}&
\end{aligned}
$$

The first inequality above comes from Lemma 15, and the second inequality is due to Lemma 17.

We can now plug these terms back in to get our final accuracy bound. Taking a union bound over the two failure probabilities, with probability at least $1 - \beta - d^{-t^2}$, when $n \geq C(\frac{t}{\xi})^2(d+2)\log d$:

$$
\begin{aligned}
\mathbb{E}\left[\left\|\hat{\theta}^P - \theta\right\|_2^2\right] \leq \left(\frac{\alpha n}{\gamma}(4B+2M)\right)^2 &+ 2\left(\frac{4B+2M}{\gamma\epsilon}\right)^2 \\
&+ \left(\frac{\gamma B}{\gamma + (1-\xi)\frac{1}{d+2}n}\right)^2 + \sigma^4\left(\frac{(1+\xi)\frac{1}{d+2}n}{(\gamma + (1-\xi)\frac{1}{d+2}n)^2}\right)^2 \\
&+ 2\left(\frac{\alpha n}{\gamma}(4B+2M) + \frac{4B+2M}{\gamma\epsilon}\right)\frac{\gamma B + Mn}{\gamma + (1-\xi)\frac{1}{d+2}n}
\end{aligned}
$$

□

Proof of Theorems 19 and 20 (Individual Rationality and Budget)

In this section we first characterize the conditions needed for individual rationality, and then compute the total budget required from the analyst to run the Private Regression Mechanism in Algorithm 4. Note that if we do not require individual rationality, it is easy to achieve a small budget: we can scale down payments as in the non-private mechanism from Section 6.3. However, once players have privacy concerns, they will no longer accept an arbitrarily small positive payment; each player must be paid enough to compensate for her privacy loss. In order to incentivize players to participate in the mechanism, the analyst will have to ensure that players receive non-negative utility from participation.

We first show that Algorithm 4 is individually rational for players with privacy costs below threshold. Note that because we allow cost parameters to be unbounded, it is not possible in general to ensure individual rationality for all players while maintaining a finite budget.

Theorem 19 (Individual Rationality)**.** *Under Assumption 4, the mechanism in Algorithm 4 is individually rational for all players with cost parameters $c_i \leq \tau_{\alpha,\beta}$ as long as,*

$$a \geq \left(\frac{\alpha n}{\gamma}(4B+2M) + \frac{\gamma B}{\gamma + (1-\xi)\frac{1}{d+2}n} + B\right)(b+2bB) + bB^2 + \tau_{\alpha,\beta}\epsilon^2,$$

regardless of the reports from players with cost coefficients above $\tau_{\alpha,\beta}$.

Proof. Let player i have privacy cost parameter $c_i \leq \tau_{\alpha,\beta}$, and consider player i's utility from participating in the mechanism. Let player i be in group $1-j$, so she is paid according to the estimator computed on the data of group j. Let $\hat{\theta}^P_j$ be the estimator output by Algorithm 4 on the reported data of group j under this strategy, and let $(\hat{\theta}^R_j)'$ be the ridge regression estimator computed within Algorithm 4 when all players in group j follow strategy $\sigma_{\tau_{\alpha,\beta}}$. Let $\hat{\theta}^R_j$ be the ridge regression estimator that would have been computed within Algorithm 4 if all players in group j had reported truthfully. For ease of notation, we will suppress the subscripts on the estimators for the remainder of the proof.

$$\begin{aligned}\mathbb{E}[u_i(x_i,y_i,\hat{y}_i)] &= \mathbb{E}[B_{a,b}((\hat{\theta}^P)^\top x_i, \mathbb{E}[\theta|x_i,\hat{y}_i]^\top x_i)|x_i,y_i] - \mathbb{E}[f_i(c_i,\epsilon)] \\ &\geq \mathbb{E}[B_{a,b}((\hat{\theta}^P)^\top x_i, \mathbb{E}[\theta|x_i,\hat{y}_i]^\top x_i)|x_i,y_i] - \tau_{\alpha,\beta}\epsilon^2 \text{ (by Assump. 4)} \\ &= B_{a,b}(\mathbb{E}[\hat{\theta}^P|x_i,y_i]^\top x_i, \mathbb{E}[\theta|x_i,\hat{y}_i]^\top x_i) - \tau_{\alpha,\beta}\epsilon^2\end{aligned}$$

We proceed by bounding the inputs to the payment rule, and thus lower-bounding the payment player i receives. The second input satisfies the following bound:

$$\mathbb{E}[\theta|x_i,\hat{y}_i]^\top x_i \leq ||\mathbb{E}[\theta|x_i,\hat{y}_i]||_2||x_i||_2 \leq B.$$

We can also bound the first input to the payment rule as follows:

$$\begin{aligned}\mathbb{E}[\hat{\theta}^P|x_i,y_i]^\top x_i &= \mathbb{E}[(\hat{\theta}^R)'|x_i,y_i]^\top x_i + \mathbb{E}[v|x_i,y_i]^\top x_i \\ &= \mathbb{E}[(\hat{\theta}^R)'|x_i,y_i]^\top x_i \\ &\leq ||\mathbb{E}[(\hat{\theta}^R)'|x_i,y_i]||_2||x_i||_2 \\ &\leq ||\mathbb{E}[(\hat{\theta}^R)' - \hat{\theta}^R|x_i,y_i]||_2 + ||\mathbb{E}[\hat{\theta}^R - \theta|x_i,y_i]||_2 + ||\mathbb{E}[\theta|x_i,y_i]||_2 \\ &\leq \frac{\alpha n}{\gamma}(4B+2M) + \frac{\gamma B}{\gamma + (1-\xi)\frac{1}{d+2}n} + B,\end{aligned}$$

where the last inequality is due to Lemma 15 and Theorem 14.

Recall that our Brier-based payment rule is $B_{a,b}(p,q) = a - b\left(p - 2pq + q^2\right)$, which is bounded below by $a - b|p| - 2b|p|\ |q| - b|q|^2 = a - |p|(b + 2b|q|) - b|q|^2$. Using the bounds we just computed on the inputs to player i's payment rule, her payment is at least

$$\pi_i \geq a - \left(\frac{\alpha n}{\gamma}(4B + 2M) + \frac{\gamma B}{\gamma + (1-\xi)\frac{1}{d+2}n} + B\right)(b + 2bB) - bB^2.$$

Thus her expected utility from participating in the mechanism is at least

$$\mathbb{E}[u_i(x_i, y_i, \hat{y}_i)] \geq a - \left(\frac{\alpha n}{\gamma}(4B + 2M) + \frac{\gamma B}{\gamma + (1-\xi)\frac{1}{d+2}n} + B\right)(b + 2bB) - bB^2 - \tau_{\alpha,\beta}\epsilon^2.$$

Player i will be ensured non-negative utility as long as

$$a \geq \left(\frac{\alpha n}{\gamma}(4B + 2M) + \frac{\gamma B}{\gamma + (1-\xi)\frac{1}{d+2}n} + B\right)(b + 2bB) + bB^2 + \tau_{\alpha,\beta}\epsilon^2.$$

□

The next theorem characterizes the total budget required by the analyst to run Algorithm 4.

Theorem 20 (Budget)**.** *The total budget required by the analyst to run Algorithm 4 when players utilize threshold equilibrium strategy $\sigma_{\tau_{\alpha,\beta}}$ is*

$$\mathcal{B} \leq n\left[a + \left(\frac{\alpha n}{\gamma}(4B + 2M) + \frac{\gamma B}{\gamma + (1-\xi)\frac{1}{d+2}n} + B\right)(b + 2bB)\right].$$

Proof. The total budget is the sum of payments to all players.

$$\begin{aligned}\mathcal{B} = \sum_{i=1}^{n} \mathbb{E}[\pi_i] &= \sum_{i=1}^{n} \mathbb{E}[B_{a,b}((\hat{\theta}^P)^\top x_i, \mathbb{E}[\theta | x_i, \hat{y}_i]^\top x_i) | x_i, y_i] \\ &= \sum_{i=1}^{n} B_{a,b}(\mathbb{E}[\hat{\theta}^P | x_i, y_i]^\top x_i, \mathbb{E}[\theta | x_i, \hat{y}_i]^\top x_i)\end{aligned}$$

Recall that our Brier-based payment rule is $B_{a,b}(p,q) = a - b\left(p - 2pq + q^2\right)$, which is bounded above by $a + b|p| + 2b|p|\ |q| = a + |p|(b + 2b|q|)$. Using the bounds computed in the proof of Theorem 19, each player i receives payment at most,

$$\pi_i \geq a + \left(\frac{\alpha n}{\gamma}(4B + 2M) + \frac{\gamma B}{\gamma + (1-\xi)\frac{1}{d+2}n} + B\right)(b + 2bB).$$

Thus the total budget is at most:

$$\mathcal{B} = \sum_{i=1}^{n} \mathbb{E}[\pi_i] \leq n\left(a + \left(\frac{\alpha n}{\gamma}(4B + 2M) + \frac{\gamma B}{\gamma + (1-\xi)\frac{1}{d+2}n} + B\right)(b + 2bB)\right).$$

□

Proof of Lemma 11 (Bound on threshold $\tau_{\alpha,\beta}$)

Lemma 11. *For a cost distribution C with conditional marginal CDF lower bounded by some function F:*

$$\min_{x_i,y_i}\left(\Pr_{c_j\sim C|x_i,y_i}[c_j \leq \tau]\right) \geq F(\tau),$$

then

$$\tau_{\alpha,\beta} \leq \max\{F^{-1}(1-\alpha\beta), F^{-1}(1-\alpha)\}.$$

Proof. We first bound $\tau^1_{\alpha,\beta}$.

$$\begin{aligned}
\tau^1_{\alpha,\beta} &= \inf_{\tau}\left(\Pr_{c\sim C}\left[|\{i : c_i \leq \tau\}| \geq (1-\alpha)n\right] \geq 1-\beta\right)\\
&= \inf_{\tau}\left(\Pr_{c\sim C}\left[|\{i : c_i \geq \tau\}| \leq \alpha n\right] \geq 1-\beta\right)\\
&= \inf_{\tau}\left(1 - \Pr_{c\sim C}\left[|\{i : c_i \geq \tau\}| \geq \alpha n\right] \geq 1-\beta\right)\\
&= \inf_{\tau}\left(\Pr_{c\sim C}\left[|\{i : c_i \geq \tau\}| \geq \alpha n\right] \leq \beta\right)
\end{aligned}$$

We continue by upper bounding the inner term of the expression.

$$\begin{aligned}
\Pr_{c\sim C}\left[|\{i : c_i \geq \tau\}| \geq \alpha n\right] &\leq \frac{\mathbb{E}[|\{i : c_i \geq \tau\}|}{\alpha n} \text{ (by Markov's inequality)}\\
&= \frac{n\ Pr[c_i \geq \tau]}{\alpha n} \text{ (by independence of costs)}\\
&= \frac{Pr[c_i \geq \tau]}{\alpha}
\end{aligned}$$

From this bound, if $\frac{Pr[c_i \geq \tau]}{\alpha} \leq \beta$, then also $\Pr_{c\sim C}\left[|\{i : c_i \geq \tau\}| \geq \alpha n\right] \leq \beta$. Thus,

$$\inf_{\tau}\left(\Pr_{c\sim C}\left[|\{i : c_i \geq \tau\}| \geq \alpha n\right] \leq \beta\right) \leq \inf_{\tau}\left(\frac{Pr[c_i \geq \tau]}{\alpha} \leq \beta\right),$$

since the infimum in the first expression is taken over a superset of the feasible region of the latter expression. Then,

$$\begin{aligned}
\tau^1_{\alpha,\beta} &\leq \inf_{\tau} \left(\frac{Pr[c_i \geq \tau]}{\alpha} \leq \beta \right) \\
&= \inf_{\tau} \left(Pr[c_i \geq \tau] \leq \alpha\beta \right) \\
&= \inf_{\tau} \left(1 - Pr[c_i \leq \tau] \leq \alpha\beta \right) \\
&= \inf_{\tau} \left(C(\tau) \geq 1 - \alpha\beta \right) \\
&\leq \inf_{\tau} \left(F(\tau) \geq 1 - \alpha\beta \right) \\
&\quad \text{(since the extremal conditional marginal bounds the unconditioned marginal)} \\
&= \inf_{\tau} \left(\tau \geq F^{-1}(1 - \alpha\beta) \right) \\
&= F^{-1}(1 - \alpha\beta)
\end{aligned}$$

Thus under our assumptions, $\tau^1_{\alpha,\beta} \leq F^{-1}(1 - \alpha\beta)$.

We now bound τ^2_{α}.

$$\begin{aligned}
\tau^2_{\alpha} &= \inf_{\tau} \left(\min_{x_i, y_i} \left(Pr_{c_j \sim C|x_i, y_i}[c_j \leq \tau] \right) \geq 1 - \alpha \right) \\
&\leq \inf_{\tau} \left(F(\tau) \geq 1 - \alpha \right) \\
&= \inf_{\tau} \left(\tau \geq F^{-1}(1 - \alpha) \right) \\
&= F^{-1}(1 - \alpha)
\end{aligned}$$

Finally,

$$\tau_{\alpha,\beta} = \max\{\tau^1_{\alpha,\beta}, \tau^2_{\alpha}\} \leq \max\{F^{-1}(1 - \alpha\beta), F^{-1}(1 - \alpha)\}.$$

□

Proof of Corollary 2 (Main result)

Corollary 1 (Main result (Formal)). *Choose* $\delta \in (0, \frac{p}{2+2p})$. *Then under Assumptions 4, 6, and 7, setting* $\alpha = n^{-\delta}$, $\beta = n^{-\frac{p}{2}+\delta(1+p)}$, $\epsilon = n^{-1+\delta}$, $\gamma = n^{1-\frac{\delta}{2}}$, $a = (6B + 2M)(1 + B)^2 n^{-\frac{3}{2}} + n^{-\frac{3}{2}+\delta}$, $b = n^{-\frac{3}{2}}$, $\xi = 1/2$, *and* $t = \sqrt{\frac{n}{4C(d+2)\log d}}$ *in Algorithm 4 ensures that with probability* $1 - d^{\Theta(-n)} - n^{-\frac{p}{2}+\delta(1+p)}$:

1. *the output of Algorithm 4 is* $O\left(n^{-1+\delta}\right)$*-jointly differentially private,*

2. *it is an $O\left(n^{-\frac{3}{2}+\delta}\right)$-approximate Bayes Nash equilibrium for a $1 - O\left(n^{-\delta}\right)$ fraction of players to truthfully report their data,*
3. *the computed estimate $\hat{\theta}^P$ is $O\left(n^{-\delta}\right)$-accurate,*
4. *it is individually rational for a $1 - O\left(n^{-\delta}\right)$ fraction of players to participate in the mechanism, and*
5. *the required budget from the analyst is $O\left(n^{-\frac{1}{2}+\delta}\right)$.*

Proof. Choose $\delta \in (0, \frac{p}{2+2p})$. Note that this ensures $\delta < 1/2$. Let $\alpha = n^{-\delta}$ and $\beta = n^{\frac{p}{2}-\delta(1+p)}$ as we have chosen. By the constraint that $\delta < \frac{p}{2+2p}$, we have ensured that $\beta = o(1)$. By Lemma 11, $\tau_{\alpha,\beta} \leq \max\{(\alpha\beta)^{-1/p}, \alpha^{-1/p}\} = (\alpha\beta)^{-1/p}$ since $\alpha, \beta = o(1)$ and $p > 1$. Then $\tau_{\alpha,\beta} = O\left(n^{1-\delta}\right)$.

By setting $\xi = 1/2$ and $t = \sqrt{\frac{n}{4C(d+2)\log d}}$, we ensure that with probability $1 - d^{-\frac{n}{4C(d+2)\log d}} = 1 - d^{\Theta(-n)}$, the bounds stated in Theorem 14 hold. With probability $1 - \beta$, at most an α-fraction of players will have cost parameters above $\tau_{\alpha,\beta}$. Taking a union bound over these two failure probabilities, the bounds in Theorems 16, 17, 18, 19, and 20 will all hold with probability at least $1 - d^{\Theta(-n)} - n^{-\frac{p}{2}+\delta(1+p)}$. For the remainder of the proof, we will assume all bounds hold, which will happen with at least the probability specified above.

First note that by Theorem 16, Algorithm 4 is 2ϵ-jointly differentially private. By our choice of ϵ, the privacy guarantee is $2n^{-1+\delta} = o(\sqrt{n})$.

Recall that by Theorem 17, it is a $\left[b\left(\frac{\alpha n}{\gamma}(4B + 2M) + \frac{\gamma B}{\gamma + (1-\xi)\frac{1}{d+2}n}\right)^2 + \tau_{\alpha,\beta}\epsilon^2\right]$-approximate Bayes-Nash equilibrium for a $(1 - \alpha)$-fraction of players to truthfully report their data. Taking B, M, ξ, and d to be constants, this strategy forms a $\Theta\left(b\left(\frac{\alpha n}{\gamma} + \frac{\gamma}{n}\right)^2 + \tau_{\alpha,\beta}\epsilon^2\right)$-approximate BNE. To achieve the desired truthfulness bound, we require (among other things) that $\tau_{\alpha,\beta}\epsilon^2 = o(\frac{1}{n})$. Given the bound on $\tau_{\alpha,\beta}$, it would suffice to have $\epsilon = o(n^{-\frac{3}{4}+\frac{\delta}{2}})$. This is satisfied by our choice of $\epsilon = n^{-1+\delta}$ because $\delta < 1/2$. After setting $b = o(\frac{1}{n})$, we will have the desired truthfulness bound if $\frac{\alpha n}{\gamma} + \frac{\gamma}{\gamma+n} = o(1)$. This implies the following constraints on γ: we require $\gamma = \omega(n\alpha) = \omega(n^{1-\delta})$ and $\gamma = o(n)$. Our choice of $\gamma = n^{1-\frac{\delta}{2}}$ satisfies these requirements. Due to our choice of $b = n^{-3/2}$, the approximation factor will be dominated by $\tau_{\alpha,\beta}\epsilon^2 = O\left(n^{-\frac{3}{2}+\delta}\right) = o(1)$. Thus truthtelling is an $O\left(n^{-\frac{3}{2}+\delta}\right) = o(1)$-approximate Bayes-Nash equilibrium for all but an $n^{-\delta} = o(1)$-fraction of players.

Recall from Theorem 18 that the computed estimate $\hat{\theta}^P$ is guaranteed to satisfy $O\left(\left(\frac{\alpha n}{\gamma}+\frac{1}{\gamma\epsilon}\right)^2+\left(\frac{\gamma}{\gamma+n}\right)^2+\left(\frac{1}{n}\right)^2+\frac{\alpha n}{\gamma}+\frac{1}{\gamma\epsilon}\right)$-accuracy. We have already established that $\frac{\alpha n}{\gamma}=o(1)$ and $\frac{\gamma}{\gamma+n}=o(1)$. Trivially, $\frac{1}{n^2}=o(1)$. We turn now to the term $\frac{1}{\gamma\epsilon}$. For this term to be $o(1)$, we require $\gamma=\omega(\frac{1}{\epsilon})=\omega\left(n^{1-\delta}\right)$. Our choice of $\gamma=n^{1-\frac{\delta}{2}}$ ensures this requirement is satisfied. Since $\frac{\alpha n}{\gamma}+\frac{1}{\gamma\epsilon}=o(1)$, then so must be $\left(\frac{\alpha n}{\gamma}+\frac{1}{\gamma\epsilon}\right)^2=o(1)$. The accuracy bound will be dominated by three terms: first $\left(\frac{\gamma}{n}\right)^2=n^{-\delta}$, second $\frac{\alpha n}{\gamma}=n^{-\frac{\delta}{2}}$, and third $\frac{1}{\gamma\epsilon}=n^{-\frac{\delta}{2}}$. Thus, Algorithm 4 outputs an estimator with accuracy $O\left(n^{-\frac{\delta}{2}}\right)=o(1)$.

Theorem 19 says that the mechanism in Algorithm 4 is individually rational for a $(1-\alpha)$-fraction of players as long as $a\geq\left(\frac{\alpha n}{\gamma}(4B+2M)+\frac{\gamma B}{\gamma+(1-\xi)\frac{1}{d+2}n}+B\right)(b+2bB)+bB^2+\tau_{\alpha,\beta}\epsilon^2$. We now expand each term of this expression to prove that our choice of a satisfies the desired bound. Consider the first term: $\frac{\alpha n}{\gamma}(4B+2M)=n^{-\frac{\delta}{2}}(4B+2M)$. This term is decreasing in n, so it can be upper bounded by its value when $n=1$. Thus $\frac{\alpha n}{\gamma}(4B+2M)\leq 4B+2M$. Now consider the second term:

$$\frac{\gamma B}{\gamma+(1-\xi)\frac{1}{d+2}n}=\frac{n^{1-\frac{\delta}{2}}B}{n^{1-\frac{\delta}{2}}+\frac{1}{2(d+2)}n}=\frac{n^{-\frac{\delta}{2}}B}{n^{-\frac{\delta}{2}}+\frac{1}{2(d+2)}}=B\left(1-\frac{1}{2(d+2)n^{-\frac{\delta}{2}}+1}\right).$$

The final term $\frac{-1}{2(d+2)n^{-\frac{\delta}{2}}+1}$ is always negative, so the entire term $\frac{\gamma B}{\gamma+(1-\xi)\frac{1}{d+2}n}$ can be bounded above by B. We can simplify the expression $b+2bB+bB^2$ as $(1+B)^2b=(1+B)^2n^{-3/2}$. Finally, as noted earlier (and due to to Lemma 11), we can upper bound $\tau_{\alpha,\beta}\epsilon^2\leq n^{-\frac{3}{2}+\delta}$. Combining all of these bounds, it would suffice to set $a\geq(6B+2M)(1+B)^2n^{-3/2}+n^{-\frac{3}{2}+\delta}$. We set a to be exactly this bound. Then it is individually rational for a $1-\alpha=1-n^{-\delta}=1-o(1)$ fraction of players to participate in the mechanism.

By Theorem 20, the budget required from the analyst is:

$$\mathcal{B}\leq n\left[a+\left(\frac{\alpha n}{\gamma}(4B+2M)+\frac{\gamma B}{\gamma+(1-\xi)\frac{1}{d+2}n}+B\right)(b+2bB)\right].$$

From our choice of $a=\Theta\left(n^{-\frac{3}{2}+\delta}\right)$ and because $\frac{\alpha n}{\gamma}+\frac{\gamma}{n}=o(1)$, the required budget is $\mathcal{B}=O\left(n(b+\tau_{\alpha,\beta}\epsilon^2)\right)=O\left(n(n^{-\frac{3}{2}}+n^{-\frac{3}{2}+\delta})\right)=O\left(n^{-\frac{1}{2}+\delta}\right)=o(1)$. □

Chapter 7

THE POSSIBILITIES AND LIMITATIONS OF PRIVATE PREDICTION MARKETS

7.1 Introduction

Betting markets of various forms—including the stock exchange [54], futures markets [88], sports betting markets [44], and markets at the racetrack [95]—have been shown to successfully collect and aggregate information. Over the last few decades, *prediction markets* designed specifically for the purpose of elicitation and aggregation, have yielded useful predictions in domains as diverse as politics [6], disease surveillance [85], and entertainment [82].

The desire to aggregate and act on the strategically valuable information dispersed among employees has led many companies to experiment with internal prediction markets. An internal corporate market could be used to predict the launch date of a new product or the product's eventual success. Among the first companies to experiment with internal markets were Hewlett-Packard, which implemented real-money markets, and Google, which ran markets using its own internal currency that could be exchanged for raffle tickets or prizes [84, 20]. More recently, Microsoft, Intel, Ford, GE, Siemens, and others have engaged in similar experiments [5, 13, 20].

Proponents of internal corporate markets often argue that the market structure helps in part because, without it, "business practices [...] create incentives for individuals not to reveal their information" [84]. However, even with a formal market structure in place, an employee might be hesitant to bet against the success of their team for fear of insulting her coworkers or angering management. If an employee has information that is unfavorable to the company, she might choose not to report it, leading to predictions that are overly optimistic for the company and ultimately contributing to an "optimism bias" in the market similar to the bias in Google's corporate markets discovered by Cowgill and Zitzewitz [20].

To address this issue, we consider the problem of designing *private* prediction markets. A private market would allow participants to engage in the market and contribute to the accuracy of the market's predictions without fear of having their information or beliefs revealed. The goal is to provide participants with a form of "plausible deniability." Although participants' trades or wagers should together influence the market's behavior and predictions, no single participant's actions should have too much influence over what others can observe. We formalize this idea using *differential privacy*, presented in Chapter 2, which can be used to guarantee that any participant's actions cannot be inferred from observations.

We begin in Section 7.3 by designing a private analog of the *weighted score wagering mechanisms* first introduced by Lambert et al. [68]. A *wagering mechanism* allows bettors to each specify a belief about the likelihood of a future event and a corresponding monetary wager. These wagers are then collected by a centralized operator and redistributed among bettors in such a way that more accurate bettors receive higher rewards. Lambert et al. [68] showed that the class of weighted score wagering mechanisms, which are built on the machinery of proper scoring rules [49], is the unique set of wagering mechanisms to satisfy a set of desired properties such as budget balance, truthfulness, and anonymity. We design a class of wagering mechanisms with randomized payments that maintain the nice properties of weighted score wagering mechanisms in expectation while additionally guaranteeing ϵ-joint differential privacy in the bettors' reported beliefs. We discuss the trade-offs that exist between the privacy of the mechanism (captured by the parameter ϵ) and the sensitivity of a bettor's payment to her own report, and show how to set the parameters of our mechanisms to achieve a reasonable level of the plausible deniability desired in practice.

We next address the problem of running private dynamic prediction markets in Section 7.4. There we consider the setting in which traders buy and sell securities with values linked to future events. For example, a market might offer a security worth \$1 if Microsoft Bing's market share increases in 2016 and \$0 otherwise. A risk neutral trader who believes that the probability of Bing's market share increasing is p would profit from buying this security at any price less than \$$p$ or (short) selling it at any price greater than \$$p$. The market price of the security is thought to reflect traders' collective beliefs about the likelihood of this event. We focus on *cost-function prediction markets* [15, 1] such as Hanson's popular logarithmic market scoring rule [55]. In a cost-function market, all trades are placed through an *automated market maker*, a centralized algorithmic agent that is always willing to buy or sell securities at some current market price that depends on the history of trade via a potential function called the cost function. We ask whether it is possible for a market maker to price trades according to a noisy cost function in a way that maintains traders' privacy without allowing traders to make unbounded profit off of the noise. Unfortunately, we show that under general assumptions, it is impossible for a market maker to achieve bounded loss and ϵ-differential privacy without allowing the privacy guarantee to degrade very quickly as the number of trades grows. In particular, the quantity e^{ϵ} must grown faster than linearly in the number of trades, making such markets impractical in settings in which privacy is valued. We suggest several avenues for future research aimed at circumventing this lower bound.

There is very little prior work on the design of private prediction markets, and to the best of our knowledge, we are the first to consider privacy for one-shot wagering mechanisms. Most closely related to our work is the recent paper of Waggoner, Frongillo, and Abernethy [99] who consider a setting in which each of a set of self-interested agents holds a private data point consisting of an observation x and corresponding label y. A firm would like to purchase the agents' data in order to learn a function to accurately predict the labels of new observations. Building on the mathematical foundations of cost-function market makers, Waggoner, Frongillo, and Abernethy propose a mechanism that provides incentives for the agents to reveal their data to the firm in such a way that the firm is able to solve its prediction task while maintaining the agents' privacy. The authors mention that similar ideas can be applied to produce privacy-preserving prediction markets, but their construction requires knowing the number of trades that will occur in advance to appropriately set parameters. The most straightforward way of applying their techniques to prediction markets results in a market maker falling in the class covered by our impossibility result, suggesting that such techniques cannot be used to derive a privacy-preserving market with bounded loss when the number trades is unknown.

7.2 Privacy with Streaming Data

The weighted score wagering mechanisms we design in Section 7.3 are *differentially private* and *joint differentially private* in the standard sense, according to Definitions 1 and 2, respectively. However, these definitions assume the input database D is fixed and finite, whereas the dynamic prediction markets we consider in Section 7.4 have input databases of possibly infinite size that arrive online in a *streaming* fashion, one element at a time.

Fortunately, differential privacy has also been considered for streaming algorithms [12, 37]. Let $\mathbb{N} = \{1, 2, 3, \ldots\}$. A stream $s \in \mathcal{D}^{\mathbb{N}}$ is a string of countable length of elements in $\mathcal{D}$, where $s_t \in \mathcal{D}$ denotes the element at position or *time* t and $s_{1,\ldots,t} \in \mathcal{D}^t$ is the length t prefix of the stream s. Two streams s and s' are said to be *neighbors* if they differ at exactly one time t.

A streaming algorithm $\mathcal{M}$ is said to be *unbounded* if it accepts streams of indefinite length, that is, if for any stream $s \in \mathcal{D}^{\mathbb{N}}$, $\mathcal{M}(s) \in \mathbb{R}^{\mathbb{N}}$. In contrast, a streaming algorithm is T*-bounded* if it accepts only streams of length at most T. Dwork et al. [37] consider only T-bounded streaming algorithms. Since we consider unbounded streaming algorithms, we use a more appropriate definition of differential privacy for streams adapted from Chan, Shi, and Song [12]. For unbounded streaming algorithms, it can be convenient to let the privacy guarantee degrade as the input stream grows in length. Chan, Shi, and Song [12] implicitly allow this in some of their results; see, for example, Corollary 4.5 in their paper. For clarity and preciseness, we explicitly capture this in our definition.

Definition 31 (Differential Privacy for Streams)**.** *For any non-decreasing function* $\epsilon : \mathbb{N} \to \mathbb{R}_+$ *and any* $\delta \geq 0$*, a streaming algorithm* $\mathcal{M} : \mathcal{D}^{\mathbb{N}} \to \mathbb{R}^{\mathbb{N}}$ *is* $(\epsilon(t), \delta)$-differentially private *if for every pair of neighboring streams* $s, s'' \in \mathcal{D}^{\mathbb{N}}$*, for every* $t \in \mathbb{N}$*, and for every subset* $\mathcal{S} \subseteq \mathbb{R}^t$*,*

$$\Pr[\mathcal{M}(s_{1,\ldots,t}) \in \mathcal{S}] \leq e^{\epsilon(t)} \Pr[\mathcal{M}(s'_{1,\ldots,t}) \in \mathcal{S}] + \delta.$$

If $\delta = 0$*, we say that* $\mathcal{M}$ *is* $\epsilon(t)$-differentially private.

Note that we allow ϵ to grow with t, but require that δ stay constant. In principle, one could also allow δ to depend on the length of the stream. However, allowing δ to increase would likely be unacceptable in scenarios in which privacy is considered important. In fact, it is more typical to require *smaller* values of δ for larger databases since for a database of size n, an algorithm could be considered (ϵ, δ)-private for δ on the order of $1/n$ even if it fully reveals a small number of randomly chosen database entries [34]. Since we use this definition only when showing an impossibility result, allowing δ to decrease in t would not strengthen our result.

7.3 Private Wagering Mechanisms

We begin with the problem of designing a one-shot wagering mechanism that incentivizes bettors to truthfully report their beliefs while maintaining their privacy. A wagering mechanism allows a set of bettors to each specify a belief about a future event and a monetary wager. Wagers are collected by a centralized operator and redistributed to bettors in such a way that bettors with more accurate predictions are more highly rewarded. Lambert et al. [68] showed that the class of *weighted score wagering mechanisms* (WSWMs) is the unique class of wagering mechanisms to satisfy a set of desired axioms such as budget balance and truthfulness. In this section, we show how to design a randomized wagering mechanism that achieves ϵ-joint differential privacy while maintaining the nice properties of WSWMs in expectation.

Standard wagering mechanisms

Wagering mechanisms, introduced by Lambert et al. [68], are mechanisms designed to allow a centralized operator to elicit the beliefs of a set of bettors without taking on any risk. In this paper we focus on *binary* wagering mechanisms, in which each bettor i submits a report $p_i \in [0,1]$ specifying how likely she believes it is that a particular event will occur, along with a wager $m_i \geq 0$ specifying the maximum amount of money that she is willing to lose. After all reports and wagers have been collected, all parties observe the realized outcome $\omega \in \{0,1\}$ indicating whether or not the event occurred. Each bettor i then receives a payment that is a function of the outcome and the reports and wagers of all bettors. This idea is formalized as follows.

Definition 32 (Wagering Mechanism [68])**.** *A* wagering mechanism *for a set of bettors $\mathcal{N} = \{1,\ldots,n\}$ is specified by a vector $\mathbf{\Pi}$ of (possibly randomized) profit functions, $\Pi_i : [0,1]^n \times \mathbb{R}_+^n \times \{0,1\} \rightarrow \mathbb{R}$, where $\Pi_i(\boldsymbol{p},\boldsymbol{m},\omega)$ denotes the total profit to bettor i when the vectors of bettors' reported probabilities and wagers are $\boldsymbol{p}$ and $\boldsymbol{m}$ and the realized outcome is ω. It is required that $\Pi_i(\boldsymbol{p},\boldsymbol{m},\omega) \geq -m_i$ for all $\boldsymbol{p}$, $\boldsymbol{m}$, and ω, which ensures that no bettor loses more than her wager.*

There are two minor differences between the definition presented here and that of Lambert et al. [68]. First, for convenience, we use Π_i to denote the *total* profit to bettor i (i.e., her payment from the mechanism minus her wager), unlike Lambert et al. [68], who use Π_i to denote the payment only. While this difference is inconsequential, we mention it to avoid confusion. Second, all previous work on wagering mechanisms has restricted attention to *deterministic* profit functions Π_i. Since randomization is necessary to attain privacy, we open up our study to *randomized* profit functions.

Lambert et al. [68] defined a set of desirable properties or axioms that deterministic wagering mechanisms should arguably satisfy. Here we adapt those properties to potentially randomized wagering mechanisms, making the smallest modifications possible to maintain the spirit of the axioms. Four of the properties (truthfulness, individual rationality, normality, and monotonicity) were originally defined in terms of expected profit with the expectation taken over some true or believed distribution over the outcome ω. We allow the expectation to be over the randomness in the profit function as well. Sybilproofness was not initially defined in expectation; we now ask that this property hold in expectation with respect to the randomness in the profit function. We define anonymity in terms of the distribution over all bettors' profits, and ask that budget balance hold for any realization of the randomness in $\mathbf{\Pi}$.

(a) **Budget balance:** The operator makes no profit or loss, i.e., $\forall \mathbf{p} \in [0,1]^n$, $\forall \mathbf{m} \in \mathbb{R}_+^n$, $\forall \omega \in \{0,1\}$, and for any realization of the randomness in $\mathbf{\Pi}$, $\sum_{i=1}^n \Pi_i(\mathbf{p},\mathbf{m},\omega) = 0$.

(b) **Anonymity:** Profits do not depend on the identify of the bettors. That is, for any permutation of the bettors σ, $\forall \mathbf{p} \in [0,1]^n$, $\forall \mathbf{m} \in \mathbb{R}_+^n$, $\forall \omega \in \{0,1\}$, the joint distribution over profit vectors $\{\Pi_i(\mathbf{p},\mathbf{m},\omega)\}_{i\in\mathcal{N}}$ is the same as the joint distribution over profit vectors $\{\Pi_{\sigma(i)}\left((p_{\sigma^{-1}(i)})_{i\in\mathcal{N}},(m_{\sigma^{-1}(i)})_{i\in\mathcal{N}},\omega\right)\}_{i\in\mathcal{N}}$.

(c) **Truthfulness:** Bettors uniquely maximize their expected profit by reporting the truth. That is, $\forall i \in \mathcal{N}$, $\forall \mathbf{p}_{-i} \in [0,1]^{n-1}$, $\forall \mathbf{m} \in \mathbb{R}_+^n$, $\forall p^*, p_i \in [0,1]$ with $p_i \neq p^*$,

$$\mathbb{E}_{\omega\sim p^*}\left[\Pi_i((p^*,\mathbf{p}_{-i}),\mathbf{m},\omega)\right] > \mathbb{E}_{\omega\sim p^*}\left[\Pi_i((p_i,\mathbf{p}_{-i}),\mathbf{m},\omega)\right].$$

(d) **Individual rationality:** Bettors prefer participating to not participating. That is, $\forall i \in \mathcal{N}$, $\forall m_i > 0$, for all $p^* \in [0,1]$, there exists some $p_i \in [0,1]$ such that $\forall \mathbf{p}_{-i} \in [0,1]^{n-1}$, $\forall \mathbf{m}_{-i} \in \mathbb{R}_+^{n-1}$, $E_{\omega\sim p^*}\left[\Pi_i((p_i,\mathbf{p}_{-i}),\mathbf{m},\omega)\right] \geq 0$.

(e) **Normality:**[1] If any bettor j changes her report, the change in the expected profit to any other bettor i with respect to a fixed belief p^* is the opposite sign of the change in expected payoff to j. That is, $\forall i,j \in \mathcal{N}$, $i \neq j$, $\forall \mathbf{p},\mathbf{p}' \in [0,1]^n$ with $p'_k = p_k$ for all $k \neq j$, $\forall p^* \in [0,1]$, $\forall \mathbf{m} \in \mathbb{R}_+^n$,

$$\mathbb{E}[\Pi_j(\mathbf{p},\mathbf{m},\omega)] < \mathbb{E}[\Pi_j(\mathbf{p}',\mathbf{m},\omega)] \implies \mathbb{E}[\Pi_i(\mathbf{p},\mathbf{m},\omega)] \geq \mathbb{E}[\Pi_i(\mathbf{p}',\mathbf{m},\omega)].$$

All expectations are taken w.r.t. $\omega \sim p^*$ and the randomness in the mechanism.

[1] Lambert et al. [67] and Chen et al. [17] used an alternative definition of normality for wagering mechanisms that essentially requires that if, from some agent i's perspective, the prediction of agent j improves, then i's expected profit decreases. This form of normality also holds for our mechanism.

(f) **Sybilproofness:** Profits remain unchanged as any subset of players with the same reports manipulate user accounts by merging accounts, creating fake identities, or transferring wagers. That is, $\forall \mathcal{S} \subset \mathcal{N}$, $\forall \mathbf{p}$ with $p_i = p_j$ for all $i, j \in \mathcal{S}$, $\forall \mathbf{m}, \mathbf{m}' \in \mathbb{R}_+^n$ with $m_i = m_i'$ for $i \notin \mathcal{S}$ and $\sum_{i \in \mathcal{S}} m_i = \sum_{i \in \mathcal{S}} m_i'$, $\forall \omega \in \{0,1\}$, two conditions hold:

$$\mathbb{E}\left[\Pi_i(\mathbf{p}, \mathbf{m}, \omega)\right] = \mathbb{E}\left[\Pi_i(\mathbf{p}, \mathbf{m}', \omega)\right] \qquad \forall i \notin \mathcal{S},$$

$$\sum_{i \in \mathcal{S}} \mathbb{E}\left[\Pi_i(\mathbf{p}, \mathbf{m}, \omega)\right] = \sum_{i \in \mathcal{S}} \mathbb{E}\left[\Pi_i(\mathbf{p}, \mathbf{m}', \omega)\right].$$

(g) **Monotonicity** The magnitude of a bettor's expected profit (or loss) increases as her wager increases. That is, $\forall i \in \mathcal{N}$, $\forall \mathbf{p} \in [0,1]^n$, $\forall \mathbf{m} \in \mathbb{R}_+^n$, $\forall M_i > m_i$, $\forall p^* \in [0,1]$, either $0 < \mathbb{E}_{\omega \sim \mathbf{p}^*}[\Pi_i(\mathbf{p}, (m_i, \mathbf{m}_{-i}), \omega)] < \mathbb{E}_{\omega \sim \mathbf{p}^*}[\Pi_i(\mathbf{p}, (M_i, \mathbf{m}_{-i}), \omega)]$ or $0 > \mathbb{E}_{\omega \sim \mathbf{p}^*}[\Pi_i(\mathbf{p}, (m_i, \mathbf{m}_{-i}), \omega)] > \mathbb{E}_{\omega \sim \mathbf{p}^*}[\Pi_i(\mathbf{p}, (M_i, \mathbf{m}_{-i}), \omega)]$.

Previously studied wagering mechanisms [68, 17, 67] achieve truthfulness by incorporating *strictly proper scoring rules* [91] into their profit functions. Scoring rules reward individuals based on the accuracy of their predictions about random variables. For a binary random variable, a scoring rule s maps a prediction or report $p \in [0,1]$ and an outcome $\omega \in \{0,1\}$ to a score. A strictly proper scoring rule incentivizes a risk neutral agent to report her true belief.

Definition 33 (Strictly proper scoring rule [91])**.** *A function* $s : [0,1] \times \{0,1\} \to \mathbb{R} \cup \{-\infty\}$ *is a* strictly proper scoring rule *if for all* $p, q \in [0,1]$ *with* $p \neq q$*,* $\mathbb{E}_{\omega \sim p}[s(p, \omega)] > \mathbb{E}_{\omega \sim p}[s(q, \omega)]$.

One common example is the Brier scoring rule [9], defined in Chapter 6 as $s(p, \omega) = 1 - 2(p - \omega)^2$. Note that for the Brier scoring rule, $s(p, \omega) \in [-1, 1]$ for all p and ω, but any strictly proper scoring rule with a bounded range can be scaled to have range $[0,1]$.

The WSWMs incorporate proper scoring rules, assigning each bettor a profit based on how her score compares to the wager-weighted average score of all bettors, as in Algorithm 5. Lambert et al. [68] showed that the set of WSWMs satisfy the seven axioms above and is the *unique* set of deterministic mechanisms that simultaneously satisfy budget balance, anonymity, truthfulness, normality, and sybilproofness.

Algorithm 5 Weighted-score wagering mechanisms [68]

```
Parameters: number of bettors $n$, strictly proper scoring rule $s$ with range in $[0,1]$

Solicit reports $\mathbf{p}$ and wagers $\mathbf{m}$
Realize state $\omega$
for $i = 1, \dots, n$ do
    Pay bettor $i$
        $$\Pi_i(\mathbf{p}, \mathbf{m}, \omega) = m_i \left( s(p_i, \omega) - \frac{\sum_{j \in \mathcal{N}} m_j s(p_j, \omega)}{\sum_{j \in \mathcal{N}} m_j} \right)$$
end for
```

Adding privacy

We would like our wagering mechanism to protect the privacy of each bettor i, ensuring that the $n-1$ other bettors cannot learn too much about i's report from their own realized profits, even if they collude. Note that paying each agent according to an independent scoring rule would easily achieve privacy, but would fail budget balance and sybilproofness. We formalize our desire to add privacy to the other good properties of weighted score wagering mechanisms using joint differential privacy.

(h) **ϵ-joint differential privacy:** The vector of profit functions satisfies ϵ-joint differential privacy, i.e., $\forall i \in \mathcal{N}$, $\forall \mathbf{p} \in [0,1]^n$, $\forall p_i' \in [0,1]$, $\forall \mathbf{m} \in \mathbb{R}_+^n$, $\forall \omega \in \{0,1\}$, and $\forall \mathcal{S} \subset \mathbb{R}_+^{n-1}$, $\Pr[\Pi_{-i}((p_i, \mathbf{p}_{-i}), \mathbf{m}, \omega) \in \mathcal{S}] \leq e^{\epsilon} \Pr[\Pi_{-i}((p_i', \mathbf{p}_{-i}), \mathbf{m}, \omega) \in \mathcal{S}]$.

This definition requires only that the report p_i of each bettor i be kept private, not the wager m_i. Private wagers would impose more severe limitations on the mechanism, even if wagers are restricted to lie in a bounded range; see Section 7.3 for a discussion. Note that if bettor i's report p_i is correlated with his wager m_i, as might be the case for a Bayesian agent [67], then just knowing m_i could reveal information about p_i. In this case, differential privacy would guarantee that other bettors can infer no more about p_i after observing their profits than they could from observing m_i alone. If bettors have immutable beliefs as assumed by Lambert et al. [68], then reports and wagers are not correlated and m_i reveals nothing about p_i.

Unfortunately, it is not possible to jointly obtain properties (a)–(h) with any reasonable mechanism. This is due to an inherent tension between budget balance and privacy. This is easy to see. Budget balance requires that a bettor i's profit is the negation of the sum of profits of the other $n-1$ bettors, i.e., $\Pi_i(\mathbf{p}, \mathbf{m}, \omega) =$

$-\sum_{j \neq i} \Pi_j(\mathbf{p}, \mathbf{m}, \omega)$. Therefore, under budget balance, the other $n-1$ bettors could always collude to learn bettor i's profit exactly. In order to obtain privacy, it would therefore be necessary for bettor i's profit to be differentially private in her own report, resulting in profits that are almost entirely noise. This is formalized in the following theorem. We omit a formal proof since it follows immediately from the argument described here.

Theorem 21. *Let $\mathbf{\Pi}$ be the vector of profit functions for any wagering mechanism that satisfies both budget balance and ϵ-joint differential privacy for any $\epsilon > 0$. Then for all $i \in \mathcal{N}$, Π_i is ϵ-differentially private in bettor i's report p_i.*

Since it is unsatisfying to consider mechanisms in which a bettor's profit is not sensitive to her own report, we require only that budget balance hold in expectation over the randomness of the profit function. An operator who runs many markets may be content with such a guarantee as it implies that he will not lose money on average.

(a′) **Budget balance in expectation:** The operator neither makes a profit nor a loss in expectation, i.e., $\forall \mathbf{p} \in [0,1]^n$, $\forall \mathbf{m} \in \mathbb{R}_+^n$, $\forall \omega \in \{0,1\}$, $\sum_{i=1}^n \mathbb{E}\left[\Pi_i(\mathbf{p}, \mathbf{m}, \omega)\right] = 0$.

Private weighted score wagering mechanisms

Motivated by the argument above, we seek a wagering mechanism that simultaneously satisfies properties (a′) and (b)–(h). Keeping Theorem 21 in mind, we would also like the wagering mechanism to be defined in such a way that each bettor i's profit is sensitive to her own report p_i. Sensitivity is difficult to define precisely, but loosely speaking, we would like it to be the case that 1) the magnitude of $\mathbb{E}\left[\Pi_i(\mathbf{p}, \mathbf{m}, \omega)\right]$ varies sufficiently with the choice of p_i, and 2) there is not too much noise or variance in a bettor's profit, i.e., $\Pi_i(\mathbf{p}, \mathbf{m}, \omega)$ is generally not too far from $\mathbb{E}\left[\Pi_i(\mathbf{p}, \mathbf{m}, \omega)\right]$.

A natural first attempt would be to employ the standard Laplace Mechanism [34] on top of a WSWM, adding independent Laplace noise to each bettor's profit. The resulting profit vector would satisfy ϵ-joint differential privacy, but since Laplace random variables are unbounded, a bettor could lose more than her wager. Adding other forms of noise does not help; to obtain differential privacy, the noise must be unbounded [36]. Truncating a bettor's profit to lie within a bounded range *after* noise is added could achieve privacy, but would result in a loss of truthfulness as the bettor's expected profit would no longer be a proper scoring rule.

Algorithm 6 Private wagering mechanism

```
Parameters: num bettors $n$, privacy param $\epsilon$, strictly proper scoring rule $s$ with range in $[0,1]$
Fix $\alpha = 1 - e^{-\epsilon}$ and $\beta = e^{-\epsilon}$
Solicit reports $\mathbf{p}$ and wagers $\mathbf{m}$
Realize state $\omega$
for $i = 1, \ldots, n$ do
    Independently draw random variable $x_i(p_i, \omega)$ such that
        $$x_i(p_i,\omega) = \begin{cases} 1 & \text{w.p. } \frac{\alpha s(p_i,\omega)+\beta}{1+\beta} \\ -\beta & \text{w.p. } \frac{1-\alpha s(p_i,\omega)}{1+\beta} \end{cases}$$
end for
for $i = 1, \ldots, n$ do
    Pay bettor $i$
        $$\Pi_i(\mathbf{p},\mathbf{m},\omega) = m_i \left( \alpha s(p_i,\omega) - \frac{\sum_{j \in \mathcal{N}} m_j x_j(p_j,\omega)}{\sum_{j \in \mathcal{N}} m_j} \right)$$
end for
```

Instead, we take a different approach. Like the WSWM, our *private wagering mechanism*, formally defined in Algorithm 6, rewards each bettor based on how good his score is compared with an aggregate measure of how good bettors' scores are on the whole. However, this aggregate measure is now calculated in a noisy manner. That is, instead of comparing a bettor's score to a weighted average of all bettors' scores, the bettor's score is compared to a weighted average of random variables that are equal to bettors' scores in expectation. As a result, each bettor's profit is, in expectation, equal to the profit she would receive using a WSWM, scaled down by a parameter α to ensure that no bettor ever loses more than her wager, as stated in the following lemma. The proof simply shows that for each i, $\mathbb{E}[x_i(p_i,\omega)] = \alpha s(p_i,\omega)$.

Lemma 18. *For any number of bettors $n > 0$ with reports $\boldsymbol{p} \in [0,1]^n$ and wagers $\boldsymbol{m} \in \mathbb{R}_+^n$, for any setting of the privacy parameter $\epsilon > 0$, for any outcome $\omega \in \{0,1\}$, the expected value of bettor i's profit $\Pi_i(\boldsymbol{p},\boldsymbol{m},\omega)$ under the private wagering mechanism with scoring rule s is equal to bettor i's profit under a WSWM with scoring rule αs.*

Proof. For each $i \in \mathcal{N}$,

$$\mathbb{E}[x_i(p_i,\omega)] = \frac{\alpha s(p_i,\omega) + \beta}{1+\beta} - \beta\frac{1 - \alpha s(p_i,\omega)}{1+\beta} = \alpha s(p_i,\omega) \tag{7.1}$$

and so

$$\mathbb{E}[\Pi_i(\mathbf{p},\mathbf{m},\omega)] = m_i\left(\alpha s(p_i,\omega) - \frac{\sum_{j\in\mathcal{N}} m_j \alpha s_j(p_j,\omega)}{\sum_{j\in\mathcal{N}} m_j}\right).$$

This is precisely the profit to bettor i in a WSWM with scoring rule αs. □

Using this lemma, we show that this mechanism does indeed satisfy joint differential privacy as well as the other desired properties.

Theorem 22. *The private wagering mechanism satisfies (a′) budget balance in expectation, (b) anonymity, (c) truthfulness, (d) individual rationality, (e) normality, (f) sybilproofness, (g) monotonicity, and (h) ϵ-joint differential privacy.*

Proof. Any WSWM satisfies budget balance in expectation (by satisfying budget balance), truthfulness, individual rationality, normality, sybilproofness, and monotonicity [68]. Since these are all defined in terms of expected profit, Lemma 18 implies that the private wagering mechanism satisfies them too.

Anonymity is easily observed since profits are defined symmetrically for all bettors.

Finally we show ϵ-joint differential privacy. We first prove that each random variable $x_i(p_i,\omega)$ is ϵ-differentially private in bettor i's report p_i which implies that the noisy aggregate of scores is private in all bettors' reports. We then apply the Billboard Lemma to show that the profit vector $\mathbf{\Pi}$ satisfies joint differential privacy.

To show that $x_i(p_i,\omega)$ is differentially private in p_i, for each of the two values that $x_i(p_i,\omega)$ can take on we must ensure that the ratio of the probability it takes this value under any report p and the probability it takes this value under any alternative report p' is bounded by e^ϵ. Fix any $\omega \in \{0,1\}$. Since s has range in $[0,1]$,

$$\frac{\Pr(x_i(p,\omega) = 1)}{\Pr(x_i(p',\omega) = 1)} = \frac{\alpha s(p,\omega) + \beta}{\alpha s(p',\omega) + \beta} \le \frac{\alpha + \beta}{\beta} = \frac{1 - e^{-\epsilon} + e^{-\epsilon}}{e^{-\epsilon}} = e^\epsilon,$$

$$\frac{\Pr(x_i(p,\omega) = -\beta)}{\Pr(x_i(p',\omega) = -\beta)} = \frac{1 - \alpha s(p,\omega)}{1 - \alpha s(p',\omega)} \le \frac{1}{1-\alpha} = \frac{1}{1 - (1 - e^{-\epsilon})} = e^\epsilon.$$

Thus $x_i(p_i,\omega)$ is ϵ-differentially private in p_i. By Theorem 4 of McSherry [72], the vector $(x_1(p_1,\omega),\ldots,x_n(p_n,\omega))$ (and thus any function of this vector) is ϵ-differentially private in the vector $\mathbf{p}$, since each $x_i(p_i,\omega)$ does not depend on the reports of anyone but i. Since we view the wagers m_i as constants, the quantity $\sum_{j\in\mathcal{N}} m_j x_j(p_j,\omega)/\sum_{j\in\mathcal{N}} m_j$ is also ϵ-differentially private in the reports $\mathbf{p}$. Call this quantity X.

To apply the Billboard Lemma, we can imagine the operator publicly announcing the quantity X to the bettors. Given access to X, each bettor is able to calculate her own profit $\Pi_i(\mathbf{p},\mathbf{m},\omega)$ using only her own input and the values α and ω. The Billboard Lemma implies that the vector of profits is ϵ-joint differentially private. □

Sensitivity of the mechanism

Having established that our mechanism satisfies properties (a′) and (b)–(h), we next address the sensitivity of the mechanism in terms of the two facets described above: range of achievable expected profits and the amount of noise in the profit function. This discussion sheds light on how to set ϵ in practice.

The first facet is quantified by Lemma 18. As α grows, the magnitude of bettors' expected profits grows, and the range of expected profits grows as well. When α approaches 1, the range of expected profits achievable through the private wagering mechanism approaches that of a standard WSWM with the same proper scoring rule.

Unfortunately, since $\alpha = 1 - e^{-\epsilon}$, larger values of α imply larger values of the privacy parameter ϵ. This gives us a clear tradeoff between privacy and magnitude of expected payments. Luckily, in practice, it is probably unnecessary for ϵ to be very small for most markets. A relatively large value of ϵ can still give bettors plausible deniability. For example, setting $\epsilon = 1$ implies that a bettor's report can only change the probability of another bettor receiving a particular profit by a factor of roughly 2.7 and leads to $\alpha \approx 0.63$, a tradeoff that may be considered acceptable in practice.

The second facet is quantified in the following theorem, which states that as more money is wagered by more bettors, each bettor's realized profit approaches its expectation. The bound depends on $||\mathbf{m}||_2/||\mathbf{m}||_1$. If all wagers are equal, this quantity is equal to $1/\sqrt{n}$ and bettors' profits approach their expectations as n grows. This is not the case at the other extreme, when there are a small number of bettors with wagers much larger than the rest. The proof uses Hoeffding's inequality to bound the difference between the quantity $m_j x_j(p_j,\omega)$ and its expectation.

Theorem 23. *For any $\delta \in [0,1]$, any $\epsilon > 0$, any number of bettors $n > 0$, any vectors of reports $\boldsymbol{p} \in [0,1]^n$ and wagers $\boldsymbol{m} \in \mathbb{R}_+^n$, with probability at least $1-\delta$, for all $i \in \mathcal{N}$, the profit Π_i output by the private wagering mechanism satisfies*

$$|\Pi_i(\boldsymbol{p},\boldsymbol{m},\omega) - \mathbb{E}[\Pi_i(\boldsymbol{p},\boldsymbol{m},\omega)]| \leq m_i\left(\frac{\|\boldsymbol{m}\|_2}{\|\boldsymbol{m}\|_1}(1+\beta)\sqrt{\frac{\ln(2/\delta)}{2}}\right).$$

Proof. For any $j \in \mathcal{N}$, consider the quantity $m_j x_j(p_j,\omega)$. From Equation 7.1, $\mathbb{E}[m_j x_j(p_j,\omega)] = m_j \alpha s(p_j,\omega)$. Additionally we can bound $m_j x_j(p_j,\omega) \in [-m_j\beta, m_j]$. Hoeffding's inequality then implies that with probability at least $1-\delta$,

$$\left|\sum_{j\in\mathcal{N}} m_j \alpha s(p_j,\omega) - \sum_{j\in\mathcal{N}} m_j x_j(p_j,\omega)\right| \leq \|\mathbf{m}\|_2(1+\beta)\sqrt{\frac{\ln(2/\delta)}{2}}.$$

From the definition of the private wagering mechanism and Lemma 18, we then have that with probability at least $1-\delta$, for any $i \in \mathcal{N}$,

$$\begin{aligned}|\Pi_i(\mathbf{p},\mathbf{m},\omega) - \mathbb{E}[\Pi_i(\mathbf{p},\mathbf{m},\omega)]| &= \frac{m_i}{\sum_{j\in\mathcal{N}} m_j}\left|\sum_{j\in\mathcal{N}} m_j \alpha s(p_j,\omega) - \sum_{j\in\mathcal{N}} m_j x_j(p_j,\omega)\right| \\ &\leq m_i \frac{\|\mathbf{m}\|_2}{\|\mathbf{m}\|_1}(1+\beta)\sqrt{\frac{\ln(2/\delta)}{2}}\end{aligned}$$

as desired. □

The following corollary shows that if all wagers are bounded in some range $[L,U]$, profits approach their expectations as the number of bettors grows.

Corollary 3. *Fix any L and U, $0 < L < U$. For any $\delta \in [0,1]$, any $\epsilon > 0$, any $n > 0$, any vectors of reports $\boldsymbol{p} \in [0,1]^n$ and wagers $\boldsymbol{m} \in [L,U]^n$, with probability at least $1-\delta$, for all $i \in \mathcal{N}$, the profit Π_i output by the private wagering mechanism satisfies*

$$|\Pi_i(\boldsymbol{p},\boldsymbol{m},\omega) - \mathbb{E}[\Pi_i(\boldsymbol{p},\boldsymbol{m},\omega)]| \leq m_i\left(\frac{U}{\sqrt{n}L}(1+\beta)\sqrt{\frac{\ln(2/\delta)}{2}}\right).$$

Keeping wagers private

Property (h) requires that bettors' reports be kept private but does not guarantee private wagers. The same tricks used in our private wagering mechanism could be applied to obtain a privacy guarantee for both reports and wagers if wagers are restricted to lie in a bounded range $[L,U]$, but this would come with a great loss in sensitivity. Under the most straightforward extension, the parameter α would need to be set to $(L/U)(1 - e^{-\epsilon/n})$ rather than $(1 - e^{-\epsilon})$, greatly reducing the scale of achievable profits and thus making the mechanism impractical in most settings.

Loosely speaking, the extra factor of L/U stems from the fact that a bettor's effect on the profit of any other bettor must be roughly the same whether he wagers the maximum amount or the minimum. The poor dependence on n is slightly more subtle. We created a private-belief mechanism by replacing each bettor j's score $s(p_j,\omega)$ in the WSWM with a random variable $x_j(p_j,\omega)$ that is ϵ-differentially private in p_j. To obtain private wagers, we would instead need to replace the full term $m_j s(p_j,\omega)/\sum_{k\in N} m_k$ with a random variable for each j. This term depends on the wagers of *all* n bettors in addition to p_j. Since each bettor's profit would depend on n such random variables, achieving ϵ-joint differential privacy would require that each random variable be ϵ/n-differentially private in each bettor's wager.

We believe that sacrifices in sensitivity are unavoidable and not merely an artifact of our techniques and analysis, but leave a formal lower bound to future work.

7.4 Limits of Privacy with Cost-Function Market Makers

In practice, prediction markets are often run using dynamic mechanisms that update in real time as new information surfaces. We now turn to the problem of adding privacy guarantees to continuous-trade markets. We focus our attention on cost-function prediction markets, in which all trades are placed through an automated market maker [55, 15, 1]. The market maker can be viewed as a streaming algorithm that takes as input a stream of trades and outputs a corresponding stream of market states from which trade prices can be computed. Therefore, the privacy guarantees we seek are in the form of Definition 31. We ask whether it is possible for the automated market maker to price trades according to a cost function while maintaining $\epsilon(t)$-differential privacy without opening up the opportunity for traders to earn unbounded profits, leading the market maker to experience unbounded loss. We show a mostly negative result: to achieve bounded loss, the privacy term $e^{\epsilon(t)}$ must grow faster than linearly in t, the number of rounds of trade.

For simplicity, we state our results for markets over a single binary security, though we believe they extend to cost-function markets over arbitrary security spaces.

Standard cost-function market makers

We consider a setting in which there is a single binary security that traders may buy or sell. After the outcome $\omega \in \{0,1\}$ has been revealed, a share of the security is worth \$1 if $\omega = 1$ and \$0 otherwise. A cost-function prediction market for this security is fully specified by a convex function C called the *cost function*. Let x_t be the number of shares that are bought or sold by a trader in the tth transaction; positive values of x_t represent purchases while negative values represent (short) sales. The market state after the first $t-1$ trades is summarized by a single value $q_t = \sum_{\tau=1}^{t-1} x_\tau$, and the tth trader is charged $C(q_t + x_t) - C(q_t) = C(q_{t+1}) - C(q_t)$. Thus the cost function can be viewed as a potential function, with $C(q_{t+1}) - C(0)$ capturing the amount of money that the market maker has collected from the first t trades. The *instantaneous price* at round t, denoted p_t, is the price per share of purchasing an infinitesimally small quantity of shares: $p_t = C'(q_t)$. This framework is summarized in Algorithm 7.

Algorithm 7 Cost-function market maker (parameters: cost function C)

```
Initialize: $q_1 = 0$
for $t = 1, 2, \ldots$ do
    Update instantaneous price $p_t = C'(q_t)$
    A trader buys $x_t \in \mathbb{R}$ shares and pays $C(q_t + x_t) - C(q_t)$
    Update market state $q_{t+1} = q_t + x_t$
end for
Realize outcome $\omega$
if $\omega = 1$ then
    for $t = 1, 2, \ldots$ do
        Market maker pays $x_t$ to the trader from round $t$
    end for
end if
```

The most common cost-function market maker is Hanson's log market scoring rule (LMSR) [55]. The cost function for the single-security version of LMSR can be written as $C(q) = b \log(e^{(q+a)/b} + 1)$ where $b > 0$ is a parameter controlling the rate at which prices change as trades are made and a controls the initial market price at state $q = 0$. The instantaneous price at any state q is $C'(q) = e^{(q+a)/b}/(e^{(q+a)/b} + 1)$.

Under mild conditions on C, all cost-function market makers satisfy several desirable properties, including natural notions of no-arbitrage and information incorporation [1]. We refer to any cost function C satisfying these mild conditions as a *standard cost function*. Although the market maker subsidizes trade, crucially its worst-case loss is bounded. This ensures that the market maker does not go bankrupt, even if traders are perfectly informed. Formally, there exists a finite bound B such that for any T, any sequence of trades $x_1, \ldots, x_T$, and any outcome $\omega \in \{0, 1\}$,

$$q_{T+1} \cdot \mathbb{1}(\omega = 1) - (C(q_{T+1}) - C(0)) \leq B,$$

where $\mathbb{1}$ is the indicator function that is 1 if its argument is true and 0 otherwise. The first term on the left-hand side is the amount that the market maker must pay (or collect from) traders when ω is revealed. The second is the amount collected from traders. For the LMSR with initial price $p_1 = 0.5$ ($a = 0$), the worst-case loss is $b \log(2)$.

The noisy cost-function market maker

Clearly the standard cost-function market maker does not ensure differential privacy. The amount that a trader pays is a function of the market state, the sum of all past trades. Thus anyone observing the stream of market prices could infer the exact sequence of past trades. To guarantee privacy while still approximating cost-function pricing, the marker maker would need to modify the sequence of published prices (or equivalently, market states) to ensure that such information leakage does not occur.

In this section, we define and analyze a *noisy* cost-function market maker. The noisy market maker prices trades according to a cost function, but uses a noisy version of the market state in order to mask the effect of past trades. In particular, the market maker maintains a noisy market state $q'_t = q_t + \eta_t$, where q_t is the true sum of trades and η_t is a (random) noise term. The cost of trade x_t is $C(q'_t + x_t) - C(q'_t)$, with the instantaneous price now $p_t = C'(q'_t)$. Since the noise term η_t must be large enough to mask the trade x_t, we limit trades to be some maximum size k. A trader who would like to buy or sell more than k shares must do this over multiple rounds. The full modified framework is shown in Algorithm 8. For now we allow the noise distribution $\mathcal{Y}$ to depend arbitrarily on the history of trade. This framework is general; the natural adaptation of the privacy-preserving data market of Waggoner, Frongillo, and Abernethy [99] to the single security prediction market setting would result in a market maker of this form, as would a cost-function market that used existing private streaming techniques for bit counting [12, 37] to keep noisy, private counts of trades.

In this framework, we can interpret the market maker as implementing a noise trader in a standard cost-function market. Under this interpretation, after a (real) trader purchases x_t shares at state q'_t, the market state momentarily moves to $q'_t + x_t =$

Algorithm 8 Noisy cost-function market maker (parameters: cost function C, distribution $\mathcal{Y}$ over noise $\{\eta_t\}$, maximum trade size k)

```
**Initialize:** $q_1 = 0$
Draw $\eta_1$ and set $q'_1 = \eta_1$
**for** $t = 1,2,\ldots$ **do**
    Update instantaneous price $p_t = C'(q'_t)$
    A trader buys $x_t \in [-k,k]$ shares and pays $C(q'_t + x_t) - C(q'_t)$
    Update true market state $q_{t+1} = q_t + x_t$
    Draw $\eta_{t+1}$ and update noisy market state $q'_{t+1} = q_{t+1} + \eta_{t+1}$
**end for**
Realize outcome $\omega$
**if** $\omega = 1$ **then**
    **for** $t = 1,2,\ldots$ **do**
        Market maker pays $x_t$ to the trader from round $t$
    **end for**
**end if**
```

$q_t + \eta_t + x_t = q_{t+1} + \eta_t$. The market maker, acting as a noise trader, then effectively "purchases" $\eta_{t+1} - \eta_t$ shares at this state for a cost of $C((q_{t+1} + \eta_t) + (\eta_{t+1} - \eta_t)) - C(q_{t+1} + \eta_t) = C(q_{t+1} + \eta_{t+1}) - C(q_{t+1} + \eta_t)$, bringing the market state to $q_{t+1} + \eta_{t+1} = q'_{t+1}$. The market maker makes this trade regardless of the impact on its own loss. These noise trades obscure the trades made by real traders, opening up the possibility of privacy.

However, these noisy trades also open up the opportunity for traders to profit off of the noise. For the market to be practical, it is therefore important to ensure that the property of bounded worst-case loss is maintained. For the noisy cost-function market maker, for any sequence of T trades $x_1, \ldots, x_T$, any outcome $\omega \in \{0,1\}$, and any *fixed* noise values $\eta_1, \ldots, \eta_T$, the loss of the market maker is

$$L_T(x_1,\ldots,x_T,\eta_1,\ldots,\eta_T,\omega) \equiv q_{T+1} \cdot \mathbb{1}(\omega = 1) - \sum_{t=1}^{T} \left(C(q'_t + x_t) - C(q'_t)\right) .$$

As before, the first term is the (possibly negative) amount that the market maker pays to traders when ω is revealed, and the second is the amount collected from traders (which no longer telescopes). Unfortunately, we cannot expect this loss to be bounded for *any* noise values; the market maker could always get extremely unlucky and draw noise values that traders can exploit. Instead, we consider a relaxed version of bounded loss which holds in expectation with respect to the noise values η_t.

In addition to this relaxation, one more modification is necessary. Note that traders can (and should) base their actions on the current market price. Therefore, if our loss guarantee only holds in expectation with respect to noise values η_t, then it is no longer sufficient to give a guarantee that is worst case over any sequences of trades. Instead, we allow the sequence of trades to depend on the realized noise, introducing a game between traders and the market maker. To formalize this, we imagine allowing an adversary to control the traders. We define the notion of a *strategy* for this adversary.

Definition 34 (Trader strategy)**.** *A* trader strategy *σ is a set of (possibly randomized) functions $\sigma = \{\sigma_1, \sigma_2, \ldots\}$, with each σ_t mapping a history of trades and noisy market states $(x_1, \ldots, x_{t-1}, q'_1, \ldots, q'_t)$ to a new trade x_t for the trader at round t.*

Let Σ be the set of all strategies. With this definition in place, we can formally define what it means for a noisy cost-function market maker to have bounded loss.

Definition 35 (Bounded loss for a noisy cost-function market maker)**.** *A noisy cost-function market maker with cost function C and distribution $\mathcal{D}$ over noise values $\eta_1, \eta_2, \ldots$ is said to have* bounded loss *if there exists a finite B such that for all strategies $\sigma \in \Sigma$, all times $T \geq 1$, and all $\omega \in \{0,1\}$,*

$$\mathbb{E}\left[L_T(x_1, \ldots, x_T, \eta_1, \ldots, \eta_T, \omega)\right] \leq B,$$

where the expectation is taken over the market's noise values $\eta_1, \eta_2, \ldots$ distributed according to $\mathcal{Y}$ and the (possibly randomized) actions $x_1, x_2, \ldots$ of a trader employing strategy σ. In this case, the loss of the market maker is said to be bounded by *B. The noisy cost-function market maker has* unbounded loss *if no such B exists.*

If the noise values were deterministic, this definition of worst-case loss would correspond to the usual one, but because traders react intelligently to the specific realization of noise, we must define worst-case loss in game-theoretic terms.

Limitations on privacy

By effectively acting as a noise trader, a noisy cost-function market maker can partially obscure trades. Unfortunately, the amount of privacy achievable through this technique is limited. In this section, we show that in order to simultaneously maintain bounded loss and achieve $\epsilon(t)$-differential privacy, the quantity $e^{\epsilon(t)}$ must grow faster than linearly as a function of the number of rounds of trade.

Before stating our result, we explain how to frame the market maker setup in the language of differential privacy. Recall from Section 7.2 that a differentially private unbounded streaming algorithm $\mathcal{M}$ takes as input a stream s of arbitrary length and outputs a stream of values that depend on s in a differentially private way. In the market setting, the stream s corresponds to the sequence of trades $\vec{x} = (x_1, x_2, \ldots)$. We think of the noisy cost-function market maker (Algorithm 8) as an algorithm $\mathcal{M}$ that, on any stream prefix $(x_1, \ldots, x_t)$, outputs the noisy market states $(q'_1, \ldots, q'_{t+1})$.[2] The goal is to find a market maker such that $\mathcal{M}$ is $\epsilon(t)$-differentially private.

One might ask whether it is necessary to allow the privacy guarantee to diminish as the the number of trades grows. When considering the problem of calculating noisy sums of bit streams, for example, Chan, Shi, and Song [12] are able to maintain a fixed privacy guarantee as their stream grows in length by instead allowing the accuracy of their counts to diminish. This approach doesn't work for us; we cannot achieve bounded loss yet allow the market maker's loss to grow with the number of trades.

Our result relies on one mild assumption on the distribution $\mathcal{Y}$ over noise. In particular, we require that the noise η_{t+1} be chosen independent of the current trade x_t.[3] We refer to this as the *trade-independent noise assumption*. The distribution of η_{t+1} may still depend on the round t, the history of trade $x_1, \ldots, x_{t-1}$, and the realizations of past noise terms, $\eta_1, \ldots, \eta_t$. This assumption is needed in the proof only to rule out unrealistic market makers that are specifically designed to monitor and infer the behavior of the specific adversarial trader that we consider, and the result likely holds even without it. However, it is not a terribly restrictive assumption as most standard ways of generating noise could be written in this form. For example, Chan, Shi, and Song [12] and Dwork et al. [37] show how to maintain a noisy count of the number of ones in a stream of bits. Both achieve this by computing the exact count and adding noise that is correlated across time but independent of the data. If similar ideas were used to choose the noise term in our setting, the trade-independent noise assumption would be satisfied. The noise employed in the mechanism of Waggoner, Frongillo, and Abernethy [99] also satisfies this assumption. Our impossibility result then implies that their market would have

[2] Announcing q'_t allows traders to infer the instantaneous price $p_t = C'(q'_t)$. It is equivalent to announcing p_t in terms of information revealed as long as C is strictly convex in the region around q'_t.

[3] The proof can be extended easily to the more general case in which the calculation of η_{t+1} is differentially private in x_t; we make the slightly stronger assumption to simplify presentation.

unbounded loss if a limit on the number of rounds of trade were not imposed. To obtain privacy guarantees, Waggoner, Frongillo, and Abernethy must assume that the number of trades is known in advance and can therefore be used to set relevant market parameters.

We now state the main result.

Theorem 24. *Consider any noisy cost-function market maker using a standard convex cost function C that is nonlinear in some region, a noise distribution $\mathcal{D}$ satisfying the trade-independent noise assumption, and a bound $k > 0$ on trade size. If the market maker satisfies bounded loss, then it cannot satisfy $(\epsilon(t), \delta)$-differential privacy for any function ϵ such that $e^{\epsilon(t)} = O(t)$ with any constant $\delta \in [0, 1)$.*

This theorem rules out bounded loss with $\epsilon(t) = \log(mt)$ for any constant $m > 0$. It is open whether it is possible to achieve $\epsilon(t) = m \log(t)$ (and therefore $e^{\epsilon(t)} = t^m$) for some $m > 1$, but such a guarantee would likely be insufficient in most practical settings.

Note that with unbounded trade size (i.e., $k = \infty$), our result would be trivial. A trader could change the market state (and hence the price) by an arbitrary amount in a single trade. To provide differential privacy, the noisy market state would then have to be independent of past trades. The noisy market price would not be reflective of trader beliefs, and the noise added could be easily exploited by traders to improve their profits. By imposing a bound on trade size, we only strengthen our negative result.

While the proof of Theorem 24 is quite technical, the intuition is simple. We consider the behavior of the noisy cost-function market maker when there is a single trader trading in the market repeatedly using a simple trading strategy. This trader chooses a *target state* q^*. Whenever the noisy market state q'_t is less than q^* (and so $p_t < p^* \equiv C'(q^*)$), the trader purchases shares, pushing the market state as close to q^* as possible. When the noisy state q'_t is greater than q^* (so $p_t > p^*$), the trader sells shares, again pushing the state as close as possible to q^*. Each trade makes a profit for the trader *in expectation* if it were the case that $\omega = 1$ with probability p^*. Since there is only a single trader, this means that each such trade would result in an expected loss with respect to p^* for the market maker. Unbounded expected loss for any p^* implies unbounded loss in the worst case—either when $\omega = 0$ or $\omega = 1$. The crux of the proof involves showing that in order achieve bounded loss against this trader, the amount of added noise η_t cannot be too big as t grows, resulting in a sacrifice of privacy.

To formalize this intuition, we first give a more precise description of the strategy σ^* employed by the single trader we consider.

Definition 36 (Target strategy)**.** *The* target strategy σ^* *with target* $q^* \in \mathbb{R}$ *chosen from a region in which C is nonlinear is defined as follows. For all rounds t,*

$$\sigma_t^*(x_1,\ldots,x_{t-1},q'_1,\ldots,q'_t) = \begin{cases} \min\{q^* - q'_t, k\} & \textit{if } q'_t \leq q^*, \\ -\min\{q'_t - q^*, k\} & \textit{otherwise}. \end{cases}$$

As described above, if $\omega = 1$ with probability p^*, a trader following this target strategy makes a non-negative expected profit on every round of trade. Furthermore, this trader makes an expected profit of at least some constant $\chi > 0$ on each round in which the noisy market state q'_t is more than a constant distance γ from q^*. The market maker must subsidize this profit, taking an expected loss with respect to p^* on each round. These ideas are formalized in Lemma 19, which lower bounds the expected loss of the market maker in terms of the probability of q'_t falling far from q^*. In this statement, D_C denotes the Bregman divergence[4] of C.

Lemma 19. *Consider a noisy cost-function market maker satisfying the conditions in Theorem 24 with a single trader following the target strategy* σ^* *with target* q^**. Suppose* $\omega = 1$ *with probability* $p^* = C'(q^*)$*. Then for any* γ *such that* $0 < \gamma \leq k$*,*

$$\mathbb{E}\left[L_T(x_1,\ldots,x_T,\eta_1,\ldots,\eta_T,\omega)\right] \geq \chi \sum_{t=1}^{T} \Pr(|q'_t - q^*| \geq \gamma),$$

where the expectation and probability are taken over the randomness in the noise values $\eta_1,\eta_2,\ldots$*, the resulting actions* $x_1,x_2,\ldots$ *of the trader, and the random outcome* ω*, and where* $\chi = \min\{D_C(q^*+\gamma,q^*), D_C(q^*-\gamma,q^*)\} > 0$.

Intuitively, Lemma 19 lower bounds the trader's (possibly complicated) per-round profit by χ if the noisy quantity is in $[q-\gamma, q+\gamma]$ and 0 otherwise. The value of χ will then the minimum of her expected profit from a trade that pushes the quantity either from $q-\gamma$ to q or from $q+\gamma$ to q. If the trader buys or sells any additional shares, it is because the noisy quantity was strictly outside of the range $[q-\gamma, q+\gamma]$.

[4] The *Bregman divergence* of a convex function F of a single variable is defined as $D_F(p,q) = F(p) - F(q) - F'(q)(p-q)$. The Bregman divergence is always non-negative. If F is strictly convex, it is strictly positive when the arguments are not equal.

We can then consider her trade in two parts: first, the $|x_t - \gamma|$ shares that moved the quantity from $q_t + \eta_t$ to $q \pm \gamma$, and second, the remaining γ shares that moved the quantity from exactly $q \pm c$ to q, from which she receives expected profit at most χ. As described in the proof of Lemma 19, her per-share profit from the first portion of the trade must be higher than per-share profit from the latter portion of the trade, so her expected profit from the entire trade is at least:

$$\mathbb{E}_p\left[\text{profit}(x_t)\right] \geq \frac{\chi}{\gamma}(|x_t - \gamma|) + \chi \geq \chi.$$

The proof of Lemma 19 makes use of the following technical lemma, which says that it is profitable in expectation to sell shares as long as the price remains above p^* or to purchase shares as long as the price remains below p^*. In this statement, q can be interpreted as the current market state and x as a new purchase (or sale); $C'(q^*)x - C(q+x) + C(q) \geq 0$ would then be the expected profit of a trader making this purchase or sale if $\omega \sim p^* = C'(q^*)$.

Lemma 20. *Fix any convex function C and any q^*, q, and x such that $q + x \geq q^*$ if $x \leq 0$ and $q + x \leq q^*$ if $x \geq 0$. Then $C'(q^*)x - C(q + x) + C(q) \geq 0$.*

Proof. Since C is convex, the assumptions in the lemma statement imply that if $x \leq 0$ then $C'(q + x) \geq C'(q^*)$, while if $x \geq 0$ then $C'(q + x) \leq C'(q^*)$. Therefore, in either case $C'(q + x)x \leq C'(q^*)x$, and

$$C'(q^*)x - C(q + x) + C(q) \geq C'(q + x)x - C(q + x) + C(q) = D_C(q, q + x) \geq 0.$$

□

Proof of Lemma 19. From the definition of the market maker's loss, we can rewrite $\mathbb{E}\left[L_T(x_1, \ldots, x_T, \eta_1, \ldots, \eta_T, \omega)\right] = \sum_{t=1}^{T} \mathbb{E}[\pi_t]$, where π_t is the expected (over just the randomness in ω) loss of the market maker from the tth trade, i.e.,

$$\pi_t = C'(q^*)x_t - C(q'_t + x_t) + C(q'_t).$$

By definition of the target strategy $\vec{s}^*$ and Lemma 20, $\pi_t \geq 0$ for all t.

Consider a round t in which $|q'_t - q^*| \geq \gamma$. Suppose first that $q'_t \geq q^* + \gamma$, so a trader playing the target strategy would sell. By definition of $\vec{s}^*$, $x_t = -\min\{q'_t - q^*, k\} \leq$

$-\gamma$. We can write

$$\begin{aligned}
\pi_t &= C'(q^*)(x_t + \gamma) - C'(q^*)\gamma - C(q'_t + x_t) + C(q'_t - \gamma) - C(q'_t - \gamma) + C(q'_t) \\
&\geq -C'(q^*)\gamma - C(q'_t - \gamma) + C(q'_t) \\
&\geq -C'(q^*)\gamma - C(q^*) + C(q^* + \gamma) \\
&= D_C(q^* + \gamma, q^*) \geq \chi,
\end{aligned}$$

where χ is defined as in the lemma statement. The first inequality follows from an application of Lemma 20 with $q = q'_t - \gamma$ and $x = x_t + \gamma$. The second follows from the convexity of C and the assumption that $q'_t \geq q^* + \gamma$.

If instead $q'_t \leq q^* - \gamma$ (so a trader playing the target strategy would buy), a similar argument can be made to show that $\pi_t \geq D_C(q^* - \gamma, q^*) \geq \chi$.

Putting this all together, we have

$$\mathbb{E}\left[L_T(x_1, \ldots, x_T, \eta_1, \ldots, \eta_T, \omega)\right] = \sum_{t=1}^{T} \mathbb{E}[\pi_t] \geq \sum_{t=1}^{T} \chi \Pr(|q'_t - q^*| \geq \gamma)$$

as desired. The fact that $\chi > 0$ follows from the fact that it is the minimum of two Bregman divergences, each of which is strictly positive since C is nonlinear (and thus strictly convex) in the region around q^* and the arguments are not equal. □

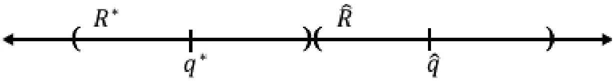
We now complete the proof of the main result.

Proof of Theorem 24. We will show that bounded loss implies $(\epsilon(t), \delta)$-differential privacy cannot be achieved with $e^{\epsilon(t)} = O(t)$ for any constant $\delta \in [0, 1)$.

Throughout the proof, we reason about the probabilities of various events conditioned on there being a single trader playing a particular strategy. All strategies we consider are deterministic, so all probabilities are taken just with respect to the randomness in the market maker's added noise $(\eta_1, \eta_2, \ldots)$.

As described above, we focus on the case in which a single trader plays the target strategy σ^* with target q^*. Define R^* to be the open region of radius $k/4$ around q^*, that is, $R^* = (q^* - k/4, q^* + k/4)$. Let $\hat{q} = q^* + k/2$ and let $\hat{R} = (\hat{q} - k/4, \hat{q} + k/4)$. Notice that R^* and $\hat{R}$ do not intersect, but from any market state $q \in R^*$ a trader could move the market state to $\hat{q}$ with a purchase or sale of no more than k shares.

For any round t, let σ^t be the strategy in which $\sigma^t_\tau = \sigma^*_\tau$ for all rounds $\tau \neq t$, but $\sigma^t_t(x_1, \ldots, x_{t-1}, q'_1, \ldots, q'_t) = \hat{q} - q'_t$ if $|\hat{q} - q'_t| \leq k$ (otherwise, σ^t_t can be defined



arbitrarily). In other words, a trader playing strategy σ^t behaves identically to a trader playing strategy σ^* on all rounds except round t. On round t, the trader instead attempts to move the market state to $\hat{q}$.

For any t, the behavior of a trader playing strategy σ^* and a trader playing strategy σ^t are indistinguishable through round $t-1$, and therefore the behavior of the market maker is indistinguishable as well. At round t, if it is the case that $q'_t \in R^*$ (and therefore $|q'_t - q^*| \leq k/4 < k$ and also $|q'_t - \hat{q}| \leq 3k/4 < k$), then a trader playing strategy σ^* would purchase $q^* - q'_t$ shares, while a trader playing strategy σ^t would purchase $\hat{q} - q'_t$. Differential privacy tells us that conditioned on such a state being reached, the probability that q'_{t+1} lies in any range (and in particular, in R^*) should not be too different depending on which of the two actions the trader takes. More formally, if the market maker satisfies $\epsilon(t)$-differential privacy, then for all rounds t,

$$\begin{aligned} e^{\epsilon(t)} &\geq \frac{\Pr(q'_{t+1} \in R^* | \sigma = \sigma^*, q'_t \in R^*) - \delta}{\Pr(q'_{t+1} \in R^* | \sigma = \sigma^t, q'_t \in R^*)} \geq \frac{\Pr(q'_{t+1} \in R^* | \sigma = \sigma^*, q'_t \in R^*) - \delta}{\Pr(q'_{t+1} \notin \hat{R} | \sigma = \sigma^t, q'_t \in R^*)} \\ &= \frac{\Pr(q'_{t+1} \in R^* | \sigma = \sigma^*, q'_t \in R^*) - \delta}{\Pr(q'_{t+1} \notin R^* | \sigma = \sigma^*, q'_t \in R^*)}. \end{aligned}$$

The first inequality follows from the definition of $(\epsilon(t), \delta)$-differential privacy. The second follows from the fact that R^* and $\hat{R}$ are disjoint. The last line is a consequence of the trade-independent noise assumption. By simple algebraic manipulation, for all t,

$$\Pr(q'_{t+1} \notin R^* | \sigma = \sigma^*, q'_t \in R^*) \geq \frac{1-\delta}{1+e^{\epsilon(t)}}. \tag{7.2}$$

We now further investigate the term on the left-hand side of this equation. For the remainder of the proof, we assume that $\sigma = \sigma^*$ and implicitly condition on this.

Applying Lemma 19 with $\gamma = k/4$, we find that the expected value of the market maker's loss after T rounds if $\omega = 1$ with probability $p^* = C'(q*)$ is lower bounded by $\chi \sum_{t=1}^{T} \Pr(q'_t \notin R^*)$ for the appropriate constant χ. This implies that for at least one of $\omega = 1$ or $\omega = 0$, $\mathbb{E}\left[L_T(x_1, \ldots, x_T, \eta_1, \ldots, \eta_T, \omega)\right] \geq \chi \sum_{t=1}^{T} \Pr(q'_t \notin R^*)$, where the expectation is just over the random noise of the market maker and the resulting actions of the trader. Since we have assumed that the market maker's loss is bounded, this implies there must exist some loss bound B such that

$$\frac{B}{\chi} \geq \sum_{t=1}^{\infty} \Pr(q'_t \notin R^*). \tag{7.3}$$

Fix any constant $\alpha \in (0, 1)$. Equation 7.3 implies that for all but finitely many t, $\Pr(q'_t \notin R^*) < \alpha$, or equivalently, for all but finitely many t, $\Pr(q'_t \in R^*) \geq 1 - \alpha$. Call the set of t for which this holds $\mathcal{T}$. Equation 7.3 also implies that

$$\begin{aligned}\frac{B}{\chi} &\geq \sum_{t=1}^{\infty} \left[\Pr(q'_{t+1} \notin R^* | q'_t \in R^*)\Pr(q'_t \in R^*) + \Pr(q'_{t+1} \notin R^* | q'_t \notin R^*)\Pr(q'_t \notin R^*)\right] \\ &\geq \sum_{t=1}^{\infty} \Pr(q'_{t+1} \notin R^* | q'_t \in R^*)\Pr(q'_t \in R^*) \geq (1-\alpha) \sum_{t \in \mathcal{T}} \Pr(q'_{t+1} \notin R^* | q'_t \in R^*).\end{aligned}$$

Combining this with Equation 7.2 yields

$$\sum_{t \in \mathcal{T}} \frac{1-\delta}{1+e^{\epsilon(t)}} \leq \frac{B}{\chi(1-\alpha)}. \tag{7.4}$$

Now suppose for contradiction that $e^{\epsilon(t)} = O(t)$. Then by definition, for some constant $m > 1$ there exists a round τ such that for all $t > \tau$, $e^{\epsilon(t)} \leq mt$. Then

$$\sum_{t \in \mathcal{T}} \frac{1-\delta}{1+e^{\epsilon(t)}} \geq \sum_{t \in \mathcal{T}, t > \tau} \frac{1-\delta}{1+e^{\epsilon(t)}} \geq \sum_{t \in \mathcal{T}, t > \tau} \frac{1-\delta}{1+mt} > \frac{1-\delta}{m} \sum_{t \in \mathcal{T}, t > \tau} \frac{1}{1+t}.$$

Since this sum is over all natural numbers t except a finite number, it must diverge, and therefore Equation 7.4 cannot hold. Therefore, we cannot have $e^{\epsilon(t)} = O(t)$. □

7.5 Concluding Remarks

We designed a class of randomized wagering mechanisms that keep bettors' reports private while maintaining truthfulness, budget balance in expectation, and other desirable properties of weighted score wagering mechanisms. The parameters of our mechanisms can be tuned to achieve a tradeoff between the level of privacy guaranteed and the sensitivity of a bettor's payment to her own report. Determining how to best make this tradeoff in practice (and more generally, what level of privacy is acceptable in differentially private algorithms) is an open empirical question.

While our results in the dynamic setting are negative, there are several potential avenues for circumventing our lower bound. The lower bound shows that it is not possible to obtain reasonable privacy guarantees using a noisy cost-function market maker when traders may buy or sell fractional security shares, as is typically assumed in the cost function literature. Indeed, the adversarial trader we consider buys and sells arbitrarily small fractions when the market state is close to its target. This behavior could be prevented by enforcing a minimum unit of purchase. Perhaps cleverly designed noise could allow us to avoid the lower bound with this additional restriction. However, based on preliminary simulations of a noisy cost-function market based on Hanson's LMSR 2003 with noise drawn using standard binary streaming approaches [37, 12], it appears an adversary can still cause a market maker using these "standard" techniques to have unbounded loss by buying one unit when the noisy market state is below the target and selling one unit when it is above.

One could also attempt to circumvent the lower bound by adding a transaction fee for each trade that is large enough that traders cannot profit off the market's noise. While the fee could always be set large enough to guarantee bounded loss, a large fee would discourage trade in the market and limit its predictive power. A careful analysis would be required to ensure that the fee could be set high enough to maintain bounded loss without rendering the market predictions useless.

BIBLIOGRAPHY

[1] Jacob Abernethy, Yiling Chen, and Jennifer Wortman Vaughan. "Efficient Market Making via Convex Optimization, and a Connection to Online Learning". In: *ACM Transactions on Economics and Computation* 1.2 (2013).

[2] Alessandro Acquisti, Curtis Taylor, and Liad Wagman. "The economics of privacy". Forthcoming: Journal of Economic Literature. 2014.

[3] "Apple Previews iOS 10, the Biggest iOS Release Ever". In: *Apple Press Info* (June 2016). URL: `https://www.apple.com/pr/library/2016/06/13Apple-Previews-iOS-10-The-Biggest-iOS-Release-Ever.html`.

[4] Raef Bassily, Adam Smith, and Abhradeep Thakurta. "Private Empirical Risk Minimization, Revisited". In: *arXiv preprint 1405.7085* (2014).

[5] Henry Berg and Todd A. Proebsting. "Hanson's Automated Market Maker". In: *Journal of Prediction Markets* 3.1 (2009), pp. 45–59.

[6] Joyce E. Berg et al. "Results from a dozen years of election futures markets research". In: *Handbook of Experimental Economic Results*. Ed. by C. A. Plott and V. Smith. 2001.

[7] J. Eric Bickel. "Some Comparisons among Quadratic, Spherical, and Logarithmic Scoring Rules". In: *Decision Analysis* 4.2 (June 2007), pp. 49–65.

[8] Avrim Blum et al. "Privacy-preserving public information for sequential games". In: *Proceedings of the 2015 Conference on Innovations in Theoretical Computer Science*. ACM. 2015, pp. 173–180.

[9] Glenn W. Brier. "Verification of Forecasts Expressed in Terms of Probability". In: *Monthly Weather Review* 78.1 (1950), pp. 1–3.

[10] Yang Cai, Constantinos Daskalakis, and Christos H. Papadimitriou. "Optimum Statistical Estimation with Strategic Data Sources". In: *arXiv preprint 1408.2539* (2014).

[11] Giacomo Calzolari and Alessandro Pavan. "On the optimality of privacy in sequential contracting". In: *Journal of Economic Theory* 130.1 (2006), pp. 168–204.

[12] T.-H. Hubert Chan, Elaine Shi, and Dawn Song. "Private and continual release of statistics". In: *ACM Transactions on Information and System Security* 14.3 (2011), p. 26.

[13] Robert Charette. "An Internal Futures Market". In: *Information Management* (2007).

[14] Kamalika Chaudhuri, Claire Monteleoni, and Anand D. Sarwate. “Differentially Private Empirical Risk Minimization”. In: *J. Mach. Learn. Res.* 12 (July 2011), pp. 1069–1109.

[15] Yiling Chen and David M. Pennock. “A Utility Framework for Bounded-Loss Market Makers”. In: *Proc. of the Conference on Uncertainty in Artificial Intelligence*. 2007.

[16] Yiling Chen, Or Sheffet, and Salil Vadhan. “Privacy games”. In: *Web and Internet Economics*. Springer, 2014, pp. 371–385.

[17] Yiling Chen et al. “Removing Arbitrage from Wagering Mechanisms”. In: *Proceedings of the 15th ACM Conference on Economics and Computation*. 2014.

[18] Yiling Chen et al. “Truthful Mechanisms for Agents That Value Privacy”. In: *Proceedings of the 14th ACM Conference on Electronic Commerce*. EC ’13. ACM. 2013, pp. 215–232.

[19] Vincent Conitzer, Curtis R Taylor, and Liad Wagman. “Hide and seek: Costly consumer privacy in a market with repeat purchases”. In: *Marketing Science* 31.2 (2012), pp. 277–292.

[20] Bo Cowgill and Eric Zitzewitz. “Corporate Prediction Markets: Evidence from Google, Ford, and Firm X”. In: *Review of Economic Studies* 82.4 (2015), pp. 1309–1341.

[21] Rachel Cummings, Federico Echenique, and Adam Wierman. “The Empirical Implications of Privacy-Aware Choice”. In: *Operations Research* 64.1 (2016). Preliminary Version appeared in the Proceedings of the 15th ACM Conference on Electronic Commerce (EC 2014), pp. 67–78.

[22] Rachel Cummings, Stratis Ioannidis, and Katrina Ligett. “Truthful Linear Regression”. In: *Proceedings of The 28th Conference on Learning Theory*. COLT ’15. 2015, pp. 448–483.

[23] Rachel Cummings, David M. Pennock, and Jennifer Wortman Vaughan. “The Possibilities and Limitations of Private Prediction Markets”. In: *Proceedings of the 17th ACM Conference on Economics and Computation*. EC ’16. 2016, pp. 143–160.

[24] Rachel Cummings et al. “Accuracy for Sale: Aggregating Data with a Variance Constraint”. In: *Proceedings of the 2015 Conference on Innovations in Theoretical Computer Science*. ITCS ’15. 2015, pp. 317–324.

[25] Rachel Cummings et al. “Coordination Complexity: Small Information Coordinating Large Populations”. In: *Proceedings of the 2016 ACM Conference on Innovations in Theoretical Computer Science*. ITCS ’16. 2016, pp. 281–290.

[26] Rachel Cummings et al. "Privacy and Truthful Equilibrium Selection for Aggregative Games". In: *Proceedings of the 11th International Conference on Web and Internet Economics*. WINE '15. 2015, pp. 286–299.

[27] Rachel Cummings et al. "The Strange Case of Privacy in Equilibrium Models". In: *Proceedings of the 17th ACM Conference on Economics and Computation*. EC '16. 2016, pp. 659–659.

[28] Amit Datta, Michael Carl Tschantz, and Anupam Datta. "Automated Experiments on Ad Privacy Settings". In: *Proceedings on Privacy Enhancing Technologies* 1.1 (2015), pp. 92–112.

[29] Andrew F Daughety and Jennifer F Reinganum. "Public goods, social pressure, and the choice between privacy and publicity". In: *American Economic Journal: Microeconomics* 2.2 (2010), pp. 191–221.

[30] Eddie Dekel, Barton L Lipman, and Aldo Rustichini. "Representing preferences with a unique subjective state space". In: *Econometrica* 69.4 (2001), pp. 891–934.

[31] Eddie Dekel, Barton L. Lipman, and Aldo Rustichini. "Temptation-Driven Preferences". In: *The Review of Economic Studies* 76.3 (2009), pp. 937–971.

[32] Ofer Dekel, Felix Fischer, and Ariel D. Procaccia. "Incentive compatible regression learning". In: *Journal of Computer and System Sciences* 76.8 (2010), pp. 759–777.

[33] Cynthia Dwork. "Differential privacy: A survey of results". In: *Theory and Applications of Models of Computation*. Springer, 2008, pp. 1–19.

[34] Cynthia Dwork and Aaron Roth. "The algorithmic foundations of differential privacy". In: *Foundations and Trends in Theoretical Computer Science* 9.34 (2014), pp. 211–407.

[35] Cynthia Dwork, Guy N. Rothblum, and Salil Vadhan. "Boosting and Differential Privacy". In: *Proceedings of the IEEE 51st Annual Symposium on Foundations of Computer Science*. FOCS '10. 2010, pp. 51–60.

[36] Cynthia Dwork et al. "Calibrating noise to sensitivity in private data analysis". In: *Proceedings of the 3rd Conference on Theory of Cryptography*. TCC '06. 2006, pp. 265–284.

[37] Cynthia Dwork et al. "Differential privacy under continual observation". In: *Proceedings of the 42nd ACM Symposium on Theory of Computing*. 2010.

[38] Cynthia Dwork et al. "Fairness through awareness". In: *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*. ACM. 2012, pp. 214–226.

[39] Cynthia Dwork et al. “On the complexity of differentially private data release: efficient algorithms and hardness results”. In: *STOC ’09*. 2009, pp. 381–390.

[40] Cynthia Dwork et al. “Our Data, Ourselves: Privacy Via Distributed Noise Generation”. In: *EUROCRYPT*. 2006, pp. 486–503.

[41] Úlfar Erlingsson, Vasyl Pihur, and Aleksandra Korolova. “RAPPOR: Randomized Aggregatable Privacy-Preserving Ordinal Response”. In: *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security*. CCS ’14. 2014, pp. 1054–1067.

[42] Lisa Fleischer and Yu-Han Lyu. “Approximately optimal auctions for selling privacy when costs are correlated with data”. In: *Proceedings of the 13th ACM Conference on Electronic Commerce*. EC ’12. 2012, pp. 568–585.

[43] Xavier Freixas, Roger Guesnerie, and Jean Tirole. “Planning under incomplete information and the ratchet effect”. In: *The review of economic studies* 52.2 (1985), pp. 173–191.

[44] John M. Gandar et al. “Informed traders and price variations in the betting market for professional basketball games”. In: *Journal of Finance* LIII.1 (1999), pp. 385–401.

[45] John Geanakoplos, David Pearce, and Ennio Stacchetti. “Psychological games and sequential rationality”. In: *Games and Economic Behavior* 1.1 (1989), pp. 60–79.

[46] Arpita Ghosh and Katrina Ligett. “Privacy and coordination: computing on databases with endogenous participation.” In: *ACM Conference on Electronic Commerce*. 2013, pp. 543–560.

[47] Arpita Ghosh and Aaron Roth. “Selling privacy at auction”. In: *Games and Economic Behavior* 91 (2015). Preliminary Version appeared in the Proceedings of the 12th ACM Conference on Electronic Commerce (EC 2011), pp. 334–346.

[48] Arpita Ghosh et al. “Buying Private Data Without Verification”. In: *Proceedings of the Fifteenth ACM Conference on Economics and Computation*. EC ’14. 2014, pp. 931–948.

[49] Tilmann Gneiting and Adrian E. Raftery. “Strictly Proper Scoring Rules, Prediction, and Estimation”. In: *Journal of the American Statistical Association* 102.477 (2007), pp. 359–378.

[50] Avi Goldfarb and Catherine Tucker. “Shifts in Privacy Concerns”. In: *American Economic Review* 102.3 (2012), pp. 349–353.

[51] Avi Goldfarb et al. *The Effect of Social Interaction on Economic Transactions: An Embarrassment of Niches?* 2013.

[52] Ronen Gradwohl. “Privacy in Implementation”. In: *CMS-EMS Discussion Paper 1561*. 2013.

[53] Ronen Gradwohl and Rann Smorodinsky. “Subjective perception games and privacy”. In: *arXiv preprint arXiv:1409.1487* (2014).

[54] Sanford J. Grossman. “On the Efficiency of Competitive Stock Markets Where Traders Have Diverse Information”. In: *The Journal of Finance* 31.2 (1976), pp. 573–585.

[55] Robin Hanson. “Combinatorial Information Market Design”. In: *Information Systems Frontiers* 5.1 (2003), pp. 105–119.

[56] Moritz Hardt and Guy N. Rothblum. “A Multiplicative Weights Mechanism for Privacy-Preserving Data Analysis”. In: *FOCS*. 2010, pp. 61–70.

[57] John C. Harsanyi. “Games with incomplete information played by Bayesian players”. In: *Management Science* 14.3 (1967), pp. 159–183.

[58] Oliver D Hart and Jean Tirole. “Contract renegotiation and Coasian dynamics”. In: *The Review of Economic Studies* 55.4 (1988), pp. 509–540.

[59] Ori Heffetz and Katrina Ligett. *Privacy and Data-Based Research*. Tech. rep. National Bureau of Economic Research, 2013.

[60] Thibaut Horel, Stratis Ioannidis, and S. Muthukrishnan. “Budget Feasible Mechanisms for Experimental Design”. In: *LATIN 2014: Theoretical Informatics*. Ed. by Alberto Pardo and Alfredo Viola. Lecture Notes in Computer Science. 2014, pp. 719–730.

[61] Justin Hsu et al. “Differential Privacy: An Economic Method for Choosing Epsilon”. In: *IEEE 27th Computer Security Foundations Symposium*. IEEE, 2014, pp. 398–410.

[62] Justin Hsu et al. “Private Matchings and Allocations”. In: *Proceedings of the 46th Annual ACM Symposium on Theory of Computing*. STOC ’14. 2014, pp. 21–30.

[63] Stratis Ioannidis and Patrick Loiseau. “Linear Regression as a Non-cooperative Game”. In: *Web and Internet Economics*. Ed. by Yiling Chen and Nicole Immorlica. Lecture Notes in Computer Science. 2013, pp. 277–290.

[64] Michael Kearns et al. “Mechanism Design in Large Games: Incentives and Privacy”. In: *Proceedings of the 5th Conference on Innovations in Theoretical Computer Science*. ITCS ’14. ACM. 2014, pp. 403–410.

[65] Erica Klarreich. *Privacy by the Numbers: A New Approach to Safeguarding Data Privacy by the Numbers: A New Approach to Safeguarding Data*. `https://www.scientificamerican.com/article/privacy-by-the-numbers-a-new-approach-to-safeguarding-data/`. [Online; accessed Oct. 21, 2016]. 2012.

[66] Donald Knuth. “Seminumerical algorithms”. In: 2nd ed. Vol. 2. Addison-Wesley Publishing Company, 1981, pp. 130–131.

[67] Nicolas S. Lambert et al. “An Axiomatic Characterization of Wagering Mechanisms”. In: *Journal of Economic Theory* 156 (2015), pp. 389–416.

[68] Nicolas S. Lambert et al. “Self-financed Wagering Mechanisms for Forecasting”. In: *Proceedings of the 9th ACM Conference on Electronic Commerce*. 2008.

[69] Katrina Ligett and Aaron Roth. “Take It or Leave It: Running a Survey when Privacy Comes at a Cost”. In: *Proceedings of the 8th International Conference on Internet and Network Economics*. WINE ’12. 2012, pp. 378–391.

[70] Ashwin Machanavajjhala et al. “Privacy: Theory Meets Practice on the Map”. In: *Proceedings of the 2008 IEEE 24th International Conference on Data Engineering*. ICDE ’08. 2008, pp. 277–286.

[71] Andreu Mas-Colell, Michael D. Whinston, and Jerry R Green. *Microeconomic theory*. Vol. 1. Oxford University Press, 1995.

[72] Frank McSherry. “Privacy integrated queries: an extensible platform for privacy-preserving data analysis”. In: *In Proceeding SIGMOD Conference*. 2009, pp. 19–30.

[73] Frank McSherry and Kunal Talwar. “Mechanism Design via Differential Privacy”. In: *Proceedings of the 48th Annual IEEE Symposium on Foundations of Computer Science*. 2007, pp. 94–103.

[74] Nolan Miller, Paul Resnick, and Richard Zeckhauser. “Eliciting Informative Feedback: The Peer-Prediction Method”. In: *Management Science* 51.9 (2005), pp. 1359–1373.

[75] Roger Myerson. “Optimal auction design”. In: *Mathematics of Operations Research* 6.1 (1981), pp. 58–73.

[76] John Nash. “Non-cooperative games”. In: *Annals of Mathematics* 54.2 (1951), pp. 286–295.

[77] Noam Nisan et al. *Algorithmic game theory*. Cambridge University Press Press, 2007.

[78] Kobbi Nissim, Claudio Orlandi, and Rann Smorodinsky. “Privacy-aware Mechanism Design”. In: *Proceedings of the 13th ACM Conference on Electronic Commerce*. EC ’12. ACM. 2012, pp. 774–789.

[79] Kobbi Nissim, Rann Smorodinsky, and Moshe Tennenholtz. “Approximately optimal mechanism design via differential privacy”. In: *Proceedings of the 2012 Conference on Innovations in Theoretical Computer Science Conference on Innovations in Theoretical Computer Science*. 2012, pp. 203–213.

[80] Kobbi Nissim, Salil Vadhan, and David Xiao. “Redrawing the boundaries on purchasing data from privacy-sensitive individuals”. In: *Proceedings of the 5th Conference on Innovations in Theoretical Computer Science*. ITCS ’14. ACM. 2014, pp. 411–422.

[81] Mallesh M. Pai and Aaron Roth. “Privacy and mechanism design”. In: *ACM SIGecom Exchanges* 12.1 (2013), pp. 8–29.

[82] David M. Pennock et al. “The real power of artificial markets”. In: *Science* 291 (2002), pp. 987–988.

[83] Javier Perote and Juan Perote-Pena. “Strategy-Proof Estimators for Simple Regression”. In: *Mathematical Social Sciences 47*. 2004, pp. 153–176.

[84] Charles Plott and Kay-Yut Chen. “Information Aggregation Mechanisms: Concept, Design and Field Implementation”. California Institute of Technology Social Science Working Paper 1131. 2002.

[85] Philip M. Polgreen, Forrest D. Nelson, and George R. Neumann. “Using Prediction Markets to Forecast Trends in Infectious Diseases”. In: *Clinical Infectious Diseases* 44.2 (2007), pp. 272–279.

[86] Lee Rainie et al. *Anonymity, Privacy and Security Online*. Tech. rep. Pew Research Center, 2013.

[87] Ryan M. Rogers and Aaron Roth. “Asymptotically Truthful Equilibrium Selection in Large Congestion Games”. In: *Proceedings of the 15th ACM Conference on Economics and Computation*. 2014.

[88] Richard Roll. “Orange Juice and Weather”. In: *The American Economic Review* 74.5 (1984), pp. 861–880.

[89] Ariel Rubinstein. *Lecture notes in microeconomic theory: the economic agent*. Princeton University Press, 2012.

[90] Laura Ryan. *Feds Investigate ‘Discrimination by Algorithm’*. Ed. by National Journal. [Online; Retrieved 8/5/2015]. Sept. 2014.

[91] Leonard J. Savage. “Elicitation of Personal Probabilities and Expectations”. In: *Journal of the American Statistical Association* 66.336 (1971), pp. 783–801.

[92] Klaus M Schmidt. “Commitment through incomplete information in a simple repeated bargaining game”. In: *Journal of Economic Theory* 60.1 (1993), pp. 114–139.

[93] Itamar Simonson and Amos Tversky. “Choice in Context: tradeoff contrast and extremeness aversion.” In: *Journal of marketing research* (1992).

[94] Curtis R Taylor. “Consumer privacy and the market for customer information”. In: *RAND Journal of Economics* 35.4 (2004), pp. 631–650.

[95] Richard H. Thaler and William T. Ziemba. “Anomalies: Parimutuel betting markets: Racetracks and lotteries”. In: *J. of Economic Perspectives* 2.2 (1988), pp. 161–174.

[96] Hal R. Varian. “Revealed preference”. In: *Samuelsonian economics and the twenty-first century* (2006), pp. 99–115.

[97] Hal R. Varian. “The Nonparametric Approach to Demand Analysis”. In: *Econometrica* 50.4 (July 1982), pp. 945–974.

[98] Roman Vershynin. “Introduction to the non-asymptotic analysis of random matrices”. In: *Compressed Sensing, theory and applications*. Ed. by Y. Eldar and G. Kutyniok. Cambridge University Press, 2012. Chap. 5, pp. 210–268.

[99] Bo Waggoner, Rafael Frongillo, and Jacob Abernethy. “A Market Framework for Eliciting Private Data”. In: *Advances in Neural Information Processing Systems 28*. 2015.

[100] Stanley L. Warner. “Randomized response: A survey technique for eliminating evasive answer bias”. In: *Journal of American Statistical Association* 60.309 (1965), pp. 63–69.

[101] David Xiao. “Is privacy compatible with truthfulness?” In: *Proceedings of the 4th Conference on Innovations in Theoretical Computer Science*. ITCS ’13. ACM. 2013, pp. 67–86.