STRATEGIC BEHAVIOR AND DATABASE PRIVACY

A Dissertation
Presented to
The Academic Faculty

by

Sara Krehbiel

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in
Algorithms, Combinatorics, and Optimization

School of Computer Science
Georgia Institute of Technology
August 2015

STRATEGIC BEHAVIOR AND DATABASE PRIVACY

Approved by:

Professor Chris Peikert, Advisor
School of Computer Science
Georgia Institute of Technology

Professor Sasha Boldyreva
School of Computer Science
Georgia Institute of Technology

Professor Vijay Vazirani
School of Computer Science
Georgia Institute of Technology

Professor Matt Baker
School of Mathematics
Georgia Institute of Technology

Professor Adam Smith
Computer Science and Engineering Department
Pennsylvania State University

Date Approved: 17 April 2015

To my dad,

Keith Krehbiel.

ACKNOWLEDGEMENTS

I am very grateful to my advisor, Chris Peikert, who has provided me with invaluable mentorship throughout my time at Georgia Tech. Chris guided me through countless major and minor research hurdles, and I always left our meetings with a renewed sense of clarity and confidence.

I am fortunate to have worked with and received valuable feedback from many other talented researchers during my time at Georgia Tech, including Nina Balcan, Georgios Piliouras, Jinwoo Shin, Rikke Bendlin, Ying Xiao, Sasha Boldyreva, Vijay Vazirani, Ruta Mehta, Matt Baker, and Adam Smith.

My friendships spanning my time at Georgia Tech all the way back to elementary school have kept me sane, strong, and happy. Jeff, I feel particularly lucky to have shared these years in Atlanta with you. You have been a rock through challenging and exciting times, and you have made my life so much richer.

Emily, you understand me better than anyone. I'm so lucky that the funniest person I know is also ready with such thoughtful insight and advice whenever I need it. Mom, your love and selflessness is unparalleled. Whenever my resilience has faltered, your unconditional support has always pushed me through. Dad, you have profoundly shaped me as a thinker, a learner, and a teacher. You have always inspired me, and I am honored to have you welcome me to the academy.

TABLE OF CONTENTS

SUMMARY

This dissertation focuses on strategic behavior and database privacy. First, we look at strategic behavior as a tool for distributed computation. We blend the perspectives of game theory and mechanism design in proposals for distributed solutions to the classical set cover optimization problem. We endow agents with natural individual incentives, and we show that centrally broadcasting non-binding advice effectively guides the system to a near-optimal state while keeping the original incentive structure intact.

We next turn to the database privacy setting, in which an analyst wishes to learn something from a database, but the individuals contributing the data want to protect their personal information. The notion of differential privacy allows us to do both by obscuring true answers to statistical queries with a small amount of noise. The ability to conduct a task differentially privately depends on whether the amount of noise required for privacy still permits statistical accuracy.

We show that it is possible to give a satisfying tradeoff between privacy and accuracy for a computational problem called independent component analysis (ICA), which seeks to decompose an observed signal into its underlying independent source variables. We do this by releasing a perturbation of a compact representation of the observed data. This approach allows us to preserve individual privacy while releasing information that can be used to reconstruct the underlying relationship between the observed variables.

In almost all of the differential privacy literature, the privacy requirement must be specified before looking at the data, and the noise added for privacy limits the statistical

utility of the sanitized data. The third part of this dissertation ties together privacy and strategic behavior to answer the question of how to determine an appropriate level of privacy when data contributors prefer more privacy but an analyst prefers more accuracy. The proposed solution to this problem views privacy as a public good and uses market design techniques to collect these preferences and then privately select and enforce a socially efficient level of privacy.

CHAPTER I

INTRODUCTION

This dissertation focuses on a variety of settings involving multiple parties playing different roles with different objectives. Several disciplines have developed in order to study such settings. We begin with a brief overview of the most relevant of these and then introduce the three problems studied.

1.1 Settings

Game theory is the formal study of how people behave when interacting with others who often have different incentives. A game is characterized by a list of parties, a set of possible actions for each party, and a description of each party's utility from the game as a function of all parties' actions. In the classical *prisoners' dilemma*, two criminals are arrested and interrogated in separate rooms. Each individual faces the choice of confessing or staying silent. Their eventual prison sentences depend on who stays silent: both go to prison for 1 year if they both stay silent, and both go to prison for 5 years if they both confess. If only one talks, he serves no time and the other goes to prison for 10 years. Coordinating and staying silent would benefit them collectively. However, confessing is the *best response* for each party individually regardless of whatever the other chooses. Therefore the unique *Nash equilibrium* [74] of the game is for both parties to confess.

In the subfield of *mechanism design*, a game designer chooses a game structure with some target outcome in mind, without necessarily knowing players' incentives. A mechanism is said to be *incentive compatible* if each party maximizes his utility by truthfully revealing his private information. For example, in a *Vickrey (or second price) auction* [97], multiple parties submit secret bids about how much they are

willing to pay for a single item, and the person who bids the most is awarded the item for a price equal to the value of the second highest bid. The highest bidder's utility for the auction (his true value for the item minus the amount he pays) does not change if he bids more than his true value, and it can only decrease if he bids less, because this may cause him to lose the item. A truthful bidder not allocated the item could only win the item by outbidding someone with a higher value, and this would result in him paying more for the item than it is worth to him. Revealing one's true value for the good is therefore a *dominant strategy* (i.e., it maximizes utility, regardless of others' types and behaviors), so Vickrey auctions are incentive compatible.

We also consider settings in which the goal of the system is not to influence the behavior of parties who have private information, but it is to keep that information private while achieving some other goal. *Cryptography* is the study of methods for secure communication in the presence of adversaries. The subfield of *secure multi-party computation* provides protocols allowing multiple parties to compute a function of their data without learning anything beyond the output of the function [103, 35, 12]. The theoretical computer science community has experienced a surge of interest in database privacy following the introduction of the notion of *differential privacy* [27, 25], a central focus of this dissertation. Informally, an analysis is differentially private if changing any single record in the input database has essentially no impact on the output of the analysis. This ensures that the outcome of the analysis cannot compromise the privacy of whoever contributed that record in the database. In contrast to other methods of database privacy, the differential privacy framework provides rigorous privacy guarantees even if an adversary has access to other information that may be arbitrarily related to the database in question.

Contributions

This thesis investigates a variety of questions concerning strategic behavior, privacy, and their intersection. In the remainder of this introductory chapter, we introduce models of dynamic behavior for distributed set cover games [8], differentially private independent component analysis [61], and markets for differential privacy [59]. These are the respective topics of Chapter 2, Chapter 3, and Chapter 4.

An additional component of my doctoral work is presented in [10], in which we propose secure threshold protocols for cryptoschemes based on hard lattice problems. Since its connections to privacy and game theory are minimal, that work is not covered in this dissertation.

1.2 Behavioral Dynamics for Distributed Set Cover

Game theory can be seen as the study of equilibrium outcomes of (fixed) games, and mechanism design is the study of how to structure games to encourage outcomes with specific analytical properties. In [8], we blend the perspectives of game theory and mechanism design, considering the classical *set cover* optimization problem in a distributed setting. We endow agents with natural individual incentives, and we centrally broadcast non-binding advice intended to guide the system to a near-optimal state while keeping the original incentive structure intact.

As a concrete example of a distributed cover game in practice, suppose a state's legislature wants to establish a number of subsidized health clinics. Residents in a county that houses such a clinic will enjoy its benefits, but they will also incur additional local taxes to pay for the clinic. Residents in a county without a clinic do not incur additional taxes, but they only receive the benefits of a clinic if there is one in a neighboring county. The state legislature would like to optimize the net benefit for the state by encouraging a particular set of counties to open clinics. However, since clinics are locally subsidized, counties individually decide whether to open a

clinic, so the legislature cannot centrally dictate a particular distribution of clinics. The agents (counties) in this example have inherent costs associated with being *on* (paying for a local clinic) or *off* (relying on an adjacent county to pay for a clinic), and these incentives are correlated with the social objective, but it is unclear whether unstructured distributed behavior will lead to a good social outcome.

Another application is engineering networks in which non-willful distributed agents are programmed to make decisions based on their surroundings. The extensive literature on cooperative control has shown that in this setting many optimization problems can be conveniently solved in a distributed fashion by endowing agents with artificial individual objective functions and cost-minimizing behavior [90]. Several such papers consider game-theoretic formulations of covering problems that are inspired by practical sensor network problems [89, 66, 87, 16]. The agents are autonomous sensors, and each geographic region corresponds to a set of sensors that could cover that region. A sensor that is *on* is charged some fixed cost, whereas a sensor that is *off* is charged a cost proportional to the number or importance of its adjacent regions that are uncovered by any other sensor. A globally optimal solution may not be known ahead of time, and it may not be possible to dictate it once sensors are placed if a central authority cannot communicate perfectly with all sensors, so we must rely on well-designed distributed behavior.

Equilibrium Quality of Cover Games

Our game generalizes the following simple vertex cover game characterized by a graph on n vertices: a vertex that is *on* experiences cost 1, and a vertex that is *off* experiences cost w for each adjacent edge whose other endpoint is *off*, i.e., each adjacent *uncovered* edge. A (pure) *Nash equilibrium* of this game is a state in which each vertex's choice to be *on* or *off* is its *best response* to the choice of all other vertices, i.e., the action that minimizes that vertex's cost. For $w > 1$, it is easy to verify that every edge will

be covered in any equilibrium. But how do agents reach such an equilibrium? In a game modeling agents that are only locally aware, it is natural to first consider the simple process of *best response dynamics*. In each round of best response dynamics, a single agent ensures he is playing a best response to the other agents by possibly updating his strategy. In a finite *potential game* such as ours, where any single move that reduces a player's individual cost also reduces some global function, best response dynamics converge to a pure Nash equilibrium [71, 77].

However, consider a star graph, in which $n-1$ vertices are each adjacent to exactly one edge, which connects to 1 central vertex. The equilibrium with only the center vertex *on* optimizes social cost, and the other equilibrium with the $n-1$ non-central vertices on is $\Theta(n)$ more costly. This ratio between the costliest equilibrium and the least costly state is called the *price of anarchy* of the game [58, 77]. Best response dynamics are not sufficient for driving agents to a low-cost state in games with high price of anarchy, because such dynamics are only guaranteed to converge to an arbitrary equilibrium. In fact, [16] (whose distributed model is captured by our general set cover game) and many other control theory papers guarantee convergence only to locally optimal stable states, and so these results do not translate to strong global performance guarantees.

Models of Dynamic Behavior

Mechanism design for distributed systems is fundamentally concerned with aligning individual incentives with social welfare to avoid socially inefficient outcomes that can arise from agents acting autonomously. Because agents are only influenced by their neighborhood, we relax the strict mechanism design goal of incentive compatibility and instead model dynamic behavior using game theoretically appealing heuristics such as best response dynamics and ideas from learning theory such as incorporating

advice from an expert. We study the *public service advertising* (PSA) and *learn-then-decide* (LTD) models of Balcan, Blum, and Mansour [6, 7]. The models share the common feature that a central authority knows some state with low social cost, and the authority broadcasts this joint strategy to encourage agents to adopt their prescribed strategies.

Specifically, the PSA model of [6] assumes that each agent independently has some probability of receiving the advertised strategy. Those that receive their prescribed strategies temporarily adopt them; those that do not receive their prescribed strategies behave in a myopic best-response manner. This model is well-suited for an engineering systems setting, where we do not expect all components to receive the central authority's signal. The learning models of [7] assume that each agent uses any of a broad class of learning algorithms to continually choose between acting according to its local best-response move and its broadcasted signal. In the LTD model, agents eventually commit to one of these options. LTD is motivated by a social setting where agents that are only locally aware are interested in exploring the advertising strategy in the hope that it will benefit them personally to follow a central expert's advice.

Cover Game Results

We show that both the PSA and LTD models keep systems out of pathologically high cost cover game equilibria. Furthermore, we give the first theoretical results for the PSA or LTD model that employ particular structural aspects of the advice vector. The following informal theorem statements summarize our results. These results are presented formally (and more generally) in Theorems 2.3.1 and 2.4.1 and in Corollary 2.3.9.

Let $\mathcal{G}$ be any cover game with n elements, constant costs and weights, constant-size sets, and a constant number of sets containing any given pair of elements. We note that LP-rounding yields a joint strategy whose social cost is within a constant factor of

OPT, the cost of the min-cost configuration of $\mathcal{G}$, and recall that the price of anarchy is as high as $\Theta(n)$.

Theorem (PSA and LTD with arbitrary advertised strategies). *For any PSA or LTD advertised joint strategy s^{ad} for game $\mathcal{G}$, PSA and LTD each converge in* $\text{poly}(n)$ *steps to a joint strategy with expected cost $O(\texttt{cost}(s^{ad})^2)$.*

Theorem (PSA with specially-designed advertised strategies). *There exists a* $\text{poly}(n)$ *algorithm for constructing a PSA advertised strategy s^{ad} of a particular form for game $\mathcal{G}$ such that except with probability $1/n$, PSA converges in* $\text{poly}(n)$ *steps to a joint strategy with cost $O(\log n) \cdot OPT$.*

The PSA and LTD models share three features that jointly help us give these positive results for covering games: 1) advertising seeds the system with a preference for globally efficient behavior, 2) best-response dynamics harnesses the fact that individual and social welfare is aligned and permits potential arguments, and 3) the randomness that dictates which agents receive signals and update orders allows for expected or high probability cost arguments when straightforward structural arguments are not possible.

1.3 Private Independent Component Analysis

Database privacy has attracted a surge of interest from the theoretical computer science community following the introduction of the notion of *differential privacy* [27, 25]. Differential privacy characterizes a privacy property for a mechanism $\mathcal{M}$ that operates on n data records of type $\mathcal{D}$:

Definition. *For $\epsilon \geq 0$, a mechanism $\mathcal{M} : \mathcal{D}^n \to Range(\mathcal{M})$ is* ϵ-differentially private *if for all* neighboring *databases $X, Y \in \mathcal{D}^n$ and for all subsets $S \subseteq Range(\mathcal{M})$,*

$$\Pr[\mathcal{M}(X) \in S] \leq \exp(\epsilon) \cdot \Pr[\mathcal{M}(Y) \in S].$$

We will see later that a restricted *neighborhood* often helps us achieve privacy, but the original and most general notion of neighboring databases are those that differ (arbitrarily) on only one record. This ensures that the outcome of the analysis cannot compromise the privacy of whomever contributed that entry to the database. In contrast to other methods of database privacy, the differential privacy framework provides rigorous privacy guarantees even if an adversary has access to other information that may be arbitrarily related to the database in question. There is a large body of literature on differentially private mechanisms that return accurate answers for many classes of statistical queries relevant to our work, including counting queries and histograms [25], arbitrary query classes with low VC dimension [14], and principal component analysis [106, 19].

Principal component analysis (PCA) transforms multi-dimensional data into a new coordinate system, in which coordinates are ordered by how much of the variance of the data they respectively account for. This process is accomplished by normalizing the columns of the data so they are mean-zero, computing the covariance of the normalized data, and then orthonormalizing the eigenvectors of the covariance matrix. Recent work shows that differentially private PCA is only possible for *incoherent* databases, whose weight is distributed roughly evenly across their rows [40].

Our work in [61] focuses on an extension of PCA called independent component analysis, which belongs to the field of *parametric statistics*. Methods in parametric statistics assume data comes from some distribution, and the statistical goal is to estimate parameters of that distribution. Despite the importance of parametric statistics both in theory and practice, the majority of research in differential privacy has focused on computing simple summary statistics or other nonparametric methods, and relatively little is known about fitting models while preserving differential privacy. Two notable exceptions are [92], which tackles private parametric inference by releasing maximum likelihood estimators, and [26], which proposes the Propose-Test-Release

framework for differentially private robust estimators. We will show how we use this framework for estimators whose robustness relies on subtle database properties.

Independent Component Analysis

Independent component analysis (ICA) is an important non-linear computational method in statistics that generalizes PCA. The classical example application of ICA is known as the *cocktail party problem*, in which several microphones capture audio from a mixture of simultaneous conversations, and the goal is to recover the speech signal for each distinct conversation. This task is known more generally as channel separation, deconvolution, or blind-source separation.

The basic ICA model assumes p independent real random *source variables* $s = (s_1, \ldots, s_p) \in \mathbb{R}^p$, each with some (non-Gaussian and possibly unknown) distribution over $\mathbb{R}$. We do not observe these variables directly, rather we observe them through p *signal variables* $x = (x_1, \ldots, x_p)$, which are linear combinations of the source variables s under a fixed nonsingular *mixing matrix* $A \in \mathbb{R}^{p \times p}$:

$$x = sA.$$

The goal of ICA is to recover (up to trivial equivalencies) the mixing matrix A from the distribution of x or some noisy approximation thereof, which then also reveals the source variables s. Jutten and Herault [48] were the first to formalize this problem, though they point out that variants had appeared earlier in various fields (the earliest such mention is in [9], according to the survey [49]).

ICA has found many applications in a number of fields with privacy concerns; see [44] for an overview of the more traditional applications. In particular, [54] uses ICA to find independent factors that affect the cash-flow of different stores in the same retail chain using cash-flow data from each store, and this is extended to a predictive econometric perspective in [67]. Many studies have found ICA to be a useful tool for extracting information from EEG, MEG, and fMRI data and for medical

diagnosis more generally [100, 98, 99, 47, 69, 13, 84, 53, 102]. ICA has also been used to effectively reduce layer size in deep neural networks [11, 63, 62, 75].

The above ICA applications have a natural privacy aspect, in that the observations x and underlying source variables s may reflect confidential information about individuals or organizations, whereas the mixing matrix represents an underlying (and non-confidential) structure that the analysis is attempting to discover. For example, the source signals and observations could respectively represent mutations in a genome and incidents of diseases, and the mixing matrix would capture how the former influences the latter. It is tempting to assume that publicly revealing *only* the computed mixing matrix should preserve the privacy of the observations. However, this argument is only heuristic and relies on assumptions that may not hold in reality, e.g., that the observations are independent and distributed exactly according to an ICA model, and that the chosen ICA algorithm correctly recovers A to high precision and without too much dependence on individual observations. Otherwise, individual data could leak into the computed A in unexpected and hard-to-characterize ways.

Solution Approach

All known ICA algorithms work in two main phases: a whitening step first calculates the mean and covariance of the data and applies a corresponding affine transformation to normalize the data, and an optimizing phase then recovers the columns of the orthonormal mixing matrix. Our approach starts with the observation that for many popular ICA algorithms (especially those with provable guarantees), the optimization phase does not require the data itself, but only its *fourth moment tensor*, i.e., the values $\mathrm{E}[x_i x_j x_k x_l]$ for all $i, j, k, l \in \{1, \ldots, p\}$. By independence of the source variables, the fourth moment has local optima exactly at the columns of A^{-1}, and a variety of methods (including gradient descent and fixed-point methods) may be used to find these optima. After recovering the orthonormal A^{-1}, an ICA algorithm finally reverses

the original affine transformation (using the mean and covariance computed during whitening) to obtain the original mixing matrix and source variables.

We therefore observe that for differentially private ICA, it suffices to release a sanitized (private) version of the appropriately normalized first, second, and fourth moment tensors, and then an analyst may run any of the above optimization algorithms on these private moment tensors. With this approach, determining how much noise to add for privacy reduces to analyzing the sensitivity of these tensors under a single row change. As noted in [40], the sensitivity of the covariance is tightly related to the *incoherence* of the data. One of our main technical contributions is that essentially the same is true of the fourth moment tensor, with an additional dependence on the *conditioning* of the covariance of the data, i.e., the ratio of its maximum and minimum singular values.

Because of the dependence of the fourth moment tensor's sensitivity on incoherence and conditioning, privacy is given with respect to a definition of neighboring databases that bounds the amount a single row change can affect the incoherence and conditioning of the database. Although not as general as a neighborhood definition permitting an arbitrary row change, our neighborhood definition is much more permissive than definitions that require the row change to be of constant Frobenius norm, in particular as required in [40]. This neighborhood definition allows us to employ the Propose-Test-Release paradigm of [26]: it (privately) checks whether the input database is adequately conditioned, and if so, it releases noisy moment tensors. Otherwise, it aborts with no output.

Demonstrating ICA Utility

In arguing the utility of our mechanism for ICA, we face two main conceptual challenges. First, for arbitrary input data, which may not be well described by any ICA model, there is no canonical notion of the most accurate mixing matrix, nor one

single objective function measuring how well a candidate mixing matrix fits the data. Second, existing ICA algorithms usually lack formal guarantees of output accuracy (i.e., how close their output comes to optimizing the objective), except sometimes in highly idealized settings where the data exactly conform to an ICA model. It is therefore unclear how to meaningfully quantify the accuracy of differentially private ICA, especially with arbitrary data.

One of our main contributions is a way of quantifying ICA accuracy in an algorithm-agnostic way, and even in the absence of output guarantees in the non-private setting. Instead of analyzing the *output* of a particular ICA algorithm on its sanitized input, we consider the *objective function* that the algorithm attempts to optimize. We show that if the original data conform closely to some ICA *reference model*, then the optima of the sanitized fourth moment tensor are close to the columns of the model's unmixing matrix. Our differentially private mechanism and notion of utility can be directly applied to many of the recent ICA algorithms with provable guarantees [32, 76, 5, 2, 3, 36]. With this approach, utility loss depends on the *spectral* norm perturbation of our fourth moment tensors while privacy requires sufficient *Frobenius* norm perturbation. We show that our noise tensor's spectral norm grows with $\sqrt{p}$ while its Frobenius norm grows with p^2, allowing us to provide surprisingly good privacy-utility tradeoffs.

1.4 Markets for Database Privacy

There is an inherent tradeoff between guaranteeing privacy and preserving the statistical utility of a database, and this tradeoff is central to the problem studied in [59]. Classically, differentially private mechanisms take the privacy parameter ϵ as an input and provide a corresponding privacy guarantee by adding random noise at various stages in the sanitization process. More noise enhances privacy but decreases accuracy. Different applications intuitively require different levels of privacy; medical records,

for example, may require a higher standard of privacy than Netflix preferences.

Despite this, the literature is largely agnostic to the choice of this privacy parameter. It is almost always assumed to be exogenously given, but it is unclear who sets ϵ and how. In many cases, the government enforces certain privacy standards, but legislators are often ill-informed of the theoretical or practical consequences of privacy policy. A trusted database curator might set ϵ (before collecting data) according to expert discretion about the relative needs for privacy versus meaningful data analysis, but he may not have good information about the privacy preferences of the individuals contributing their private data or the value of accuracy for an analyst. Indeed, we would expect data contributors and analysts to inflate their stated respective needs for privacy and accuracy if there is no downside to doing so. The work in [59] is motivated by two questions:

1. How can we extend rigorous privacy analysis to a setting in which a mechanism operates at some level of privacy that the mechanism chooses *endogenously* as a function of its inputs?

2. How can a privacy-preserving mechanism use the privacy and utility preferences of the parties involved (the data contributors and analyst) to set a value of ϵ that achieves a desirable privacy/utility tradeoff?

An arbitrary value of ϵ may be suboptimal in that an analyst may be willing to pay data contributors to accept a weaker privacy guarantee in order to improve the accuracy of a statistical analysis, or data contributors may be willing to pay more for stronger privacy. To avoid this type of economic inefficiency, this work considers tools and solution concepts from economics to develop a privacy mechanism that finds an equilibrium level of privacy by making monetary transfers simulating a market.

This approach raises a third question: When designing a mechanism that sets its own level of ϵ according to market forces driven by privacy and accuracy preferences,

should data contributors be paid for use of their data subject to some privacy guarantee, or should they pay an analyst for a guarantee of privacy? Many of the related works consider settings in which companies or researchers want access to a body of data. An individual truly *owns* data about him[1], so those who want access should compensate him for his resulting loss of privacy. For example, [65] develop an opt-in survey protocol where an analyst must pay participants in order to run a differentially private study, motivated by the need to entice volunteers for medical studies.

However, if we accept the general idea of commoditizing privacy by letting others buy access to private data about us, then a rigorous privacy guarantee should be seen as a valuable feature of a product or service that *already* collects our data with consent. For example, Netflix may have standard (but imperfect! [73]) privacy agreements with its users, but it could offer a premium service with a strong and rigorous guarantee of differential privacy for data collected from its users. Netflix could then privately release user data in attempt to improve their service, and users could be comfortable knowing that their entire viewing histories are not likely to become public knowledge.

Prior Works

Several recent works have studied mechanisms that consider the incentives of data contributors who value privacy [101, 79, 78, 20, 65, 30, 80]. In all of these mechanisms, ϵ must be chosen exogenously, before looking at the data contributors' private data or privacy valuations. In contrast, Ghosh and Roth [34] propose a model, called the *insensitive value model*, and present two mechanisms in this model that 1) solicit privacy and accuracy preferences from data contributors and an analyst, 2) determine an appropriate value of ϵ by organizing data contributors by privacy preferences and (roughly) conducting a second price auction, 3) charge the analyst a payment in exchange for a noisy statistic on the data, and 4) distribute this payment among

[1]Many works refer to these individuals as *data owners*; our choice of the term *data contributor* is deliberately ownership-agnostic.

data contributors to compensate for their ϵ loss of privacy. In an alternate model of [33], an analyst proposes a differentially private computation designed so the privacy parameter ϵ decreases with the number of data contributors who voluntarily opt in.

A weakness of the insensitive value model of [34] and the opt-in model of [33] is that while differential privacy is guaranteed at any level output by the mechanism, privacy is with respect to the private data only. This means that if individuals' data are correlated with their privacy preferences or participation decisions, the mechanism may indirectly leak information about private data. To address this concern, [34] propose a stronger *sensitive value model* that requires differential privacy with respect to privacy valuations as well as private data. However, they also give a negative result that seems to indicate that the sensitive value model admits no mechanisms that satisfy all desired properties.

Contributions

Chapter 4 presents three contributions to the existing literature on mechanisms that seek to internally determine a meaningful value of ϵ:

1) Stronger negative results. We extend the negative result of Ghosh and Roth [34] in the sensitive value model and extend it by relaxing assumptions. In particular, their result holds when we relax the cost function assumptions atnd accuracy requirements. Significantly, only economically trivial mechanisms are possible even in the setting that data contributors have positive value for a privacy guarantee and must pay the analyst for such a guarantee. However, we will see that these results are due to a worst-case, non-endogenous differential privacy requirement, not the requirement that privacy preferences be kept private, and this motivates us to take a closer look at the privacy requirements for mechanisms that choose privacy as a function of inputs.

2) Formalization of endogenous privacy. The standard definition of differential privacy gives a guarantee for the mechanism on *any* pair of databases parametrized by a *single, data-independent* value of ϵ. One could show that a mechanism that chooses ϵ internally as a function of the input data is $\epsilon_\circ$-differentially private for some fixed $\epsilon_\circ$, but the definition provides no way to connect the mechanism's chosen, *input-dependent* level of ϵ to a privacy guarantee. This incompatibility between the existing definition (see Section 1.3) and the new setting is rectified with a new definition of *endogenous differential privacy*:

Definition. *A mechanism* $\mathcal{M} : \mathcal{D}^n \to \mathit{Range}(\mathcal{M})$ *is* endogenously differentially private *if for all neighboring databases* $X, Y \in \mathcal{D}^n$*, for all* ϵ *in the privacy support of* $\mathcal{M}$ *on* X*, and for all* $S \subseteq \mathit{Range}(\mathcal{M})$*,*

$$\Pr[\mathcal{M}(X) \in S] \leq \exp(\epsilon) \cdot \Pr[\mathcal{M}(Y) \in S].$$

3) A general-purpose privacy market. Using the endogenous privacy relaxation, Section 4.4 prevents the first privacy markets that 1) elicit truthful privacy preferences from the data contributors, 2) use these preferences to set ϵ endogenously, and 3) give privacy guarantees for the privacy preferences as well as the data itself. These mechanisms model privacy as a *public good* and employ tools from classical markets for public goods [97, 22, 37, 38], adding carefully calibrated random noise for endogenous differential privacy.

These markets for privacy as a public good raise an important open question: Can we develop meaningful markets in which analysts must buy differentially private data access, compensating data contributors for their privacy loss, as assumed in the models of [34]? The new negative results suggest that the new endogenous privacy definition may permit markets in which money flows from analyst to data contributors. However, it is a non-trivial challenge in this case to model realistic preferences that can be truthfully elicited.

Other Work

After discussing these new results in behavioral dynamics with mechanism design [8], differential privacy [61], and mechanism design as a tool for solving a privacy problem [59], we conclude in Chapter 5 with a discussion of other game theory problems for which privacy may be a useful feature. We present some preliminary results for computing approximate equilibria of large anonymous games in a way that hides individual players' behavior in the context of related work. We also propose new questions about when direct revelation mechanisms can have meaningful approximate versions that preserve the privacy of players' types.

CHAPTER II

A GAME OF DISTRIBUTED SET COVERING

2.1 Overview

In [8], we model set cover problems as games, and we use models from learning theory to describe local decision making by agents in these games. We are interested in demonstrating convergence not to arbitrary local equilibria but to states whose cost is competitive with the global optimum. We accomplish this by incorporating a globally-informed central authority into natural models of dynamic behavior.

This chapter is organized as follows. In Section 2.2, we cover the necessary game theory and behavioral dynamics definitions, we give the formal model of the set cover games we study, and we introduce other notation used throughout. Set cover is of particular interest in control theory for its natural applications to sensor networks. The centralized optimization problem is NP-hard. The problem admits good approximations, but these approximations do not necessarily represent equilibria of an underlying distributed game, in which elements are modeled as independent, cost-minimizing agents.

In Section 2.3, we introduce the public service advertising (PSA) model of [6]. The PSA model begins with agents playing arbitrary strategies, and a central authority broadcasts a strategy for each agent. Some of these agents receive and follow this strategy temporarily while others converge to a steady state by a series of myopic, local optimization steps known as *best-response dynamics*. Eventually, even those agents who initially followed their advertised strategies switch to best-responding and the system converges to a new equilibrium. Theorem 2.3.1 gives our results for advertising effectiveness in PSA, roughly stating that for any advertised strategy, the

expected cost of the final state at the end of PSA is the square of the cost of the advertised strategy. Our proof associates the costs of the outcome of PSA with the cost of the advertised strategy by leveraging the simplicity of best-response dynamics to charge each component of the final cost to some component of the advertised cost. The subtlest part of the analysis uses the fact that agents receive the signal with independent probability, and this probabilistic reasoning forces an expected cost bound. The bottleneck reflected in our final bound, however, is due to a crude use of a structural assumption that some parameter Δ bounds the number of sets containing a given pair of elements, and in particular, we use Δ to bound the number of sets containing at least two elements that are *on* in the advertised strategy.

We then show how to construct an advice vector specifically designed for the PSA model. These advice vectors are not only low-cost but efficient in that each *on* element uniquely covers many sets. Because it is very likely that some of the sets will have *each* element receptive to their advertised strategies in the first advertising phase of PSA, the efficiency property of our advertised strategy will cause all sets to be covered by the end of the phase. Indeed, Theorem 2.3.8 states that for such an advertised strategy, the cost at the end of PSA is of the same order as the cost of the advertised strategy with high probability. This substantially improves our results for general advice vectors, and it constitutes the first result for the PSA model that uses the *structure* of a carefully chosen advice vector to prove high probability and near-optimal outcomes.

In Section 2.4 we introduce the related learn-then-decide (LTD) model of [7]. In contrast to PSA, agents *explore* in the first phase, independently choosing during each update round to either best-respond or follow their advertised strategy. Agents *exploit* one of these options in the second phase, arbitrarily committing to consistently play either their advertised strategy or their best response. Theorem 2.4.1 relates the cost of the outcome of LTD to that of an arbitrary advertised strategy. In addition to the

techniques used for the analogous Theorem 2.3.1 in the PSA model, we have to use additional techniques relying on the agents' random update order to prove that agents switching to their advertised strategy in the exploit phase does not increase cost too much.

Our work is in the context of several other works have considered games related to our set cover game [89, 66, 87, 16, 15, 28, 17, 18, 4, 43, 42], which give results that are strictly weaker or incomparable to our results. Other approaches to avoiding high-cost equilibria [91, 52, 51, 55, 56, 31, 57, 82, 45, 72, 64] either result in weak global performance guarantees or central authorities that are much stronger than those of the PSA and LTD models. We note that the models of [24, 29] do not accommodate our game.

2.2 Preliminaries

2.2.1 Game Theory Definitions

We represent a general game as a triple $\mathcal{G} = \langle N, (S_i), (\texttt{cost}_i)\rangle$, where N is a set of n agents, S_i is the (finite) action space of agent $i \in N$, and $\texttt{cost}_i$ denotes the cost function of agent i. The joint action space of the agents is $S = S_1 \times \cdots \times S_n$. For a joint action $s \in S$, we denote by s_{-i} the actions of all agents $j \neq i$. Agents' cost functions map joint actions to non-negative real numbers, i.e. $\texttt{cost}_i : S \to \mathbb{R}^+$ for all $i \in N$. In this chapter, we define a *social cost function*, $\texttt{cost} : S \to \mathbb{R}$, simply as the sum of individual agents' costs. The optimal social cost is denoted

$$OPT := \min_{s \in S} \texttt{cost}(s).$$

The *best-response set* of agent i to any joint action s_{-i} of all other agents is the set of actions that minimize i's cost, i.e.

$$BR_i(s_{-i}) := \{\arg\min_{s_i \in S_i} \texttt{cost}_i(s_i, s_{-i})\}.$$

Best-response dynamics is a process in which at each time step, an arbitrary[1] agent not already playing a best-response move updates his action to one in his current best-response set. A joint action $s \in S$ is a *pure Nash equilibrium* if $s_i \in BR_i(s_{-i})$ for every $i \in N$. Let $\mathcal{N}(\mathcal{G})$ denote the pure Nash equilibria of game $\mathcal{G}$.

A game $\mathcal{G}$ is called an *exact potential game* [71] if there exists a potential function $\Phi : S \to \mathbb{R}$ such that for all $i \in N$, $s_{-i} \in S_{-i}$, and $s_i, s_i' \in S_i$,

$$\texttt{cost}_i(s_i', s_{-i}) - \texttt{cost}_i(s_i, s_{-i}) = \Phi(s_i', s_{-i}) - \Phi(s_i, s_{-i}).$$

For general potential games, only the signs of both sides of these equations must be equal. While not all games have a pure Nash equilibrium, all finite potential games do, and best-response dynamics in such games always converges to a pure Nash equilibrium [71, 77]. However, the convergence time can be exponentially large in terms of the number of agents in general.

Two well known concepts for quantifying the inefficiency of equilibria relative to non-equilibria are *price of anarchy* and *price of stability*, defined respectively as

$$\text{PoA} := \max_{s \in \mathcal{N}(\mathcal{G})} \frac{\texttt{cost}(s)}{OPT} \qquad \text{PoS} := \min_{s \in \mathcal{N}(\mathcal{G})} \frac{\texttt{cost}(s)}{OPT}.$$

2.2.2 Set Cover Games

A cover game $\mathcal{G} = \langle [n], (S_i), (\texttt{cost}_i) \rangle$ is specified by a collection of sets $\mathcal{F} \subseteq 2^{[n]}$, costs c_i for $i \in [n]$, and weights w_σ for $\sigma \in \mathcal{F}$. Each agent has action space $S_i = \{on,\ off\}$. A joint strategy $s \in S$ induces a bipartition of agents that are *on* and *off*, respectively. Dropping the s when clear from context, we let

$$L(s) := \{i \in [n] \ : \ s_i = on\}, \quad R(s) := \{i \in [n] \ : \ s_i = off\}, \quad \text{and}$$

$$\mathcal{F}_R(s) := \{\sigma \in \mathcal{F} \ : \ \sigma \subseteq R(s)\}.$$

[1] We will focus on best-response dynamics with *uniformly random* update order.

We call sets in $\mathcal{F}_R$ *uncovered*. Figure 1 in Section 2.3 graphically illustrates this and other notation introduced later. An agent pays either his cost of being *on* or the weights of all uncovered sets he participates in:

$$\texttt{cost}_i(s) := \begin{cases} c_i & \text{if } s_i = \mathit{on} \\ \displaystyle\sum_{\sigma \in \mathcal{F}_R \,:\, i \in \sigma} w_\sigma & \text{if } s_i = \mathit{off}. \end{cases} \tag{2.2.1}$$

For $\sigma \subseteq [n], \mathcal{F}' \subseteq \mathcal{F}$, we define for shorthand $c(\sigma) := \sum_{i \in \sigma} c_i$ and $w(\mathcal{F}') := \sum_{\sigma \in \mathcal{F}'} w_\sigma$. Then the social cost has the following simple form:

$$\texttt{cost}(s) := \sum_{i \in [n]} \texttt{cost}_i(s) = c(L) + \sum_{\sigma \in \mathcal{F}_R} |\sigma| \cdot w_\sigma. \tag{2.2.2}$$

Our results are given in terms of some additional game parameters. Denote

$$c_{\max} := \max_{i \in [n]} c_i, \quad c_{\min} := \min_{i \in [n]} c_i, \quad w_{\max} := \max_{\sigma \in \mathcal{F}} w_\sigma, \quad w_{\min} := \min_{\sigma \in \mathcal{F}} w_\sigma.$$

For expository simplicity, we consider costs and weights which are bounded above and below by constants, i.e., $c_{\max},\ c_{\min},\ w_{\max},\ w_{\min} = \Theta(1)$, although we can push these quantities through the analysis to give results for general costs and weights, as shown in Claim 2.3.6. We also define

$$F_{\max} := \max_{\sigma \in \mathcal{F}} |\sigma|.$$

Note that when $F_{\max} = 2$, the game can be specified by a simple graph with vertex costs and edge weights, where an *on* vertex covers its incident edges. Our results when $F_{\max} = 2$ are stronger than in the general case (see Theorems 2.3.1 and 2.4.1). For $F_{\max} > 2$, a given pair of elements may appear in multiple sets. Our results depend on the maximum number of sets containing any given pair of elements, so we define:

$$\Delta := \max_{i,j \in [n], i \neq j} \left| \{\sigma \in \mathcal{F} \ :\ i, j \in \sigma\} \right|.$$

It is sometimes useful to consider sets covered by a unique element. For joint strategy s and $\ell \in L(s)$, let

$$\mathcal{F}^*_\ell := \{\sigma \in \mathcal{F} \ :\ \ell \in \sigma,\ \sigma \backslash \{\ell\} \subseteq R\}.$$

A joint strategy in which every *on* element uniquely covers many sets has a high *core minimum*:

$$\delta^* := \min_{\ell \in L} |\mathcal{F}^*_\ell|.$$

Section 2.3.1 shows how to construct joint strategies with high core minimum, and we show that advertising such joint strategies in PSA yields especially strong guarantees.

In this work, we primarily focus on the case when $F_{\max} = O(1)$. We note that this holds in many practical applications of interest. In current wireless sensor technology, for example, the maximum sensing range is around a hundred meters [105], while the size of sensors has a lower bound. Hence, the number of sensors that can cover a given geographical area is bounded above. Furthermore, a good sensor network should have low overlap in sensing areas, and even $F_{\max} = 2$ can be achieved by carefully designing locations of sensors [104].

Packing Interpretation Observe that c_i expresses how costly it is for agent i to cover the sets that contain him. For example, if $c_{\max} < w_{\min}$, it will always be cheaper for an agent to be *on* than to participate in any uncovered sets, so every set will be covered in equilibrium. The socially optimal equilibria are necessarily the minimum cost set covers. When $F_{\max} = 2$, $c_i = c$ for all i, and $c < w_{\min}$, the equilibria and socially optimal equilibria are the minimal and minimum vertex covers, respectively.

We note that if we simply redefine the costs so that i pays c_i if he is *off* and pays the sum of the weights of the *fully-covered* sets he participates in if he is *on*, this game is a packing analog of the original cover game. All of our results for PSA and LTD apply to such packing games by simply replacing *on* with *off* and vice versa in all of our results and proofs. Note that the equilibria when $c_{\max} < w_{\min}$ are configurations in which no set is fully covered. When additionally $F_{\max} = 2$ and $c_i = c$ for all i, the set of *on* agents in any equilibrium is a maximal independent set, and the lowest-cost equilibria are the maximum independent sets.

2.2.3 Cost Bounds in Cover Games

We bound the cost increase due to best-response dynamics in a cover game as follows:

Fact 2.2.1. *For arbitrary joint strategy s that evolves to s' via best-response dynamics, we have:*

$$\texttt{cost}(s') \leq F_{\max} \cdot \texttt{cost}(s).$$

Proof. Observe that the cover game is an exact potential game with potential function

$$\Phi(s) := c(L) + w(\mathcal{F}_R). \tag{2.2.3}$$

Combining this potential function with the social cost formula, we have $\Phi(s) \leq \texttt{cost}(s) \leq F_{\max} \cdot \Phi(s)$ for any joint strategy s. Potential never increases in best-response dynamics, so $\Phi(s') \leq \Phi(s)$. Then we have $\texttt{cost}(s') \leq F_{\max} \cdot \Phi(s') \leq F_{\max} \cdot \Phi(s) \leq F_{\max} \cdot \texttt{cost}(s)$. □

Although best-response dynamics converge to a pure Nash equilibrium in finite potential games, the star graph example from the introduction reveals that our class of cover games has a price of anarchy of $\Omega(n)$, motivating the need for efficient dynamics with better guarantees than convergence to arbitrary equilibria.

As a step in that direction, we observe that a centralized, poly-time LP-rounding algorithm can find a low-cost configuration s^{ad} for the cover game. Specifically, let

$$x^* := \arg\min_x \{\sum_{i=1}^{n} c_i \cdot x_i \quad \texttt{s.t.} \quad \sum_{i \in \sigma} x_i \geq 1 \ \forall \sigma \in \mathcal{F},\ x_i \in [0,1]\},$$

and then for all $i \in [n]$, set s_i^{ad} to *on* if $x_i^* \geq 1/F_{\max}$ and *off* otherwise.

Fact 2.2.2. *The configuration s^{ad} obtained from LP-rounding has*

$$\texttt{cost}(s^{ad}) \ \leq \ F_{\max} \lceil c_{\max}/w_{\min} \rceil \cdot OPT.$$

Proof. Let s^* be some joint strategy that achieves optimal social cost, and let s' be the joint strategy obtained by turning *on* an arbitrary element in each set σ that is uncovered in s^*. Then:

$$\begin{aligned}\texttt{cost}(s^{ad}) \;&\leq\; F_{\max} \cdot \sum_i c_i \cdot x_i^* \;\leq\; F_{\max} \cdot \texttt{cost}(s') \\ &\leq\; F_{\max}\lceil c_{\max}/w_{\min}\rceil \cdot \texttt{cost}(s^*) \;=\; F_{\max}\lceil c_{\max}/w_{\min}\rceil \cdot OPT. \qquad \square\end{aligned}$$

Of course, when considering games with constant $F_{\max}, c_{\max}, w_{\min}$, this LP-rounding procedure provides us with advertising strategies with cost $O(OPT)$.

2.3 Set Cover in the Public Service Advertising Model

The first model we study in this work is the public service advertising (PSA) model in [6] in which a central authority broadcasts a strategy for each agent, which some agents receive and temporarily follow. Agent behavior is described in two phases:

1: Play begins in an arbitrary state, and a central authority advertises joint action $s^{ad} \in S$. Each agent receives the proposed strategy independently with probability $\alpha \in (0, 1)$. Agents that receive this signal are called receptive. They play their advertising strategies throughout Phase 1, and non-receptive agents undergo best-response dynamics to settle on a joint strategy that is a Nash equilibrium given the fixed behavior of receptive agents. We call this joint strategy s'.

2: All agents participate in best-response dynamics until convergence to some Nash equilibrium s''.

Since our cover game is a finate potential game and all such games eventually converge to a Nash equilibrium under best-response dynamics, both phases are guaranteed to terminate. Furthermore, since potential decreases under best-response dynamics and is bounded above and below by functions of n, costs, and weights (recall Equation (2.2.3)),

polynomial-time convergence of PSA is guaranteed for any game with poly(n) costs and weights.

Notation. In this and the following section, we let $L = L(s^{ad})$ and $R = R(s^{ad})$. We let L_{on} denote the set of elements that are *on* both in s^{ad} and in s' (at the end of Phase 1), and L_{off} denotes the elements that are *off* in s' but not in s^{ad}. Analogously, R_{off} denotes the elements that are *off* both in s^{ad} and s', and R_{on} denotes the elements that are *on* in s' but not in s^{ad}. We let $\mathcal{F}_{bad}$ denote the sets covered in s^{ad} but uncovered in s'. Note that every set in $\mathcal{F}_{bad}$ contains an element in L.

Figure 1 provides a notational example when the PSA inputs are a 4-vertex star graph with $c_i = w_\sigma = 1$ for all i, σ and the graph's unique optimal advice vector s^{ad}. Note that $\mathcal{F}_R$ is empty and L contains the single center vertex for the optimal s^{ad}, which has cost 1. The figure depicts some s' with cost 4 at the end of Phase 1. The vertices in L_{off} and R_{on} are Phase 1 best-responders, and they are in equilibrium in s' given the fixed behavior of the R_{off} agent following s^{ad}. This partial equilibrium leaves one edge uncovered; the dashed line represents this set in $\mathcal{F}_{bad}$.

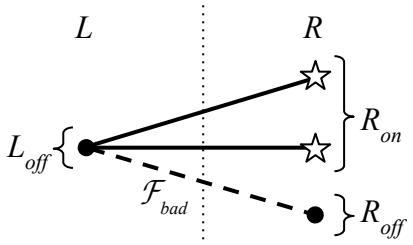

Figure 1: Example of notation

Overview of PSA results. Theorem 2.3.1 formalizes the general result of this section, relating the expected cost of the PSA outcome in vertex cover ($F_{\max} = 2$) and set cover games to that of an arbitrary advertised strategy. We assume $\Theta(1)$ costs and weights for expository simplicity, although we present the general costs and weights case without proof in Claim 2.3.6. At a high level, the proof of Theorem 2.3.1

associates the costs of the outcome of PSA with the cost of the advertised strategy by charging each component of the final cost to some component of the advertised cost, harnessing the simplicity of best-response dynamics. The subtlest part of the analysis uses the fact that agents receive the signal with independent probability, and this probabilistic reasoning forces an expected cost bound. However, the bottleneck reflected in our final bound for $F_{\max} = O(1)$ is due to a crude use of our structural assumption that at most Δ sets contain a given pair of elements. Using advertised strategies obtained from LP-rounding, Corollary 2.3.2 puts these results in terms of the lowest cost configuration of any particular set cover game. In Section 2.3.1, we strengthen these results using advertised strategies specifically constructed for the PSA model.

Theorem 2.3.1. *For a cover game with constant costs and weights, and for any advertising strategy s^{ad}, the expected cost at the end of PSA is*

$$\mathrm{E}[\texttt{cost}(s'')] = \begin{cases} O(1) \cdot \texttt{cost}(s^{ad}) & \text{if } F_{\max} = 2 \\ O(\Delta) \cdot \texttt{cost}(s^{ad})^2 & \text{if } F_{\max} = O(1) \\ O\left(\dfrac{\Delta F_{\max}^2}{\alpha^{2F_{\max}}}\right) \cdot \texttt{cost}(s^{ad})^2 & \text{otherwise.} \end{cases}$$

If s^{ad} is obtained from the LP-rounding $O(F_{\max})$-approximation algorithm described in Section 2.2.3, the following corollary is immediate from the proof of Theorem 2.3.1:

Corollary 2.3.2. *For a cover game with constant costs and weights, there exists a poly-time algorithm to find an advertising strategy s^{ad} such that the expected cost at the end of PSA is*

$$\mathrm{E}[\texttt{cost}(s'')] = \begin{cases} O(1) \cdot OPT & \text{if } F_{\max} = 2 \\ O(\Delta) \cdot OPT^2 & \text{if } F_{\max} = O(1) \\ O(\Delta F_{\max}^3/\alpha^{2F_{\max}}) \cdot OPT^2 & \text{otherwise.} \end{cases}$$

Before proving Theorem 2.3.1, we make some observations guiding the structure of the proof. We first note that since Phase 2 is simple best-response dynamics, Fact 2.2.1 assures that the cost of s'', the final equilibrium, is at most a factor $F_{\max}$ greater than the cost of s', the state at the end of Phase 1. Therefore we focus on bounding the expected cost of s' relative to that of the advertised strategy s^{ad}. The only social cost at the end of Phase 1 beyond that of the advertised strategy is due to the weight of sets that are uncovered in s' but covered in s^{ad} and the cost of elements that are *on* in s' but not in s^{ad}. These terms are bounded in Lemmas 2.3.3 and 2.3.4, respectively.

Proof of Theorem 2.3.1. Using the new notation, we can bound the expected cost at the end of Phase 1 as:

$$\mathrm{E}[\texttt{cost}(s')] \leq \texttt{cost}(s^{ad}) + \mathrm{E}[c(R_{on})] + F_{\max} \cdot \mathrm{E}[w(\mathcal{F}_{bad})].$$

The two lemmas following this proof give the following bounds:

$$w(\mathcal{F}_{bad}) \leq c(L) \qquad \text{(Lemma 2.3.3)}$$

$$\mathrm{E}[|R_{on}|] \leq \begin{cases} |\mathcal{F}_R| + O(1) \cdot |L| & \text{if } F_{\max} = 2 \\ |\mathcal{F}_R| + O(\Delta) \cdot |L| + \Delta |L|^2 & \text{if } F_{\max} = O(1) \\ |\mathcal{F}_R| + O\left(\frac{\Delta F_{\max}}{\alpha^{2F_{\max}}}\right) \cdot |L| + \Delta |L|^2 & \text{otherwise.} \end{cases} \qquad \text{(Lemma 2.3.4)}$$

Since $\texttt{cost}(s^{ad}) \geq c(L) + w(\mathcal{F}_R) = \Theta(|L|) + \Theta(|\mathcal{F}_R|)$, we have $|L|, |\mathcal{F}_R| = O(1) \cdot \texttt{cost}(s^{ad})$. We use Fact 2.2.1 to readily relate $\texttt{cost}(s')$ to $\texttt{cost}(s'')$, as required by the theorem. □

We now prove the lemmas required in the proof of Theorem 2.3.1 above. Lemma 2.3.3 simply charges the weight of each set in $\mathcal{F}_{bad}$ (sets uncovered in s' but covered in s^{ad}) to an agent that is *on* in s^{ad}.

Lemma 2.3.3. *For a cover game with constant costs and weights, and any PSA advertising strategy s^{ad},*

$$w(\mathcal{F}_{bad}) \leq c(L).$$

Proof. Note that each set in $\mathcal{F}_{bad}$ must contain a best-responding element in L_{off}, so

$$w(\mathcal{F}_{bad}) \leq \sum_{\ell \in L_{off}} \sum_{\sigma \in \mathcal{F}_{bad}: \ell \in \sigma} w_\sigma \leq \sum_{\ell \in L_{off}} c_\ell \leq c(L). \qquad \square$$

In Lemma 2.3.4 below, we bound the expected number of agents that are *on* in s' but not in s^{ad}. We do this by recognizing that each agent in R_{on} is playing a best response in s' and so is participating in at least one of the following sets in which it is the only *on* element in s': a set uncovered in s^{ad}, a set with exactly one *on* agent in s^{ad} who is *off* in s', or a set with multiple *on* agents in s^{ad} who are all *off* in s'. Our bottleneck for $F_{\max} > 2$ is in (crudely) bounding the number of sets with multiple agents *on* in s^{ad} using our hypergraph parameter Δ. We employ more sophisticated analysis to bound the *expected* number of sets with exactly one *on* agent in s^{ad} who is *off* in s'. This bound uses the fact that each agent is receptive in Phase 1 of PSA with *independent* constant probability along with algorithmic arguments that allow us to decouple related events (and a technical proposition, Proposition 2.3.5).

Lemma 2.3.4. *For a cover game with constant costs and weights, and for any PSA advertising strategy s^{ad},*

$$\mathrm{E}[|R_{on}|] \leq \begin{cases} |\mathcal{F}_R| + O(1) \cdot |L| & \text{if } F_{\max} = 2 \\ |\mathcal{F}_R| + O(\Delta) \cdot |L| + \Delta |L|^2 & \text{if } F_{\max} = O(1) \\ |\mathcal{F}_R| + O\left(\frac{\Delta F_{\max}}{\alpha^{2F_{\max}}}\right) \cdot |L| + \Delta |L|^2 & \text{otherwise.} \end{cases}$$

Proof. Each $r \in R_{on}$ plays a best-response move in s', so it must participate in a set in which it is the only *on* element (otherwise *off* would be a best response). Therefore, we can associate each $r \in R_{on}$ with a distinct set σ containing r with $\sigma \cap L \subseteq L_{off}$. Define the following partition of $\{\sigma \in \mathcal{F} : |\sigma \cap R| > 0, \sigma \cap L \subseteq L_{off}\}$.

$$\mathcal{F}^{(=0)} := \{\sigma \in \mathcal{F} \ : \ |\sigma \cap R| > 0, \ |\sigma \cap L| = 0\}$$

$$\mathcal{F}^{(=1)} := \{\sigma \in \mathcal{F} \ : \ |\sigma \cap R| > 0, \ |\sigma \cap L| = 1, \ \sigma \cap L \subseteq L_{off}\}$$

$$\mathcal{F}^{(>1)} := \{\sigma \in \mathcal{F} \ : \ |\sigma \cap R| > 0, \ |\sigma \cap L| > 1, \ \sigma \cap L \subseteq L_{off}\}$$

Observe that by definition,

$$|\mathcal{F}^{(=0)}| = |\mathcal{F}_R|. \tag{2.3.2}$$

To begin bounding $|\mathcal{F}^{(=1)}|$, recall that $\mathcal{F}^*_\ell := \{\sigma : \sigma \cap L = \ell\}$. Then:

$$\mathrm{E}[|\mathcal{F}^{(=1)}|] \leq \sum_{\ell \in L} |\mathcal{F}^*_\ell| \cdot \Pr[\ell \in L_{\mathit{off}}]. \tag{2.3.3}$$

Observe that $\ell \in L$ will never be *off* in s' if it participates more than $c_{\max}/w_{\min}$ sets where it is the unique L element and all other elements are *off* in s'. We use this fact to bound the probability that $\ell \in L_{\mathit{off}}$ by bounding the probability that many of the sets in $\mathcal{F}^*_\ell$ contain only other elements that are following their advertised strategy of *off*. However, there may be overlap in the R nodes of the sets in $\mathcal{F}^*_\ell$, so these probabilities are dependent. To circumvent this, we take some subset $\hat{\mathcal{F}}^*_\ell \subseteq \mathcal{F}^*_\ell$ such that no pair of sets in $\hat{\mathcal{F}}^*_\ell$ have common elements in R. Then, it follows that

$$\begin{aligned}
\Pr[\ell \in L_{\mathit{off}} \mid \ell \in L] &\leq \Pr[|\{\rho \in \mathcal{F}^*_\ell : \rho \backslash \{\ell\} \subseteq R_{\mathit{off}}\}| \leq c_{\max}/w_{\min}] \\
&\leq \Pr[|\{\rho \in \mathcal{F}^*_\ell : \text{ all } \rho\backslash\{\ell\} \text{ are receptive}\}| \leq c_{\max}/w_{\min}] \\
&\leq \Pr[|\{\rho \in \widehat{\mathcal{F}}^*_\ell : \text{ all } \rho\backslash\{\ell\} \text{ are receptive}\}| \leq c_{\max}/w_{\min}] \\
&\leq \Pr\left[\sum_{i=1}^{|\widehat{\mathcal{F}}^*_\ell|} X_i \leq c_{\max}/w_{\min}\right],
\end{aligned} \tag{2.3.4}$$

where $X_i \in \{0,1\}$ denotes the random variable indicating the event that for the i-th set $\rho \in \widehat{\mathcal{F}}^*_\ell$, all elements in $\rho\backslash\{\ell\}$ are receptive. We can replace the X_i in this last expression with $Y_i \in \{0,1\}$ denoting a random variable with $\Pr[Y_i = 1] = \alpha^{F_{\max}}$, because each X_i has $\Pr[X_i = 1] \geq \alpha^{F_{\max}}$. Note that by definition of $F_{\max}$ and Δ, we can choose $\hat{\mathcal{F}}^*_\ell$ with $|\mathcal{F}^*_\ell| \leq \Delta(F_{\max} - 1)|\hat{\mathcal{F}}^*_\ell|$. Then combining Inequalities (2.3.3) and (2.3.4), assuming $|\hat{\mathcal{F}}^*_\ell| \geq \lfloor c_{\max}/w_{\min} \rfloor$, and using Proposition 2.3.5, stated and proved at the

conclusion of this proof, we have:

$$\begin{aligned}
\mathrm{E}\left[|\mathcal{F}^{(=1)}|\right] &\leq \sum_{\ell\in L}|\mathcal{F}^*_\ell| \cdot \Pr\left[\sum_{i=1}^{|\widehat{\mathcal{F}}^*_\ell|} Y_i \leq c_{\max}/w_{\min}\right] \\
&\leq \sum_{\ell\in L}|\mathcal{F}^*_\ell| \sum_{i=0}^{\left\lfloor \frac{c_{\max}}{w_{\min}}\right\rfloor} \binom{|\widehat{\mathcal{F}}^*_\ell|}{i}(1-\alpha^{F_{\max}})^{|\widehat{\mathcal{F}}^*_\ell|-i}(\alpha^{F_{\max}})^i \\
&\leq (F_{\max}-1)\Delta \sum_{\ell\in L}\sum_{i=0}^{\left\lfloor \frac{c_{\max}}{w_{\min}}\right\rfloor} |\widehat{\mathcal{F}}^*_\ell| \binom{|\widehat{\mathcal{F}}^*_\ell|}{i}(1-\alpha^{F_{\max}})^{|\widehat{\mathcal{F}}^*_\ell|-i}(\alpha^{F_{\max}})^i \\
&= O\left(\frac{\Delta F_{\max}}{\alpha^{2F_{\max}}}\right)\cdot|L|. \qquad (2.3.5)
\end{aligned}$$

For $\ell \in L$ with $|\hat{\mathcal{F}}^*_\ell| < \lfloor c_{\max}/w_{\min}\rfloor$, then $|\mathcal{F}^*_\ell| \leq O(\Delta F_{\max})$, which is dominated by the corresponding term in the previous expression. Then our final $\mathcal{F}^{(=1)}$ bound is:

$$\mathrm{E}[|\mathcal{F}^{(=1)}|] = O\left(\frac{\Delta F_{\max}}{\alpha^{2F_{\max}}}\right)\cdot|L| \qquad (2.3.6)$$

Since each $\sigma \in \mathcal{F}^{(>1)}$ contains at least three elements, there are no such sets when $F_{\max} = 2$. For the general case, we use the definition of Δ and the fact that there are $\binom{|L|}{2} \leq L^2$ pairs of agents in L:

$$|\mathcal{F}^{(>1)}| \leq \begin{cases} 0 & \text{if } F_{\max} = 2 \\ \Delta\cdot|L|^2 & \text{otherwise.} \end{cases} \qquad (2.3.7)$$

Finally, since $|R_{on}| \leq |\mathcal{F}^{(=0)}|+|\mathcal{F}^{(>1)}|+|\mathcal{F}^{(=1)}|$ by construction, Inequalities (2.3.2), (2.3.6), and (2.3.7) together give the desired conclusion of Lemma 2.3.4, noting that $\Delta = 1$ when $F_{\max} = 2$. □

Proposition 2.3.5. *For any $a \in (0,1)$ and $0 < c \leq d$,*

$$\sum_{i=0}^{\lfloor c\rfloor} d\binom{d}{i}(1-a)^{d-i}a^i = O\left(\frac{\lceil c\rceil}{a^2(1-a)^2}\right).$$

Proof. This is immediate in the case that $c < 1$ because $d(1-a)^d = O(1/a)$ for all $d \geq 0$ as long as $a \in (0,1)$. Hence, assume $c \geq 1$. Let $\bar{a} = \max(a, 1-a)$ and define $\xi \in (0,1)$ to be the largest real number satisfying

$$(e/\xi)^\xi < \sqrt{1/\bar{a}},$$

where it is not hard to check that $\xi = \Omega((1-\bar{a})^2)$. For the case with $d \leq c/\xi$, $c \leq d$ gives

$$\sum_{i=0}^{\lfloor c \rfloor} d\binom{d}{i}(1-a)^{d-i}a^i \leq d\sum_{i=0}^{d}\binom{d}{i}(1-a)^{d-i}a^i = d \leq c/\xi = O(c/(1-\bar{a})^2).$$

Now consider when $d > c/\xi$. Observe that

$$d\sum_{i=0}^{\lfloor c \rfloor}\binom{d}{i}(1-a)^{d-i}a^i \;\leq\; d\cdot\bar{a}^d\sum_{i=0}^{\lfloor c \rfloor}\binom{d}{i} \;\leq\; d\cdot\bar{a}^d\sum_{i=0}^{\lfloor c \rfloor}\frac{d^i}{i!}$$

Observe that $d\cdot\bar{a}^{d/2} = O(1/(1-\bar{a}))$, $d^i/i!$ is increasing with respect to i for $i < c < d$, $x! = \Omega((x/e)^x)$, $c < \xi \cdot d$, and $(e/\xi)^\xi < \sqrt{1/\bar{a}}$. Then we can complete our proof of Proposition 2.3.5 as follows:

$$\begin{aligned}
d\cdot\bar{a}^d\sum_{i=0}^{\lfloor c \rfloor}\frac{d^i}{i!} &= O(1/(1-\bar{a}))\cdot\bar{a}^{d/2}\sum_{i=0}^{\lfloor c \rfloor}\frac{d^i}{i!} \\
&= O(c/(1-\bar{a}))\cdot\bar{a}^{d/2}\cdot\frac{d^{\lfloor c \rfloor}}{\lfloor c \rfloor!} \\
&= O(c/(1-\bar{a}))\cdot\bar{a}^{d/2}\Big(\frac{d\cdot e}{\lfloor c \rfloor}\Big)^{\lfloor c \rfloor} \\
&= O(c/(1-\bar{a}))\cdot\bar{a}^{d/2}\Big(\frac{d\cdot e}{\xi\cdot d}\Big)^{\xi\cdot d} \\
&= O(c/(1-\bar{a}))\cdot\bar{a}^{d/2}\cdot\bar{a}^{-d/2} \\
&= O(c/(1-\bar{a})),
\end{aligned}$$

□

Theorem 2.3.1 with arbitrary costs and weights. Requiring constant costs and weights allows for simplifications such as $c(L) = O(|L|)$ in the previous proofs. Following these proof exactly (using bounds such as $c(L) \leq c_{\max}\cdot|L|$), we can obtain

the following results for arbitrary costs and weights, although the calculation is routine and omitted for brevity:

Claim 2.3.6. *For a cover game with arbitrary costs and weights, and for any advertising strategy s^{ad}, the expected cost at the end of PSA is*

$$\mathrm{E}[\texttt{cost}(s'')] \;=\; \begin{cases} O\left(\left\lceil \frac{c_{\max}}{w_{\min}} \right\rceil \frac{c_{\max}}{c_{\min}}\right) \texttt{cost}(s^{ad}) & \text{if } F_{\max} = 2 \\ O\left(\Delta \left\lceil \frac{c_{\max}}{w_{\min}} \right\rceil \frac{c_{\max}}{c_{\min}^2}\right) \texttt{cost}(s^{ad})^2 & \text{if } F_{\max} = O(1) \\ O\left(\frac{\Delta F_{\max}^2}{\alpha^{2F_{\max}}} \left\lceil \frac{c_{\max}}{w_{\min}} \right\rceil \frac{c_{\max}}{c_{\min}^2}\right) \texttt{cost}(s^{ad})^2 & \text{otherwise.} \end{cases}$$

2.3.1 Better Results for Careful Advertising

For improved performance guarantees, we look for strategies that are efficient in a particular sense. Recall that the core minimum δ^* of a given strategy profile s is the minimum number of sets uniquely covered by any particular *on* element in s. We say that an advertising strategy s^{ad} satisfies Condition $(\star)$ if for all $x \geq \frac{\delta^*}{\Delta(F_{\max}-1)}$,

$$\left(\left\lfloor \frac{c_{\max}}{w_{\min}} \right\rfloor + 1\right) x^{\lfloor c_{\max}/w_{\min} \rfloor} \left(1 - \alpha^{F_{\max}}\right)^{x - \lfloor c_{\max}/w_{\min} \rfloor} \;\leq\; \frac{1}{n^2}. \qquad (\star)$$

Intuitively, this condition ensures that each element that is *on* in s^{ad} is in many sets in which it is the unique element that is *on* in s^{ad}. When this is the case, it is very likely that in Phase 1, some of these sets will have *each* element (except perhaps the single element *on* in s^{ad}) receptive to advertising. This unique *on* element will turn *on* in Phase 1, and every set will be covered. We achieve the precise condition by reverse engineering our analysis starting with this goal.

Fact 2.3.7. *For a cover game with constant costs, weights, and $F_{\max}$, there exists a polynomial-time algorithm that computes a strategy s^{ad} satisfying Condition $(\star)$ with $\texttt{cost}(s^{ad}) = O(\Delta \log n) \cdot OPT$.*

Proof. Consider the following algorithm, which is clearly poly-time:

1. Let s^* be the strategy with social cost $O(1) \cdot OPT$ obtained by LP-rounding (Fact 2.2.2).

2. Greedily turn *off* every agent that is the unique *on* element in fewer than $B\Delta \log n$ sets in s^*, for some sufficiently large constant B depending on $c_{\max}/w_{\min}$, α, and $F_{\max}$. Call the result s^{ad}.

Then for $x \geq \delta^*(s^{ad})/(\Delta(F_{\max}-1)) \geq B\log n/(F_{\max}-1) = \Theta(\log n)$, we have

$$\begin{aligned}(\lfloor c_{\max}/w_{\min}\rfloor + 1)x^{\lfloor c_{\max}/w_{\min}\rfloor}(1-\alpha^{F_{\max}})^{x-\lfloor c_{\max}/w_{\min}\rfloor} &= O(x^{\lfloor c_{\max}/w_{\min}\rfloor}(1-\alpha^{F_{\max}})^x) \\ &= O(1/n^d)\end{aligned}$$

for arbitrarily large constant d (depending on sufficiently large constant B), so s^{ad} satisfies Condition ($\star$). Furthermore, $\texttt{cost}(s^{ad}) = O(\Delta \log n) \cdot OPT$ because at most $OPT/c_{\min}$ agents are *on* in s^*, and then turning any one *off* results in at most $F_{\max}B\Delta \log n$ sets becoming uncovered. □

Theorem 2.3.8 formalizes a high probability and stronger version of Theorem 2.3.1 for the general set cover game in PSA when s^{ad} satisfies Condition ($\star$). Note that this result requires no assumptions on the costs, weights, or $F_{\max}$ of the hypergraph.

Theorem 2.3.8. *For any cover game, and for any advertising strategy s^{ad} satisfying Condition ($\star$), with probability at least $1 - 1/n$ the cost at the end of PSA is*

$$\texttt{cost}(s'') = O(F_{\max}) \cdot \texttt{cost}(s^{ad}).$$

Using the greedily constructed advertising strategy described in the proof of Fact 2.3.7, we have the following immediate corollary in the case that costs, weights, and $F_{\max}$ are constant:

Corollary 2.3.9. *For a cover game with constant costs, weights, and $F_{\max}$, there exists a poly-time algorithm to find an advertising strategy s^{ad} such that with probability at least $1 - 1/n$ the cost at the end of PSA is*

$$\texttt{cost}(s'') = O(\Delta \log n) \cdot OPT.$$

Proof of Theorem 2.3.8. Using the notation from Theorem 2.3.1, we bound the cost at the end of Phase 1 as:

$$\mathrm{E}[\texttt{cost}(s')] \leq \texttt{cost}(s^{ad}) + c(R_{on}) + F_{\max} \cdot w(\mathcal{F}_{bad}).$$

Lemma 2.3.10 below proves that for s^{ad} satisfying Condition ($\star$), all agents in L turn *on* in Phase 1 (and so $w(\mathcal{F}_{bad}) = 0$) with probability at least $1 - 1/n$, and under this event, the cost of agents in R_{on} is bounded by $w(\mathcal{F}_R) \leq \texttt{cost}(s^{ad})$. This proves that $\texttt{cost}(s') = O(\texttt{cost}(s^{ad}))$ with all but at most $1/n$ probability, and then Theorem 2.3.8 follows from Fact 2.2.1. □

Lemma 2.3.10. *For any cover game, and for advertising strategy s^{ad} satisfying Condition ($\star$), then $\mathcal{F}_{bad} = \emptyset$ and $c(R_{on}) \leq w(\mathcal{F}_R)$ with probability at least $1 - 1/n$.*

Proof. As in the proof of Lemma 2.3.4 (and using the same notation), for any $\ell \in L$ there is some subset $\widehat{\mathcal{F}}^*_\ell \subseteq \mathcal{F}^*_\ell$ such that no pair of sets in $\widehat{\mathcal{F}}^*_\ell$ have common elements in R and $|\widehat{\mathcal{F}}^*_\ell| \geq \frac{|\mathcal{F}^*_\ell|}{\Delta(F_{\max}-1)} \geq \frac{\delta^*}{\Delta(F_{\max}-1)}$. Applying the bound on $\Pr[\ell \in L_{off} \mid \ell \in L]$ derived in the proof Lemma 2.3.4 as a starting point,

$$\begin{aligned}
\Pr[\ell \in L_{off} \mid \ell \in L] &\leq \sum_{i=0}^{\left\lfloor \frac{c_{\max}}{w_{\min}} \right\rfloor} \binom{|\widehat{\mathcal{F}}^*_\ell|}{i} (1-\alpha^{F_{\max}})^{|\widehat{\mathcal{F}}^*_\ell|-i} (\alpha^{F_{\max}})^i \\
&\leq \sum_{i=0}^{\left\lfloor \frac{c_{\max}}{w_{\min}} \right\rfloor} |\widehat{\mathcal{F}}^*_\ell|^i (1-\alpha^{F_{\max}})^{|\widehat{\mathcal{F}}^*_\ell|-i} \\
&\leq \left(\left\lfloor \frac{c_{\max}}{w_{\min}} \right\rfloor + 1 \right) |\widehat{\mathcal{F}}^*_\ell|^{\left\lfloor \frac{c_{\max}}{w_{\min}} \right\rfloor} (1-\alpha^{F_{\max}})^{|\widehat{\mathcal{F}}^*_\ell| - \left\lfloor \frac{c_{\max}}{w_{\min}} \right\rfloor},
\end{aligned}$$ □

and by the assumption that s^{ad} satisfies Condition ($\star$), the above expression is at most $1/n^2$. By union bound, $\Pr[L_{off} = \emptyset] \geq 1 - 1/n$ and hence $\mathcal{F}_{bad} = \emptyset$ at the end of Phase 1 with at least this probability.

Assume this event, and observe that for each best-responding $r \in R_{on}$, c_r is no greater than the total weight of all sets containing r as the unique *on* agent. Since we assume all nodes in L are *on*, these sets are a subset of $\mathcal{F}_R$. Further, since there is no

overlap in these sets between different agents in R_{on}, we can sum over all $r \in R_{on}$ to derive $c(R_{on}) \leq w(\mathcal{F}_R)$. This completes the proof of Lemma 2.3.10.

2.4 Set Cover in the Learn-Then-Decide Model

Next we study the set cover game in the learn-then-decide (LTD) model of [7]. In contrast to PSA, agents in LTD are neither strictly receptive nor strictly best-responders in the initial exploration phase, but they choose one of these options for the final exploitation phase. The PSA model is appropriate for an engineering setting such as sensor networks, where devices may be programmed to respond in Phase 1 to a signal that only reaches some devices due to technical constraints. On the other hand, the LTD model is better for a social setting in which agents may be skeptical of the central authority, and so they experiment in Phase 1, sometimes following the advertised strategy and other times applying a best-response strategy.

1: Play begins in an arbitrary state, and a central authority advertises joint action $s^{ad} \in S$. Agent i is associated with fixed probability $p_i \geq \beta \in (0, 1)$, where β is constant. Agents are chosen to update uniformly at random for each of T^* time steps. When i updates, he plays s_i^{ad} with probability p_i or a best-response move with probability $1 - p_i$. The state at time T^* is denoted s'.

2: At time T^*, all agents in random order individually commit arbitrarily to s_i^{ad} or the best-response strategy. Finally, agents take turns in random order playing their chosen strategy until best-responders reach a Nash equilibrium s'' given the fixed behavior of s^{ad} followers.

Overview of results. Theorem 2.4.1 relates the cost of the outcome of LTD to that of any advertised strategy. At a high level, Phase 1 of LTD is similar enough to PSA that we can borrow previous techniques to bound the cost of s' in Lemma 2.4.3. To do this, we have to make a high probability assumption on the order of updates

in Phase 1. Specifically, we let $\mathcal{E} = \mathcal{E}(T', T^*)$ for $1 < T' < T^*$ denote the event that every element in L updates at least once before time T' *after* every element in R has updated at least once, and then each element in R again updates at some time $t \in [T', T^*]$. Note that we can choose $T', T^* \in \text{poly}(n)$ such that $\Pr[\mathcal{E}] \geq 1 - 1/n^{F_{\max}}$ if $F_{\max} = O(1)$. Because Phase 2 in LTD is not simple best-response dynamics, the potential argument that cost stays low in Phase 2 of PSA does not apply. Instead, we develop new techniques in Lemma 2.4.4, and this causes us to lose an extra Δ factor relative to Theorem 2.3.1 for PSA. Corollary 2.4.2 puts these results in terms of the global optimum.

We do not make routine efforts to obtain LTD analogs of the results for arbitrary $F_{\max}$, costs, and weights given for PSA (Theorem 2.3.1, Corollary 2.3.2, Claim 2.3.6) since these results appear to be far from tight.

Theorem 2.4.1. *For a cover game with constant costs and weights, there exists a $T^* \in poly(n)$ such that for any advertising strategy s^{ad}, the expected cost at the end of LTD is*

$$\mathrm{E}[\texttt{cost}(s'')] = \begin{cases} O(1) \cdot \texttt{cost}(s^{ad}) & \text{if } F_{\max} = 2 \\ O(\Delta^2) \cdot \texttt{cost}(s^{ad})^2 & \text{if } F_{\max} = O(1). \end{cases} \tag{2.4.1}$$

If s^{ad} is obtained from the LP-rounding approximation algorithm described in Section 2.2.3, the following corollary is immediate from Theorem 2.4.1:

Corollary 2.4.2. *For a cover game with constant costs and weights, there exists a $T^* \in poly(n)$ and a poly-time algorithm to find an advertising strategy s^{ad} such that the expected cost at the end of LTD is*

$$\mathrm{E}[\texttt{cost}(s'')] = \begin{cases} O(1) \cdot OPT & \text{if } F_{\max} = 2 \\ O(\Delta^2) \cdot OPT^2 & \text{if } F_{\max} = O(1). \end{cases}$$

In proving Theorem 2.4.1, we find that although LTD differs from PSA in both phases, we can analyze Phase 1 of LTD in a manner similar to Phase 1 of PSA by

defining an event that occurs with high probability in Phase 1 of LTD and then modifying the techniques of Theorem 2.3.1 to bound the cost of the state at the end of Phase 1 relative to that of the advertised strategy (Lemma 2.4.3). However, showing that the cost stays low in Phase 2 (Lemma 2.4.4) imposes new challenges that we circumvent using the fact that update order is random, and this causes us to lose an additional Δ factor.

Proof of Theorem 2.4.1. Note $\texttt{cost}(s) \leq c_{\max} \cdot n + w_{\max} \cdot F_{\max} \cdot |\mathcal{F}| = O(n^{F_{\max}})$ for any $s \in S$. Then:

$$\begin{aligned} \mathrm{E}[\texttt{cost}(s'')] &= \Pr[\mathcal{E}] \cdot \mathrm{E}[\texttt{cost}(s'') \mid \mathcal{E}] + \Pr[\mathcal{E}^c] \cdot \mathrm{E}[\texttt{cost}(s'') \mid \mathcal{E}^c] \\ &\leq \mathrm{E}[\texttt{cost}(s'') \mid \mathcal{E}] + \frac{1}{n^{F_{\max}}} \cdot O\left(n^{F_{\max}}\right) \\ &= \mathrm{E}[\texttt{cost}(s'') \mid \mathcal{E}] + O(1), \end{aligned} \tag{2.4.3}$$

□

so it suffices to bound $\mathrm{E}[\texttt{cost}(s'') \mid \mathcal{E}]$. Lemma 2.4.3 bounds the expected social cost at the end of Phase 1 under the event $\mathcal{E}$, and a bound on the increase in social cost in Phase 2 is given in Lemma 2.4.4. Together, these lemmas imply Theorem 2.4.1.

Lemma 2.4.3. *For a cover game with constant costs and weights, for any LTD advertising strategy s^{ad}, and for event $\mathcal{E}$ as defined above,*

$$\mathrm{E}[\texttt{cost}(s') \mid \mathcal{E}] = \begin{cases} O(1) \cdot \texttt{cost}(s^{ad}) & \text{if } F_{\max} = 2 \\ O(\Delta) \cdot \texttt{cost}(s^{ad})^2 & \text{if } F_{\max} = O(1). \end{cases}$$

Proof. Recall that as in the proof of Theorem 2.3.1, $\texttt{cost}(s') = \texttt{cost}(s^{ad}) + O(|R_{on}|) + O(w(\mathcal{F}_{bad}))$. Since $\texttt{cost}(s^{ad}) \geq c(L) + w(\mathcal{F}_R) = \Theta(|L|) + \Theta(|\mathcal{F}_R|)$, we have $|L|, |\mathcal{F}_R| = O(1) \cdot \texttt{cost}(s^{ad})$, so it suffices to bound $w(\mathcal{F}_{bad})$ and $|R_{on}|$ in terms of $|L|$ and $|\mathcal{F}_R|$.

Lemma 2.3.3 bounds $w(\mathcal{F}_{bad}) \leq c(L)$ in the PSA model by obvserving that elements in $L_{off} \cap \mathcal{F}_{bad}$ are best-responding at the end of Phase 1. This clean argument does

not apply in LTD, since Phase 1 terminates at a pre-determined time T^*, and so agents that played best-response most recently are not necessarily still in best-response. Instead, we charge the weight of each set in $\mathcal{F}_{bad}$ that is entirely contained in L to its element that played best-response most recently. We then bound the weight of the other $\mathcal{F}_{bad}$ sets, which each contain some *off* element in R, using the analysis of Lemma 2.3.4.

Let $\mathcal{F}_{bad\subseteq L} := \{\sigma : \sigma \subseteq L_{off}\}$. Attribute the weight of $\sigma \in \mathcal{F}_{bad\subseteq L}$ to its element ℓ that updated most recently before the end of Phase 1. Because $\ell \in L_{off}$ played best-response most recently, the weight of all sets in $\mathcal{F}_{bad\subseteq L}$ attributed to ℓ is at most c_ℓ. Summing over all $\ell \in L_{off} \subseteq L$ gives $w(\mathcal{F}_{bad\subseteq L}) \leq c(L) = O(1) \cdot |L|$.

Now let $\mathcal{F}_{bad\not\subseteq L} := \mathcal{F}_{bad} \backslash \mathcal{F}_{bad\subseteq L}$, whose sets each have elements in both L and R, all *off* in s'. Recall the definitions of $\mathcal{F}^{(=1)}$ and $\mathcal{F}^{(>1)}$ in the proof of Lemma 2.3.4 and observe that $\mathcal{F}_{bad\not\subseteq L} \subseteq \mathcal{F}^{(=1)} \cup \mathcal{F}^{(>1)}$. Assuming $\mathcal{E}$, we can modify the analysis of $|\mathcal{F}^{(=1)}|$ in Lemma 2.3.4 to get an analogous result. To do this, apply the analysis in Inequalities (2.3.4), replacing "$\rho\backslash\{\ell\} \subseteq R_{off}$" with "$\rho\backslash\{\ell\}$ are all *off* when ℓ last updates in Phase 1," and replacing "all $\rho\backslash\{\ell\}$ are receptive" with "each $r \in \rho\backslash\{\ell\}$ plays s_r^{ad} when it last updates before the last update of ℓ in Phase 1." Note that r plays s^{ad} at any given update in Phase 1 with probability $p_r \geq \beta$ in LTD, so replacing PSA probability α with β, Equation (2.3.6) holds for $\mathrm{E}[|\mathcal{F}^{(=1)}|]$ in LTD. This dominates $w(\mathcal{F}_{bad\subseteq L}) = O(1) \cdot |L|$, so adding Inequality (2.3.7) for $|\mathcal{F}^{(>1)}|$ gives:

$$\mathrm{E}[w(\mathcal{F}_{bad}) \mid \mathcal{E}] = \begin{cases} O(1) \cdot |L| & \text{if } F_{\max} = 2 \\ O(\Delta) \cdot |L| + \Delta |L|^2 & \text{if } F_{\max} = O(1). \end{cases} \tag{2.4.4}$$

This modification to the proof of Lemma 2.3.4 also bounds $|R_{on}|$ in the LTD model assuming $\mathcal{E}$:

$$E\left[|R_{on}| \mid \mathcal{E}\right] = \begin{cases} |\mathcal{F}_R| + O(1) \cdot |L| & \text{if } F_{\max} = 2 \\ |\mathcal{F}_R| + O(\Delta) \cdot |L| + \Delta |L|^2 & \text{if } F_{\max} = O(1). \end{cases} \tag{2.4.5}$$

Together, Equations (2.4.4) and (2.4.5) give Lemma 2.4.3. □

From Fact 2.2.1 and $F_{\max} = O(1)$, we have $\texttt{cost}(s'') \leq O(\Phi(s'') - \Phi(s')) + O(\texttt{cost}(s'))$. The bound on expected potential change given in Lemma 2.4.4 below therefore implies our desired bound on the cost at the end of LTD in Theorem 2.4.1. Our potential change bound employs new probabilistic reasoning using agents' random update order to bound the total impact of updates that increase potential. We do this by bounding the number of sets uncovered by some of these updates, and the last step in this reasoning decouples dependent events by creating an R-disjoint subset of sets as in the bound on $|\mathcal{F}^{(=1)}|$ in the proof of Lemma 2.3.4. This is responsible for the extra Δ term compared to the bounds for the PSA model.

Lemma 2.4.4. *For a cover game with constant costs and weights, for any LTD advertising strategy s^{ad}, and for event $\mathcal{E}$ as defined above,*

$$\mathrm{E}[\Phi(s'') - \Phi(s') \mid \mathcal{E}] = \begin{cases} O(1) \cdot \texttt{cost}(s^{ad}) & \text{if } F_{\max} = 2 \\ O(\Delta^2) \cdot \texttt{cost}(s^{ad})^2 & \text{if } F_{\max} = O(1). \end{cases}$$

Proof. Since best-response moves do not increase the potential function Φ, we only consider updates of agents following the advertising strategy s^{ad} in Phase 2. Since each such agent changes strategies at most once in Phase 2, it suffices to consider a single *off*-*on* move for each agent in L (i.e., the agent changes her strategy from *off* to *on*) and a single *on*-*off* move for each agent in R_{on}. For each $\ell \in L$, an *off*-*on* move increases potential by at most c_ℓ, so

$$\textit{off-on} \text{ moves increase potential by } \leq c(L) = O(|L|). \tag{2.4.6}$$

Let $R_{on-off} := \{r \in R_{on} : r \text{ turns } \textit{off} \text{ in Phase 2, following } s^{ad}\}$. The potential increase due to the first Phase 2 update of $r \in R_{on-off}$ at time t is the weight of sets that become uncovered by this update. For $r \in R_{on-off}$, let $\mathcal{F}_r^*$ be the collection of

sets containing r such that all of their other elements are *off* at time t. The potential increases by at most $w(\mathcal{F}_r^*)$ at time t, so noting that $\mathcal{F}_{r_1}^* \cap \mathcal{F}_{r_2}^* = \emptyset$ if $r_1 \neq r_2$, we have:

$$\textit{on-off} \text{ moves increase potential by } \leq w(\textstyle\sum_{r \in R_{on-off}} |\mathcal{F}_r^*|) = O\left(\left|\cup_{r \in R_{on-off}} \mathcal{F}_r^*\right|\right). \tag{2.4.7}$$

To bound $|\cup_{r \in R_{on-off}} \mathcal{F}_r^*|$, we partition $\cup_{r \in R_{on-off}} \mathcal{F}_r^* = \mathcal{F}^{(L_{off})} \cup \mathcal{F}^{(L_{on}^*)} \cup \mathcal{F}^{(L_{on})}$ for:

$$\mathcal{F}^{(L_{off})} := \{\sigma \in \cup_{r \in R_{on-off}} \mathcal{F}_r^* \ : \ \sigma \cap L \subseteq L_{off}\}$$
$$\mathcal{F}^{(L_{on}^*)} := \{\sigma \in \cup_{r \in R_{on-off}} \mathcal{F}_r^* \ : \ \sigma \cap L = \{\ell\} \subseteq L_{on}\}$$
$$\mathcal{F}^{(L_{on})} := \{\sigma \in \cup_{r \in R_{on-off}} \mathcal{F}_r^* \ : \ |\sigma \cap L| > 1, \sigma \cap L_{on} \neq \emptyset\}.$$

Recall the definitions of $\mathcal{F}^{(>1)}, \mathcal{F}^{(=1)}$ from Lemma 2.3.4. Note that $\mathcal{F}^{(L_{off})} \backslash \mathcal{F}_R \subseteq \mathcal{F}^{(>1)} \cup \mathcal{F}^{(=1)}$ since every such set has an element in R and at least one element in L, all of which are *off* in s'. We modify the analysis of in Lemma 2.3.4 in the same manner used to justify Equation (2.4.4) in Lemma 2.4.3 to obtain:

$$\mathrm{E}[|\mathcal{F}^{(L_{off})}| \mid \mathcal{E}] \ \leq \begin{cases} |\mathcal{F}_R| + O(1) \cdot |L| & \text{if } F_{\max} = 2 \\ |\mathcal{F}_R| + O(\Delta) \cdot |L| + \Delta |L|^2 & \text{if } F_{\max} = O(1). \end{cases} \tag{2.4.8}$$

We now show that random updates in Phase 2 limits the expected number of sets $\sigma \in \mathcal{F}^{(L_{on}^*)}$, each containing an element $r_\sigma \in R_{on-off}$ and a unique element $\ell_\sigma \in L$, which is *on* at the end of Phase 1. Our key observation is that if the (first and only) *on-off* move of r_σ uncovers σ in Phase 2, then ℓ_σ must have turned *off* earlier in Phase 2. Note that ℓ_σ can turn *off* only if doing so uncovers at most $c_{\max}/w_{\min}$ sets in which ℓ_σ participates. Hence, all but at most $c_{\max}/w_{\min}$ sets of the following type must have an element in R_{off} that updates (and, in particular, turns *on*) before ℓ_σ updates, and therefore also before r_σ updates:

$$\mathcal{F}_{\ell_\sigma}^{(R_{off})} = \{\rho \in \mathcal{F}_{\ell_\sigma}^* \ : \ \rho \backslash \{\ell_\sigma\} \subseteq R_{off}\}.$$

We use these observations to bound the probability that σ containing some $r_\sigma \in R_{on-off}$ and a unique $\ell_\sigma \in L$ with $\ell_\sigma \in L_{on}$ is uncovered by r_σ, where randomness is taken over the location of r_σ in an arbitrary fixed update order of the other agents in R. There are at least $|\mathcal{F}_{\ell_\sigma}^{(R_{off})}|/\Delta$ elements that are the first updating R agent in some set $\rho \in \mathcal{F}_{\ell_\sigma}^{(R_{off})}$, and r_σ can update before at most $c_{\max}/w_{\min}$ of them in order for ℓ_σ to have a chance to turn *off*. Therefore, we can bound the probability that σ is uncovered by r_σ as:

$$\Pr[\sigma \in \mathcal{F}^*_{r_\sigma} \mid r_\sigma \in \sigma \cap R_{on-off}, \sigma \cap L = \{\ell_\sigma\} \subseteq L_{on}, \mathcal{E}] \leq \frac{c_{\max}/w_{\min} + 1}{|\mathcal{F}_{\ell_\sigma}^{(R_{off})}|/\Delta + 1},$$

By union bound over all $r \in \sigma \cap R_{on-off}$ that could uncover σ with an *on-off* move, and for $\ell_\sigma \in L$,

$$\begin{aligned}\Pr[\sigma \in \cup_{r \in (\sigma \cap R_{on-off})} \mathcal{F}^*_r \mid \sigma \cap L = \{\ell_\sigma\} \subseteq L_{on}, \mathcal{E}] &\leq (F_{\max} - 1) \cdot \frac{c_{\max}/w_{\min} + 1}{|\mathcal{F}_{\ell_\sigma}^{(R_{off})}|/\Delta + 1} \\ &= O\left(\frac{\Delta}{|\mathcal{F}_{\ell_\sigma}^{(R_{off})}| + 1}\right).\end{aligned}$$

Now recall that $\mathcal{F}^*_\ell$ denotes the sets uniquely covered by ℓ in s^{ad}, and there is a subset $\hat{\mathcal{F}}^*_\ell \subseteq \mathcal{F}^*_\ell$ of size at least $\frac{|\mathcal{F}^*_\ell|}{\Delta(F_{\max}-1)}$ with disjoint elements in R. Note that for any $\sigma \in \mathcal{F}$, we have a lower bound $\beta^{F_{\max}}$ on the probability that all $\sigma \cap R$ followed s^{ad} and turned *off* in their last update in Phase 1. Thus we can argue that given $\mathcal{E}$, the random variable $|\mathcal{F}_\ell^{(R_{off})}|$ has (first-order) dominance over the binomial random

variable $X \sim B\left(\frac{|\mathcal{F}^*_\ell|}{\Delta(F_{\max}-1)}, \beta^{F_{\max}}\right)$. Using this, we have

$$\begin{aligned}
\mathrm{E}[|\mathcal{F}^{(L^*_{on})}| \mid \mathcal{E}] &\leq \sum_{\ell \in L} \sum_{\sigma \in \mathcal{F}^*_\ell} \Pr[\sigma \in \cup_{r \in (\sigma \cap R_{on-off})} \mathcal{F}^*_r \mid \mathcal{E}] \\
&\leq \sum_{\ell \in L} \sum_{\sigma \in \mathcal{F}^*_\ell} O\left(\mathrm{E}\left[\frac{\Delta}{|\mathcal{F}^{(R_{off})}_{\ell_\sigma}| + 1}\right]\right) \\
&\leq \sum_{\ell \in L} \sum_{\sigma \in \mathcal{F}^*_\ell} O(\Delta) \cdot \mathrm{E}\left[\frac{1}{X+1}\right] \\
&\leq \sum_{\ell \in L} \sum_{\sigma \in \mathcal{F}^*_\ell} O(\Delta) \cdot \frac{\Delta(F_{\max}-1)}{|\mathcal{F}^*_\ell|} \\
&= O(\Delta^2) \cdot |L|, \qquad (2.4.9)
\end{aligned}$$

where the third inequality uses $\beta = \Theta(1)$ and the fact that $\mathrm{E}[1/(1+Y)] \leq \frac{1}{np}$ for binomial random variable $Y \sim B(n,p)$. Recall that $\Delta = 1$ in the special case that $F_{\max} = 2$.

Finally, note that it is trivial to bound $|\mathcal{F}^{(L_{on})}|$ in a manner similar to the bound for $|\mathcal{F}^{(>1)}|$ in Lemma 2.3.4, using the fact that any pair of L can participate in $\leq \Delta$ sets:

$$|\mathcal{F}^{(L_{on})}| \leq \begin{cases} 0 & \text{if } F_{\max} = 2 \\ \Delta |L|^2 & \text{if } F_{\max} = O(1). \end{cases} \qquad (2.4.10)$$

Inequalities (2.4.6), (2.4.7), (2.4.8), (2.4.9), and (2.4.10) together give the desired conclusion. □

We note that our $(\texttt{cost}(s^{ad}))^2$ term is due to the crude bound of $\Delta|L|^2$ on the number of sets with multiple L agents (as in Lemmas 2.4.3 and 2.3.4 bounding cost at the end of Phase 1 of LTD and PSA). The additional Δ^2 factor in LTD compared to PSA is due to our new decoupling technique used in Inequality (2.4.9), which bounds the number of sets with a unique L element that is *on* at the beginning of Phase 2 and becomes uncovered by an R element following its advertising strategy.

Tightness of Our Results and Open Questions

In the case of the vertex cover setting, in which all sets are of size two, our results are essentially tight. Furthermore, such a setting arises in practical wireless sensing networks [104]. In the more general set cover setting, we still get strong results assuming constant size sets, although our results may not be tight. An additional benefit of our constant set size assumption, i.e., $F_{\max} = O(1)$, is that it allows us to give a poly-time procedure for both computing a good advice strategy and then letting the dynamics converge to an equilibrium that is within the $O(\log n)$ factor of optimal. We remark that were it not for the constant set size assumption, this result would be optimal, since [85] show that finding an $o(\log n)$-approximation of the general set cover problem is NP-hard.

Since there exists a poly-time algorithm for $O(\log n)$-approximation of the general set cover problem [21], it is conceivable that different analysis permitting arbitrary set sizes and possibly using a different characterization of a good advice strategy could give this optimal result. This is indeed an interesting open question: for arbitrary $F_{\max}$, do the dynamics models with the set cover games studied in this work converge to an equilibrium that is within the $O(\log n)$ factor of optimal? If the answer is yes, our dynamics may provide an alternative $O(\log n)$-approximation algorithm for the set cover problem.

Related work subsequent to the conference version of this work has analyzed similar settings with variants on the dynamics and games studied here (see, e.g., [83, 46]). Our techniques may continue to be of broader interest for analyzing other classic optimization problems in a distributed fashion.

CHAPTER III

DIFFERENTIALLY PRIVATE INDEPENDENT COMPONENT ANALYSIS

3.1 Overview

In [61], we provide a method for conducting independent component analysis (ICA) while maintaining differential privacy with respect to an underlying database. We prove privacy for a liberal definition of neighboring databases. Our method is compatible with a broad class of ICA algorithms, and we show that the noise we add for privacy does not degrade the accuracy of these algorithms much.

The basic ICA model assumes p independent real random *source variables* $s = (s_1, \ldots, s_p) \in \mathbb{R}^p$, each with some (non-Gaussian and possibly unknown) distribution over $\mathbb{R}$. We observe do not observe these variables directly, rather we observe them through p *signal variables* $x = (x_1, \ldots, x_p)$, which are linear combinations of the source variables s under a fixed nonsingular *mixing matrix* $A \in \mathbb{R}^{p \times p}$ such that $x = sA$. The goal of ICA is to recover (up to trivial equivalencies) the mixing matrix A from the distribution of x, which then also reveals the source variables s. In practice, we receive random samples from the approximate distribution of x, and we estimate A from these samples. Our work is motivated by the observation that many ICA applications have a natural privacy aspect, in that the observations x and underlying source variables s may reflect confidential information about individuals, whereas the mixing matrix represents an underlying (and non-confidential) structure that the analysis is attempting to discover. For example, the source signals and observations could respectively represent mutations in a genome and incidents of diseases, and the mixing matrix would capture how the former influences the latter.

After covering ICA, differential privacy, and other relevant preliminaries in Section 3.2, we present our mechanism and prove its privacy in Section 3.3. In Section 3.4, we give an algorithm-agnostic characterization and proof of our mechanism's utility for ICA, and we specialize these results for a provable ICA algorithm in Section 3.5. The rest of this section provides an overview of our main approach and techniques.

Approach

Our approach begins with the observation that all known ICA algorithms work in two main phases. The first phase calculates the mean and covariance of the data, and it applies a corresponding affine transformation to obtain *whitened* data, which is zero-centered and isotropic. In other words, this phase reduces to the case where each source variable s_i has expectation 0 and variance 1, and the mixing matrix A is orthonormal, i.e., $A^{-1} = A^T$. The second phase recovers the columns of A^{-1} via some method of multivariate optimization on the one-dimensional marginals $\langle x, u \rangle = xu^T$ over unit vectors $u \in \mathbb{R}^p$. The basic idea is that if u is a column of A^{-1} (or its negation), then $\langle x, u \rangle = sAu^T$ is exactly one of the source variables (or its negation); otherwise, $\langle x, u \rangle$ is some normalized mixture of two or more source variables.

For many popular ICA algorithms (especially those with provable guarantees), the optimization phase requires only an oracle the fourth moments of the marginals $\langle x, u \rangle$ and not the data itself. Such an oracle can be implemented using the *fourth moment tensor* of the data, i.e., the values $\mathrm{E}[x_i x_j x_k x_l]$ for all $i, j, k, l \in \{1, \ldots, p\}$. By independence of the source variables, the fourth moment of $\langle x, u \rangle$ has local optima exactly at the columns of A^{-1}, and a variety of methods may be used to find these optima, including gradient descent or fixed-point methods. Once the orthonormal A^{-1} is recovered, the algorithms finally reverse the original affine transformation using the mean and covariance to obtain the original mixing matrix and source variables.

For privacy and utility, it therefore suffices to 1) release a sanitized (private) version

of the appropriately normalized first, second, and fourth moment tensors, 2) use the noisy fourth moment in the optimization phase of any ICA algorithm to recover the (approximate) columns of A^{-1}, and 3) reverse the original affine transformation using the noisy first and second moments.

Privacy

With this approach, determining how much noise to add for privacy reduces to analyzing the sensitivity of these tensors under a single row change. As noted in [40], the sensitivity of the covariance is tightly related to the *incoherence* of the data. One of our main technical contributions (Lemma 3.3.7) is that essentially the same is true of the fourth moment tensor, with an additional dependence on *conditioning* (see Section 3.2.3).

In proving the sensitivity of the fourth moment tensors of adequately conditioned databases, we observe that a change in one row of the database affects the isotropic fourth moment tensor via the row change itself and changes to the first and second moments (Inequality (3.3.4)). In particular, the hardest change to control is the composition of the square root of the new covariance matrix with the original fourth moment tensor. Roughly speaking, our main technical lemma (Lemma 3.3.8) shows that under a slight rotation of basis vectors, the *Frobenius* norm of the fourth moment tensor does not change substantially. Note that bounding the change in spectral norm from this rotation is trivial, but privacy requires a bound on Frobenius norm. A Frobenius norm bound directly obtained from a spectral norm bound would yield poly(p) loss. Instead, our approach is a coordinate-dependent calculation (though our result is of course coordinate-free), ultimately giving an optimal bound as a function of p (up to constant factors). The proof uses essentially elementary techniques, but requires a careful counting argument and a matrix perturbation bound from [68].

Because of the dependence of the fourth moment tensor's sensitivity on incoherence

and conditioning, privacy is given with respect to a definition of *neighboring databases* that bounds the amount a single row change can affect the incoherence and conditioning of the database (Definition 3.3.1). Although not as general as a neighborhood definition permitting arbitrary row change, our neighborhood definition is much more permissive than definitions that require the row change to be of constant Frobenius norm, in particular as required in [40].

This neighborhood definition allows us to employ the *Propose-Test-Release* paradigm of [26]: it (privately) checks whether the input database is adequately conditioned, and if so, it releases noisy first, second, and fourth moment tensors. Otherwise, it aborts with no output.

Utility

In arguing the utility of our mechanism for ICA, we face two main conceptual challenges. First, for arbitrary input data, which may not be well described by any ICA model, there is no canonical notion of the most accurate mixing matrix, nor one single objective function measuring how well a candidate mixing matrix fits the data. Second, existing ICA algorithms usually lack formal guarantees of output accuracy (i.e., how close their output comes to optimizing the objective), except sometimes in highly idealized settings where the data exactly conform to an ICA model. It is therefore unclear how to meaningfully quantify the accuracy of differentially private ICA, especially when the data are arbitrary.

One of our main contributions is a way of quantifying accuracy in an algorithm-agnostic way, and even in the absence of output guarantees in the non-private setting. Instead of analyzing the *output* of a particular ICA algorithm on its sanitized input, we consider the *objective function* that the algorithm attempts to optimize. We show that if the original data conform closely to some ICA *reference model* (Definition 3.4.1), then the optima of the sanitized fourth moment tensor are close to the columns of the

model's unmixing matrix.

To show accuracy of this type for reasonable amounts of noise, we exploit an arbitrage between the Frobenius and spectral norms. Utility loss depends on the *spectral* norm perturbation of our fourth moment tensors. Using tensor generalizations of some spectral norm bounds in random matrix theory, we show that we can add quite a large amount of noise (proportional to the Frobenius norm difference) to each entry of the isotropic fourth moment tensor without much effect on the spectral norm. In particular, the spectral norm of the noise tensor grows with $\sqrt{p}$, rather than p^2 for the Frobenius norm. Thus, we are able to provide surprisingly good privacy-utility tradeoffs.

Our framework and notion of utility can be directly applied to many of the recent ICA algorithms with provable guarantees [32, 76, 5, 3]. In Section 3.5, we specifically analyze the accuracy of our mechanism when used in conjunction with the recent SVD-based ICA algorithm of [2], where utility guarantees are not as obvious. We note that the only provably good algorithm for ICA which our methods do not extend to is the Fourier transform method of [36].

3.2 Preliminaries

In this chapter, all vectors are row vectors. For $x \in \mathbb{R}^p$, we let $x^{\otimes k}$ denote the $p \times \cdots \times p$ symmetric kth order tensor whose $(i_1, \ldots, i_k)$th entry is $x_{i_1} x_{i_2} \cdots x_{i_k}$, for all $i_1, \ldots, i_k \in [p]$. In particular, $x^{\otimes 2} = x \otimes x$ is the usual outer product, which gives a $p \times p$ matrix. For a kth order real tensor M, define the Frobenius norm $\|M\|_{\ell_2} := (\sum_{i_1,\ldots,i_k} M_{i_1,\ldots,i_k}^2)^{1/2}$, where the sum ranges over all indices of M. We let $\mathrm{Lap}(b)$ and $\mathcal{N}(\sigma^2)$ respectively refer to the (mean zero) Laplace and Gaussian distributions with probability density functions:

$$f_{\mathrm{Lap}}(x) := \frac{1}{2b} \exp(-|x|/b) \qquad f_{\mathcal{N}}(x) := \frac{1}{\sigma\sqrt{2\pi}} \exp(-x^2/(2\sigma^2))$$

3.2.1 Independent Component Analysis

ICA assumes data are generated as follows: the observed data $x = (x_1, \ldots, x_p) \in \mathbb{R}^p$ are generated by p fixed linear combinations of p unobserved independent random variables $s = (s_1, \ldots, s_p) \in \mathbb{R}^p$. These linear combinations are specified by a matrix $A \in \mathbb{R}^{p \times p}$ such that $x = sA$. An ICA model is therefore fully characterized by a *mixing matrix* A and a distribution for each of the *source variables* s_i. We call the corresponding observed random variable x an *ICA-generative random variable*. The statistical goal of ICA is to recover A and s from samples of x. Initially, x may have higher dimension than s, but preprocessing can reduce this to the case where A is square and non-singular. In general, the ICA problem is well-specified only if the source variables s_i are independent and at most one of them is Gaussian. If we normalize s so that $\mathrm{E}[s_i] = 0$ and $\mathrm{E}[s_i^2] = 1$ for each $i \in [p]$, then A is uniquely defined up a permutation of columns and their signs. Deducing A^{-1} allows us to decouple the observed random variables as $xA^{-1} = s$.

Standard ICA algorithms begin with a *whitening* step: the observed signals are placed in *isotropic* position by subtracting their means and applying a linear transformation given by an inverse square root of their centered covariance matrix. For a normalized ICA-generative random variable $x = sA$, the optima of the fourth moments of x are columns of A^{-1} (up to sign), and there are $2p$ such local optima. Several ICA algorithms use gradient descent to determine the columns of A^{-1}, sequentially. An important recent work [2] projects the $p \times p \times p \times p$ fourth moment tensor in such a way that the eigenvectors of a single $p \times p$ matrix give the columns of A^{-1} all at once.

In practice, an ICA algorithm's input is an isotropic fourth moment tensor of a database $X \in \mathbb{R}^{n \times p}$ whose rows are n samples of the observed variable x. We use the

following notation for a random row vector x or database X:

$$\mu_x := \mathrm{E}[x] \qquad \mu_X := \frac{1}{n}\sum_{i\in[n]} x_i$$

$$\Sigma_x := \mathrm{E}\left[(x-\mu_x)^{\otimes 2}\right] \qquad \Sigma_X := \frac{1}{n}\sum_{i\in[n]}(x_i-\mu_X)^{\otimes 2}$$

$$M_x := \mathrm{E}\left[\left((x-\mu_x)\Sigma_x^{-1/2}\right)^{\otimes 4}\right] \qquad M_X := \frac{1}{n}\sum_{i\in[n]}\left((x_i-\mu_X)\Sigma_X^{-1/2}\right)^{\otimes 4}$$

We refer to Σ_x, Σ_X as the *(centered) second moment tensors* or *covariance matrices*, and M_x, M_X as the *(isotropic) fourth moment tensors*, of the random variable x and database X, respectively.

3.2.1.1 Fourth Order Tensors

Any fourth order (not necessarily symmetric) tensor M defines a 4-linear form:

$$M(u,v,w,x) := \sum_{i,j,k,l} M_{ijkl}\, u_i v_j w_k x_l.$$

This form defines a spectral (or operator) norm:

$$\|M\|_{op} := \max_{u,v,w,x\in\mathbb{S}^{p-1}} M(u,v,w,x)$$

where $\mathbb{S}^{p-1}$ is the Euclidean unit sphere in p coordinates. We will let f denote the following quartic form associated with a fourth order tensor M:

$$f(u) := M(u,u,u,u) = \sum_{i,j,k,l} M_{ijkl}\, u_i u_j u_k u_l.$$

The quartic form also induces a spectral norm:

$$\|M\|_{op} := \max_{u\in\mathbb{S}^{p-1}} f(u).$$

When M is symmetric (i.e., the entries for every permutation of any four indices are the same), the spectral norms defined by the multilinear and quartic forms of M are the same. It is also easy to see that any fourth moment tensor of a random variable or of a database is symmetric.

We write $Df_u = \nabla f(u) \in \mathbb{R}^p$ for the gradient of f at $u \in \mathbb{R}^p$. When M is symmetric, we can express this gradient easily:

$$(Df_u)_l = 4 \sum_{ijk} M_{ijkl}\, u_i u_j u_k.$$

We define the linear form $Df_u(\cdot) = \langle Df_u, \cdot \rangle = 4M(u, u, u, \cdot)$; note therefore that $Df_u(u) = 4f(u)$.

The Hessian matrix of second derivatives of f at u is defined as $D^2 f_u = \nabla^2 f(u) \in \mathbb{R}^{p \times p}$. We write $\lambda_1(D^2 f_u)$ to refer to the largest eigenvalue of $D^2 f_u$, taken *orthogonal* to u.

3.2.2 Differential Privacy

Let $X \in \mathbb{R}^{n \times p}$ denote an n-row database. The definition of differential privacy [27, 25] is with respect to some definition of *neighboring databases* (most generally, those that differ arbitrarily in one row):

Definition 3.2.1. *For $\epsilon, \delta \geq 0$, a mechanism $\mathcal{M} : \mathbb{R}^{n \times p} \to Range(\mathcal{M})$ is* (ϵ, δ)-differentially private *if for all neighboring databases $X, Y \in \mathbb{R}^{n \times p}$ and for all subsets $S \subseteq Range(\mathcal{M})$,*

$$\Pr[\mathcal{M}(X) \in S] \leq \exp(\epsilon) \cdot \Pr[\mathcal{M}(Y) \in S] + \delta.$$

Typically, the first (and in our case, essentially the only) step in proving differential privacy for a mechanism is to determine the *sensitivity* of a query or class of queries on neighboring databases. We use two simple and powerful mechanisms that ensure differential privacy by adding noise proportional to sensitivity:

Proposition 3.2.2 (Laplace mechanism). *For any query $q \colon \mathbb{R}^{n \times p} \to \mathbb{R}$ with sensitivity $\Delta_q \in \mathbb{R}$ such that $|q(Y) - q(X)| \leq \Delta_q$ for all neighbors $X, Y \in \mathbb{R}^{n \times p}$, the* Laplace mechanism *is $(\epsilon, 0)$-differentially private:*

$$\mathcal{M}(X) := q(X) + Lap(\Delta_q / \epsilon).$$

When queries are multidimensional and (ϵ, δ)-differential privacy is sufficient, we add Gaussian noise:

Proposition 3.2.3 (Gaussian mechanism). *For any (T-dimensional) query* $q\colon \mathbb{R}^{n\times p} \to \mathbb{R}^T$ *with sensitivity* $\Delta_q \in \mathbb{R}$ *such that* $\|q(Y) - q(X)\|_{\ell_2} \le \Delta_q$ *for all neighbors* $X, Y \in \mathbb{R}^{n\times p}$*, the* Gaussian mechanism *with* $\sigma = \Delta_q\sqrt{\log(1.25/\delta)}/\epsilon$ *is* (ϵ, δ)*-differentially private:*

$$\mathcal{M}(X) := q(X) + \mathcal{N}(\sigma^2 I_T).$$

The following well-known lemma establishes that private mechanisms remain private under composition:

Proposition 3.2.4 (Composition). *Let* $\epsilon_i, \delta_i \ge 0$ *for* $i \in [k]$*. If each* $\mathcal{M}_i$ *is* (ϵ_i, δ_i)*-differentially private, then the algorithm* $\mathcal{M}(A) := (\mathcal{M}_1(A), \ldots, \mathcal{M}_k(A))$ *releasing the concatenated results of each algorithm is* $(\sum_{i\in[k]} \epsilon_i, \sum_{i\in[k]} \delta_i)$*-differentially private.*

3.2.3 Incoherence and Conditioning

To bound the sensitivity of moment tensors, we consider the *incoherence* ν_X and *condition parameter* κ_X of a database $X \in \mathbb{R}^{n\times p}$, defined as follows:

$$\nu_X := \frac{\max_{i\in[n]}\|x_i - \mu_X\|_{\ell_2}^2}{\|X - \mathbf{1}\mu_X\|_{\ell_2}^2} \qquad\qquad \kappa_X := \mathrm{Tr}(\Sigma_X)/\lambda_p(\Sigma_X)$$

Note that $\nu_X \in [1/n, 1]$, where a smaller value indicates greater incoherence: when $\nu_X = 1/n$, each row contributes an equal proportion of the total (Frobenius norm) weight of the database. Also note that every database satisfies $\max_{i\in[n]}\|x_i - \mu_X\|_{\ell_2}^2 = n \cdot \nu_X \cdot \mathrm{Tr}(\Sigma_X)$ since $\|X - \mathbf{1}\mu_X\|_{\ell_2}^2 = n \cdot \mathrm{Tr}(\Sigma_X)$.

Similarly, $\kappa_X \ge p$ since $\mathrm{Tr}(\Sigma_X) = \sum_{j\in[p]} \lambda_j(\Sigma_X)$. A small value of κ_X (close to p) indicates roughly equal eigenvalues. Note that κ_X is at most p times greater than $\lambda_1(\Sigma_X)/\lambda_p(\Sigma_X)$, the usual condition number. We consider a well-conditioned database to be one for which $\nu_X\kappa_X$ is sufficiently small. We discuss acceptable parameter ranges after presenting our main privacy result in Theorem 3.3.4.

3.3 A Private Mechanism for ICA

Our mechanism (Algorithm 1) computes and releases noisy versions of the sample mean, centered covariance, and isotropic fourth moment tensor of any adequately conditioned input database. Step 3 checks whether the database is adequately conditioned and aborts if not. This can be seen as an instance of the Propose-Test-Release paradigm of [26]. For adequately conditioned databases, the mechanism adds *symmetric* noise to the (symmetric) moment tensors, which will be important for our utility analysis, but it does not affect our application of the Gaussian mechanism for privacy.

Algorithm 1 Private ICA

Input: Database $X \in \mathbb{R}^{n \times p}$, parameters $\rho_\nu, \epsilon_0, \epsilon, \delta > 0$.
(Constants C_0, C^* are from Definition 3.3.1 and the proof of Lemma 3.3.8.)

1: Compute the incoherence and condition parameters ν_X and κ_X.
2: Draw γ_0 from $\mathrm{Lap}(1/(C_0 \cdot \rho_\nu \cdot \ln n))$.
3: If $\nu_X \kappa_X + \gamma_0 \geq 1/(2C^* \rho_\nu)$, then **abort** with no output.
4: Compute the mean μ_X, centered covariance Σ_X, and isotropic fourth moment M_X, and then compute σ_1, σ_2, and σ_4 satisfying the bounds in Theorem 3.3.4.
5: Let $\Gamma_1 \in \mathbb{R}^p$ be a random vector with independent entries drawn from $\mathcal{N}(\sigma_1^2)$.
6: Let $\Gamma_2 \in \mathbb{R}^{p \times p}$ be a random symmetric matrix with independent entries (up to symmetry) drawn from $\mathcal{N}(\sigma_2^2)$.
7: Let $\Gamma_4 \in \mathbb{R}^{p \times p \times p \times p}$ be a random symmetric tensor with independent entries (up to symmetry) drawn from $\mathcal{N}(\sigma_4^2)$.
8: **return** $(\mu_X + \Gamma_1, \Sigma_X + \Gamma_2, M_X + \Gamma_4)$.

The privacy of our mechanism is with respect to the following definition of neighboring databases. Two databases are neighbors under our definition if they differ in only one row and if this row difference does not affect the incoherence or condition parameters by too much:

Definition 3.3.1 (Neighborhood). *For parameters $\rho_\nu \geq 1, \epsilon_0 \geq 0$, two databases $X, Y \in \mathbb{R}^{n \times p}$ are* neighbors *if they differ in only one row and if:*

1. $\nu_Y / \nu_X \in [1/\rho_\nu, \rho_\nu]$, *and*
2. $|\nu_Y \kappa_Y - \nu_X \kappa_X| \leq \epsilon_0/(C_0 \cdot \rho_\nu \cdot \ln n)$ *for an absolute constant C_0 determined by Theorem 3.3.4.*

Our definition is more restrictive than the standard neighborhood definition from the differential privacy literature (e.g., [25]), under which *any* two databases differing in only one row are neighbors. We discuss the practical significance of this restriction after formalizing the main privacy guarantee of our mechanism in Theorem 3.3.4. First, we note that privacy of the abort step follows from the realization that the abort decision alone is an instance of the Laplace mechanism:

Fact 3.3.2 (from Proposition 3.2.2). *The abort decision in Step 3 of Algorithm 1 is ϵ_0-differentially private.*

The Laplace noise for privacy also ensures that this step is likely to reject inadequately conditioned databases:

Fact 3.3.3. *Algorithm 1 aborts on any database $X \in \mathbb{R}^{n \times p}$ with $\nu_X \kappa_X > 1/(C^* \rho_\nu)$ with all but $n^{-C_0/(2C^*)}$ probability (for C_0, C^* from the mechanism).*

These facts essentially reduce the proof of privacy to bounding the sensitivity (under our neighborhood definition) of moment tensors for databases that are adequately conditioned. The three lemmas that follow provide these bounds. With them, the following privacy guarantee for *all* databases is almost immediate:

Theorem 3.3.4 (Privacy). *For any desired $c \geq 0$, there exist universal constants C_1, C_2, C_4 such that for all sufficiently large n, Algorithm 1 is $(\epsilon_0 + 3\epsilon, 3\delta + n^{-c})$-differentially private under Definition 3.3.1 for*

$$\sigma_1 \geq C_1 \cdot \sqrt{\rho_\nu \cdot \nu_X \cdot \mathrm{Tr}(\Sigma_X)/n} \cdot \sqrt{\log(1.25/\delta)}/\epsilon \tag{3.3.1}$$

$$\sigma_2 \geq C_2 \cdot \rho_\nu \cdot \nu_X \cdot \mathrm{Tr}(\Sigma_X) \cdot \sqrt{\log(1.25/\delta)}/\epsilon \tag{3.3.2}$$

$$\sigma_4 \geq C_4 \cdot \Big(\rho_\nu \nu_X \kappa_X \cdot \|M_X\|_{\ell_2} + p^2 \sqrt{\rho_\nu \nu_X \kappa_X / n} \cdot \|M_X\|_{op} + (\rho_\nu \nu_X \kappa_X)^2 \cdot n \Big) \cdot \sqrt{\log(1.25/\delta)}/\epsilon. \tag{3.3.3}$$

Proof. Lemmas 3.3.5, 3.3.6, 3.3.7 bound the sensitivities of μ_X, Σ_X, M_X for databases X that satisy $\nu_X \kappa_X \leq 1/(C^* \rho_\nu)$. Restricting to such databases, the Gaussian mechanism and composition lemma show that it is $(3\epsilon, 3\delta)$-differentially private to release the noisy moment tensors with $\sigma_1, \sigma_2, \sigma_4$ as above. Fact 3.3.3 guarantees that any other database survives the abort step with probability at most $n^{-C_0/(2C^*)}$. Assuming that row changes in these databases have arbitrary effect on the moment tensors, this contributes additive error n^{-c} (for appropriate choice of neighborhood constant C_0) to the privacy guarantee. With Fact 3.3.2 and the composition lemma, we have our final privacy guarantee for all databases. □

Parameters for privacy and utility. The strength of our mechanism's privacy guarantee is quantified not only by $(\epsilon_0 + 3\epsilon, 3\delta + n^{-c})$ from the theorem above, but also by the extent to which our neighborhood definition permits large changes in a single row. We do not allow arbitrary row changes. For example, X and Y differing on only one row may *not* be neighbors under Definition 3.3.1 if the differing row in Y contributes a much greater proportion of the total magnitude of Y than the same row in X, or if the differing row in Y has large magnitude in the direction of an eigenvalue of Σ_X, because such changes could mean that even if X is well conditioned, Y may not be. Although certain single row changes are prohibited, our definition is significantly more permissive than the definition in the related work of [40], which requires $\|y_1 - x_1\|_{\ell_2} \leq 1$ for neighboring databases. For comparison, our definition with ρ_ν a small polynomial in p permits $\|y_1 - x_1\|_{\ell_2} = \text{poly}(p)$ when the database entries are constant size, κ_X is a (small) polynomial in p, and n is a large enough polynomial in p. Note that in this setting, the naive private mechanism that directly perturbs the input database X would require $\text{poly}(p)$ noise, which would destroy any utility.

In order for the utility bounds in Section 3.4 to be meaningful, we would like to

have privacy for some $\sigma_1, \sigma_2, \sigma_4 = o(1)$. Note that the third term in the σ_4 bound requires $\nu_X = o(1/(p \cdot \rho_\nu \cdot \sqrt{n}))$. If we additionally assume that X has $O(1)$ entries and that $\mathrm{Tr}(\Sigma_X)$ is poly(p), then all other terms are $o(1)$ assuming n is a sufficiently large polynomial in p. We must have $\rho_\nu = o(\sqrt{n}/p)$ for this bound on ν_X, so choosing ρ_ν a small polynomial in p as in the previous paragraph creates a permissive and feasible neighborhood definition. Finally, note that the mechanism is very unlikely to abort on databases with $\nu_X \kappa_X = o(1/(\rho_\nu \sqrt{n}))$, so the abort step does not restrict the set of databases for which we can provide meaningful ICA utility, beyond what we already require in our utility analysis.

Since we are releasing the covariance matrix, from which the singular values of X can be computed, our mechanism is subject to the lower bound of [41]. Their notion of coherence is slightly different from ours, and our neighborhood definition is more permissive. Roughly speaking, we are within small poly(p) factors of the lower bound, even while additionally sanitizing the fourth moment tensor.

For the remainder of this section, we give the bounds needed by our privacy theorem, Theorem 3.3.4, on the moment tensor sensitivities assuming the high probability bound $\nu_X \kappa_X \leq 1/(C^* \rho_\nu)$ given by Fact 3.3.3.

Lemma 3.3.5 (Sensitivity of means). *For any neighbors $X, Y \in \mathbb{R}^{n \times p}$ with $\nu_X \kappa_X \leq 1/(C^* \rho_\nu)$,*

$$\|\mu_Y - \mu_X\|_{\ell_2} = O\left(\sqrt{\rho_\nu \cdot \nu_X \cdot \mathrm{Tr}(\Sigma_X)/n}\right).$$

Proof. Note that $\mu_Y - \mu_X = \frac{1}{n}(y_1 - x_1)$. To bound this, we use the triangle inequality and individually bound the centered norms of x_1 and y_1 using coherence and the neighborhood definition.

$$\begin{aligned}
\|\mu_Y - \mu_X\|_{\ell_2} &\leq \frac{1}{n}\|(y_1 - \mu_Y) - (x_1 - \mu_X) + (\mu_Y - \mu_X)\|_{\ell_2} \\
&\leq \frac{1}{n-1}\left(\|y_1 - \mu_Y\|_{\ell_2} + \|x_1 - \mu_X\|_{\ell_2}\right) \\
&\leq \frac{1}{n-1}\left(\sqrt{\nu_Y \cdot n \cdot \mathrm{Tr}(\Sigma_Y)} + \sqrt{\nu_X \cdot n \cdot \mathrm{Tr}(\Sigma_X)}\right).
\end{aligned}$$

Our bound must be only with respect to X so that it can be used to set the noise parameter in our mechanism. To bound $\mathrm{Tr}(\Sigma_Y) = \|Y - \mathbf{1}^n\mu_Y\|_{\ell_2}^2/n$ in terms of database X, we use the above bound for $\|\mu_Y - \mu_X\|_{\ell_2}$, the incoherence condition, and two applications of the Cauchy-Schwartz inequality:

$$\begin{aligned}
\mathrm{Tr}(\Sigma_Y) &= \frac{1}{n}\|y_1 - \mu_Y\|_{\ell_2}^2 + \frac{1}{n}\sum_{i=2}^{n}\|(x_i - \mu_X) + (\mu_X - \mu_Y)\|_{\ell_2}^2 \\
&\leq \nu_Y \cdot \mathrm{Tr}(\Sigma_Y) + 2 \cdot \mathrm{Tr}(\Sigma_X) + \frac{2(n-1)}{n} \cdot \|\mu_Y - \mu_X\|_{\ell_2}^2 \\
&\leq \nu_Y \cdot \mathrm{Tr}(\Sigma_Y) + 2 \cdot \mathrm{Tr}(\Sigma_X) + \frac{2\rho_\nu\nu_X}{(n-1)} \cdot \left(\sqrt{\mathrm{Tr}(\Sigma_Y)} + \sqrt{\mathrm{Tr}(\Sigma_X)}\right)^2 \\
&\leq (1 + 4/(n-1)) \cdot \rho_\nu \cdot \nu_X \cdot \mathrm{Tr}(\Sigma_Y) + (2 + 4\rho_\nu\nu_X/(n-1)) \cdot \mathrm{Tr}(\Sigma_X) \\
&\leq \frac{2 + 4\rho_\nu\nu_X/(n-1)}{1 - \rho_\nu\nu_X - 4\rho_\nu\nu_X/(n-1)} \cdot \mathrm{Tr}(\Sigma_X).
\end{aligned}$$

Noting that $\kappa_X \geq p$ by definition, the conditioning assumption that $\nu_X\kappa_X \leq 1/(C^*\rho_\nu)$ is enough to bound the above coefficient of $\mathrm{Tr}(\Sigma_X)$ by some absolute constant. Our bound on the difference in means follows:

$$\begin{aligned}
\|\mu_Y - \mu_X\|_{\ell_2} &\leq \frac{\sqrt{n}}{n-1}\left(\sqrt{\rho_\nu \cdot \nu_X \cdot C \cdot \mathrm{Tr}(\Sigma_X)} + \sqrt{\nu_X \cdot \mathrm{Tr}(\Sigma_X)}\right) \\
&\leq C_1 \cdot \sqrt{\frac{\rho_\nu \cdot \nu_X \cdot \mathrm{Tr}(\Sigma_X)}{n}}.
\end{aligned}$$

□

Lemma 3.3.6 (Sensitivity of covariance). *For any neighbors $X, Y \in \mathbb{R}^{n\times p}$ with $\nu_X\kappa_X \leq 1/(C^*\rho_\nu)$,*

$$\|\Sigma_Y - \Sigma_X\|_{\ell_2} = O\left(\rho_\nu \cdot \nu_X \cdot \mathrm{Tr}(\Sigma_X)\right).$$

Proof. The difference $\Sigma_Y - \Sigma_X$ comprises an update in the first row and a mean update. We bound the Frobenius norm difference of these two updates separately.

$$\begin{aligned}
\Sigma_Y - \Sigma_X &= \frac{1}{n}\sum_{i=1}^{n}\left((y_i - \mu_Y)^{\otimes 2} - (x_i - \mu_X)^{\otimes 2}\right) \\
&= \frac{1}{n}\left((y_1 - \mu_Y)^{\otimes 2} - (x_1 - \mu_Y)^{\otimes 2}\right) + \frac{1}{n}\sum_{i=1}^{n}\left((x_i - \mu_Y)^{\otimes 2} - (x_i - \mu_X)^{\otimes 2}\right).
\end{aligned}$$

Now let $\delta_\mu := \mu_Y - \mu_X$, and note that

$$(x_i - \mu_Y)^{\otimes 2} = (x_i - \mu_X - \delta_\mu)^{\otimes 2} = (x_i - \mu_X)^{\otimes 2} - (x_i - \mu_X) \otimes \delta_\mu - \delta_\mu \otimes (x_i - \mu_X) + \delta_\mu^{\otimes 2}.$$

Then we can simplify the summation term of the covariance difference equation:

$$\begin{aligned}&\frac{1}{n}\sum_{i=1}^{n} \big((x_i - \mu_Y)^{\otimes 2} - (x_i - \mu_X)^{\otimes 2}\big) \\ &\qquad = \frac{1}{n}\sum_{i=1}^{n} \big(- (x_i - \mu_X) \otimes \delta_\mu - \delta_\mu \otimes (x_i - \mu_X) + \delta_\mu^{\otimes 2}\big) = \delta_\mu^{\otimes 2}.\end{aligned}$$

We similarly expand $(x_1 - \mu_Y)^{\otimes 2}$ in the row change term of the covariance difference equation to a sum of tensors of $x_1 - \mu_X$ and δ_μ. Then we complete the bound on covariance sensitivity using incoherence, the bound on mean sensitivity, and the earlier argument that $\operatorname{Tr}(\Sigma_Y) \le C \cdot \operatorname{Tr}(\Sigma_X)$ for an absolute constant C.

$$\begin{aligned}&\|\Sigma_Y - \Sigma_X\|_{\ell_2} \\ &\le \frac{1}{n}\|(y_1 - \mu_Y)^{\otimes 2} - (x_1 - \mu_X)^{\otimes 2} + (x_1 - \mu_X) \otimes \delta_\mu + \delta_\mu \otimes (x_1 - \mu_X) - \delta_\mu^{\otimes 2}\|_{\ell_2} + \|\delta_\mu^{\otimes 2}\|_{\ell_2} \\ &\le \frac{1}{n} \cdot C\rho_\nu \nu_X n \operatorname{Tr}(\Sigma_X) + \frac{1}{n} \cdot 2\sqrt{\nu_X n \operatorname{Tr}(\Sigma_X)} \cdot C_1 \sqrt{\frac{\rho_\nu \nu_X \operatorname{Tr}(\Sigma_X)}{n}} \\ &\qquad + (\frac{1}{n} + 1) \cdot C_1^2 \cdot \frac{\rho_X \nu_X \operatorname{Tr}(\Sigma_X)}{n} \\ &\le C_2 \cdot \rho_\nu \cdot \nu_X \cdot \operatorname{Tr}(\Sigma_X).\end{aligned}$$

□

Lemma 3.3.7 (Sensitivity of fourth moment). *For any neighbors $X, Y \in \mathbb{R}^{n \times p}$ with $\nu_X \kappa_X \le 1/(C^* \rho_\nu)$,*

$$\|M_Y - M_X\|_{\ell_2} \le O\big(\rho_\nu \nu_X \kappa_X \cdot \|M_X\|_{\ell_2} + p^2 \sqrt{\rho_\nu \nu_X \kappa_X / n} \cdot \|M_X\|_{op} + (\rho_\nu \nu_X \kappa_X)^2 \cdot n\big).$$

Proof. Denote the fourth moment tensor of a database X shifted by some means vector (not necessarily its own) μ and scaled by some covariance matrix Σ as follows:

$$M_{(X,\mu,\Sigma)} := \frac{1}{n}\sum_{i \in [n]} \big((x_i - \mu)\Sigma^{-1/2}\big)^{\otimes 4}.$$

Assume without loss of generality that a neighboring database Y differs on the first row. We decompose the Frobenius norm of the change between M_X and M_Y into the sum of changes accumulated in three steps:

$$\begin{aligned}\|M_Y - M_X\|_{\ell_2} \;\leq\; & \|M_{(X,\mu_Y,\Sigma_X)} - M_X\|_{\ell_2} \\ & + \|M_{(Y,\mu_Y,\Sigma_X)} - M_{(X,\mu_Y,\Sigma_X)}\|_{\ell_2} \\ & + \|M_Y - M_{(Y,\mu_Y,\Sigma_X)}\|_{\ell_2}. \end{aligned} \tag{3.3.4}$$

The first term is the change due to updating M_X to be centered with respect to the column means of database Y. The second term is the change due to updating the single differing row from x_1 to y_1. At this stage, both databases are centered by μ_Y and scaled with Σ_X. The last term is the change due to rescaling the whole μ_Y-centered database with Σ_Y instead of Σ_X.

Letting $\delta_\mu = \mu_Y - \mu_X$, we bound the second term of Expression (3.3.4) using incoherence and the earlier lemma for mean sensitivity:

$$\begin{aligned}\|M_{(Y,\mu_Y,\Sigma_X)} - M_{(X,\mu_Y,\Sigma_X)}\|_{\ell_2} &= \frac{1}{n}\|((y_1-\mu_Y)\Sigma_X^{-1/2})^{\otimes 4} - ((x_1-\mu_Y)\Sigma_X^{-1/2})^{\otimes 4}\|_{\ell_2} \\ &\leq \frac{1}{n}(\|y_1-\mu_Y\|_{\ell_2}^4 + \|x_1-\mu_X-\delta_\mu\|_{\ell_2}^4)\|\Sigma_X^{-1}\|_{op}^2 \\ &= \frac{1}{n}\cdot O((\rho_\nu\cdot\nu_X\cdot n\cdot \mathrm{Tr}(\Sigma_X))^2)/\lambda_p(\Sigma_X)^2 \\ &= O((\rho_\nu\nu_X\kappa_X)^2\cdot n).\end{aligned}$$

Now it is enough to bound the first and third terms of Expression (3.3.4). We first bound the third term in Lemma 3.3.8, which uses a result from [68] to show that the rescaling matrix $\Sigma_Y^{1/2}\Sigma_X^{-1/2}$ is close to the identity and then uses Proposition 3.3.9, which proves that a near identity perturbation results in only a small Frobenius norm change in fourth moment tensor. Then after bounding the first term in Lemma 3.3.10, the fourth moment sensitivity follows almost immediately using triangle inequality and the observation that $1/(1-C^*\rho_\nu\nu_X\kappa_X/2) = O(1)$ when $\nu_X\kappa_X \leq 1/(C^*\rho_\nu)$.

Lemma 3.3.8 (Rescaling). *For any neighbors $X, Y \in \mathbb{R}^{n\times p}$ with $\nu_X \kappa_X < 1/(C^*\rho_\nu)$,*

$$\|M_Y - M_{(Y,\mu_Y,\Sigma_X)}\|_{\ell_2} \le \frac{C^*\rho_\nu\nu_X\kappa_X}{2} \cdot \|M_Y\|_{\ell_2}.$$

Proof. Let $T := \Sigma_Y^{1/2}\Sigma_X^{-1/2}$. With some calculation, we can see that for $u \in \mathbb{R}^p$:

$$M_{(Y,\mu_Y,\Sigma_X)}(u,u,u,u) = M_Y(Tu,Tu,Tu,Tu).$$

For any symmetric fourth order tensor M we have $\|M\|_{\ell_2}^2 = \sum_{ijkl} M(e_i,e_j,e_k,e_l)^2$. In particular,

$$\|M_Y - M_{(Y,\mu_Y,\Sigma_X)}\|_{\ell_2}^2 = \sum_{ijkl}(M_Y(e_i,e_j,e_k,e_l) - M_Y(Te_i,Te_j,Te_k,Te_l))^2. \quad (3.3.5)$$

Lemma 3.3.6 already tells us that neighboring databases have close centered covariance matrices, so we expect $T := \Sigma_Y^{1/2}\Sigma_X^{-1/2} = I + E$ with $\|E\|_{\ell_2}$ small. We give an explicit bound for $\|E\|_{\ell_2}$ using the following result, derived from Mathias [68]:

Theorem (from [68]). *Let H be a $p \times p$ Hermitian matrix that is positive definite. Then for $\eta > 0$ and Hermitian G such that $\|H^{-1/2}GH^{-1/2}\|_{\ell_2} = \eta$, we have:*

$$\|(H+G)^{1/2}H^{-1/2} - I\|_{\ell_2} \le \eta + O(\eta^2).$$

For neighbors $X, Y \in \mathbb{R}^{n\times p}$, let $H = \Sigma_X$ and $G = \Sigma_Y - \Sigma_X = \frac{1}{n}\left((y_1-\mu_Y)^{\otimes 2} - (x_1-\mu_X)^{\otimes 2}\right)$, and then bound η as follows:

$$\begin{aligned}
\eta &= \frac{1}{n}\|((y_1-\mu_Y)\Sigma_X^{-1/2})^{\otimes 2} - ((x_1-\mu_X)\Sigma_X^{-1/2})^{\otimes 2}\|_{\ell_2} \\
&\le \frac{1}{n}\left(\|(y_1-\mu_Y)\Sigma_X^{-1/2}\|_{\ell_2}^2 + \|(x_1-\mu_X)\Sigma_X^{-1/2}\|_{\ell_2}^2\right) \\
&\le \frac{1}{n}(\|y_1-\mu_Y\|_{\ell_2}^2 + \|x_1-\mu_X\|_{\ell_2}^2)/\lambda_p(\Sigma_X) \\
&\le (C\cdot\rho_\nu + 1)\cdot\nu_X\kappa_X.
\end{aligned}$$

Now write $T = I + E$ with $\|E\|_{\ell_2} \le (C\cdot\rho_\nu+1)\cdot\nu_X\kappa_X + O((C\cdot\rho_\nu+1)^2\cdot\nu_X^2\kappa_X^2)$. Given the universal constant C_P from the following proposition, we can choose a

universal constant C^* sufficiently large relative to the other constants so that the assumption $\nu_X \kappa_X \leq 1/(C^* \rho_\nu)$ implies $\|E\|_{\ell_2} \leq \frac{C^*}{2\sqrt{C_P}} \cdot \rho_\nu \nu_X \kappa_X \leq \frac{1}{2\sqrt{C_P}} < 1$. Then we can apply Proposition 3.3.9 below using M_Y for M and $\Sigma_Y^{1/2} \Sigma_X^{-1/2}$ for T to bound Equation (3.3.5) using the above bound $\|E\|_{\ell_2} \leq \frac{C^*}{2\sqrt{C_P}} \rho_\nu \nu_X \kappa_X$, completing the proof of Lemma 3.3.8. □

Proposition 3.3.9 (for Lemma 3.3.8). *There exists some universal constant $C_P \geq 1$ such that for any symmetric fourth order tensor $M \in \mathbb{R}^{p \times p \times p \times p}$ and any $T = I + E \in \mathbb{R}^{p \times p}$ with $\|E\|_{\ell_2} \leq 1$,*

$$\sum_{ijkl} (M(Te_i, Te_j, Te_k, Te_l) - M(e_i, e_j, e_k, e_l))^2 \leq C_P \cdot \|E\|_{\ell_2}^2 \|M\|_{\ell_2}^2.$$

Proof. First observe that for fixed i, j, k, l,

$$M(Te_i, Te_j, Te_k, Te_l) - M(e_i, e_j, e_k, e_l) = \Big(\sum_{i'j'k'l'} T_{i'i} T_{j'j} T_{k'k} T_{l'l} M_{i',j',k',l'} \Big) - M_{ijkl}. \tag{3.3.6}$$

By assumption, the diagonal terms of T are the only large ones. In particular, we can liberally use the coarse bound $E_{ij} \leq \|E\|_{\ell_2} \leq 1$ to control and simplify many of the coefficients of entries of M in Equation (3.3.6). Throughout, C, C', C'' are absolute constants that may differ across expressions.

Let $\mathcal{I}_i$ denote the set of 4-tuples $(i'jkl)$ with $i' \neq i$, and similarly for $\mathcal{I}_j, \mathcal{I}_k, \mathcal{I}_l$; let $\mathcal{I}_{ij}$ denote the set of $(i'j'kl)$ with $i' \neq i, j' \neq j$, and so on; let $\mathcal{I}_{ijk}$ denote the (i', j', k', l) with $i' \neq i, j' \neq j, k' \neq k$, and so on; let $\mathcal{I}_{ijkl}$ denote the (i', j', k', l') with $i' \neq i, j' \neq j, k' \neq k, l'$. Let $\mathscr{I}$ denote the constant size disjoint union of all such $\mathcal{I}$. Now apply the Cauchy-Schwartz inequality to Equation (3.3.6) to separately consider the contribution of each $\mathcal{I} \in \mathscr{I}$:

$$\sum_{ijkl}(M(Te_i,Te_j,Te_k,Te_l)-M(e_i,e_j,e_k,e_l))^2$$
$$\leq C\cdot\sum_{ijkl}(T_{ii}T_{jj}T_{kk}T_{ll}-1)^2M_{ijkl}^2+C\cdot\sum_{ijkl}\sum_{\mathcal{I}\in\mathscr{I}}(\sum_{(i'j'k'l')\in\mathcal{I}}T_{i'i}T_{j'j}T_{k'k}T_{l'l}M_{i'j'k'l'})^2.$$

It is easy to bound the first term using $T=I+E$:

$$\begin{aligned}C\cdot\sum_{ijkl}(T_{ii}T_{jj}T_{kk}T_{ll}-1)^2M_{ijkl}^2&=C\cdot\sum_{ijkl}((1+E_{ii})(1+E_{jj})(1+E_{kk})(1+E_{ll})-1)^2M_{ijkl}^2\\&\leq C'\cdot\sum_{ijkl}(|E_{ii}|+|E_{jj}|+|E_{kk}|+|E_{ll}|)^2M_{ijkl}^2\\&\leq C''\cdot\|E\|_{\ell_2}^2\|M\|_{\ell_2}^2.\end{aligned}$$

Now consider the second term for fixed $ijkl$. Let $E_{\cdot i}$ denote the ith column of E and let $M_{\cdot jkl}$ denote the vector whose ith entry is M_{ijkl}. The contribution of the $(i'jkl)\in\mathcal{I}_i$ is as follows.

$$\begin{aligned}C\cdot(\sum_{(i'jkl)\in\mathcal{I}_i}T_{i'i}T_{jj}T_{kk}T_{ll}M_{i'jkl})^2&=C\cdot(\sum_{i'\neq i}E_{i'i}(1+E_{jj})(1+E_{kk})(1+E_{ll})M_{i'jkl})^2\\&\leq C'\cdot(\sum_{i'\in[p]}E_{i'i}M_{i'jkl})^2\\&=C'\cdot\langle|E_{\cdot i}|,M_{\cdot jkl}\rangle^2\\&\leq C''\cdot\|E_{\cdot i}\|_{\ell_2}^2\|M_{\cdot jkl}\|_{\ell_2}^2.\end{aligned}$$

Summing over all $ijkl$, we have

$$C\cdot\sum_{ijkl}(\sum_{(i'jkl)\in\mathcal{I}_i}T_{i'i}T_{jj}T_{kk}T_{ll}M_{i'jkl})^2\leq C''\cdot\|E\|_{\ell_2}^2\|M\|_{\ell_2}^2.$$

We can similarly do this for $j'\neq i,k'\neq k,l'\neq l$. Analogously, sets with two, three, and four indices differing from $ijkl$ respectively contribute $C'\cdot\|E\|_{\ell_2}^4\|M\|_{\ell_2}^2$, $C'\cdot\|E\|_{\ell_2}^6\|M\|_{\ell_2}^2$, and $C'\cdot\|E\|_{\ell_2}^8\|M\|_{\ell_2}^2$. With $\|E\|_{\ell_2}\leq 1$, these higher order terms are dominated by the lower order terms, and we get our final bound. □

We finally bound the first term of Expression (3.3.4), the portion of the difference in neighboring isotropic fourth moment tensors due to the databases being centered with respect to different means.

Lemma 3.3.10 (Recentering). *For any neighbors $X, Y \in \mathbb{R}^{n\times p}$ with $\nu_X \kappa_X < 1/(C^* \rho_\nu)$,*

$$\|M_{(X,\mu_Y,\Sigma_X)} - M_X\|_{\ell_2} = O(p^2 \cdot \sqrt{\rho_\nu \nu_X \kappa_X / n} \cdot \|M_X\|_{op}).$$

Proof. Let $\delta_\mu = \mu_Y - \mu_X$ as before, and define $\hat{x}_i := (x_i - \mu_X)\Sigma_X^{-1/2}$ and $\hat{\delta}_\mu := \delta_\mu \Sigma_X^{-1/2}$. Then:

$$\begin{aligned} M_{(X,\mu_Y,\Sigma_X)} - M_X &= \frac{1}{n}\sum_{i\in[n]} \left[((x_i - \mu_X - \delta_\mu)\Sigma_X^{-1/2})^{\otimes 4} - ((x_i - \mu_X)\Sigma_X^{-1/2})^{\otimes 4}\right] \\ &= \frac{1}{n}\sum_{i\in[n]} \left[(\hat{x}_i - \hat{\delta}_\mu)^{\otimes 4} - \hat{x}_i^{\otimes 4}\right]. \end{aligned}$$

Each term in the summand can be written as the sum of 4-way outer products of $\hat{\delta}_\mu$ and $\hat{x}_i$ that each have $\hat{\delta}_\mu$ as at least one of the outer product terms. We can therefore decompose this expression into the sum of four symmetric tensors, each of which is a sum of these outer product terms. Let M_1 denote the tensor given by the sum of the four types of outer products that have one copy of $\hat{\delta}_\mu$ and three of $\hat{x}_i$. Denote the sums of the outer product terms with 2, 3, and 4 copies of $\hat{\delta}_\mu$ by M_2, M_3, M_4, respectively. Then

$$M_{(X,\mu_Y,\Sigma_X)} - M_X = M_1 + M_2 + M_3 + M_4.$$

Let $\hat{x}$ denote the random variable that takes value $\hat{x}_i$ with probability $1/n$ for each $i \in [n]$. Since $\hat{x}$ is an isotropic variable, its covariance is I, so for $u \in \mathbb{S}^{p-1}$, $1 \le \mathrm{E}\left[|\hat{x}u^\top|^2\right] \le \mathrm{E}\left[|\hat{x}u^\top|^4\right]^{1/2} = \|M_X\|_{op}^{1/2}$ by Hölder's inequality. Similarly, for $u \in \mathbb{S}^{p-1}$ and $s = 1, 2, 3$ we have:

$$\mathrm{E}\left[|\hat{x}u^\top|^s\right] \le \mathrm{E}\left[|\hat{x}u^\top|^4\right]^{s/4} \le \|M_X\|_{op}^{s/4} \le \|M_X\|_{op}.$$

Now $\|M_1\|_{op} \leq 4\|\hat{\delta}_\mu\|_{op}\|M_X\|_{op}$, because for any $u \in \mathbb{S}^{p-1}$,

$$\begin{aligned}
|M(u,u,u,u)| &\leq \frac{1}{n}\sum_{i=1}^{n} 4|\hat{\delta}_\mu u^\top||\hat{x}_i u^\top|^3 \\
&\leq 4\|\hat{\delta}_\mu\|_{op} \cdot \mathrm{E}\left[|\hat{x}u^\top|^3\right] \\
&\leq 4\|\hat{\delta}_\mu\|_{op}\|M_X\|_{op}.
\end{aligned}$$

We similarly bound the operator norms of the other tensors:

$$\begin{aligned}
\|M_2\|_{op} &\leq 6\|\hat{\delta}_\mu\|_{op}^2 \,\mathrm{E}\left[|\hat{x}u^\top|^2\right] \leq 6\|\hat{\delta}_\mu\|_{op}^2\|M_X\|_{op}, \\
\|M_3\|_{op} &\leq 4\|\hat{\delta}_\mu\|_{op}^3 \,\mathrm{E}\left[|\hat{x}u^\top|\right] \leq 4\|\hat{\delta}_\mu\|_{op}^3\|M_X\|_{op}, \\
\|M_4\|_{op} &\leq \|\hat{\delta}_\mu\|_{op}^4 \leq \|\hat{\delta}_\mu\|_{op}^4\|M_X\|_{op}.
\end{aligned}$$

By sensitivity of means, we have

$$\|\hat{\delta}_\mu\|_{op} \leq \|\delta_\mu\|_{\ell_2}\|\Sigma_X^{-1}\|_{op}^{1/2} = O(\sqrt{\rho_\nu \nu_X \kappa_X / n}).$$

For $\nu_X \kappa_X \leq 1/(C^* \rho_\nu)$, the higher order terms can be absorbed into a larger constant in the first term:

$$\begin{aligned}
\|M_{(X,\mu_Y,\Sigma_X)} - M_X\|_{op} &\leq \|M_1\|_{op} + \|M_2\|_{op} + \|M_3\|_{op} + \|M_4\|_{op} \\
&= O(\|\hat{\delta}_\mu\|_{op}\|M_X\|_{op}) \\
&= O(\sqrt{\rho_\nu \nu_X \kappa_X / n} \cdot \|M_X\|_{op}).
\end{aligned}$$

Finally, we use the very crude bound that no entry of a symmetric tensor can be greater than the spectral norm to get our bound on Frobenius norm. □

With the previous bounds on change in fourth moment tensor due to row change, mean change, and rescaling, we bound the change in fourth moment tensor as follows:

$$\begin{aligned}
\|M_Y - M_X\|_{\ell_2} &\leq \|M_{(X,\mu_Y,\Sigma_X)} - M_X\|_{\ell_2} + \|M_{(Y,\mu_Y,\Sigma_X)} - M_{(X,\mu_Y,\Sigma_X)}\|_{\ell_2} + \|M_Y - M_{(Y,\mu_Y,\Sigma_X)}\|_{\ell_2} \\
&\leq C' \cdot p^2\sqrt{\rho_\nu \nu_X \kappa_X / n} \cdot \|M_X\|_{op} + C'' \cdot (\rho_\nu \nu_X \kappa_X)^2 \cdot n + C''' \cdot \rho_\nu \nu_X \kappa_X \cdot \|M_Y\|_{\ell_2}
\end{aligned}$$

Note that $\|M_Y\|_{\ell_2} \leq \|M_X\|_{\ell_2} + \|M_Y - M_X\|_{\ell_2}$. To replace M_Y with M_X in the asymptotic bound of Lemma 3.3.7, we substitute this triangle inequality and then manipulate the above inequality in the natural way, observing that $C/(1 - C'''\rho_\nu\nu_X\kappa_X) = O(1)$ given the conditioning assumption that $\nu_X\kappa_X \leq 1/C^*\rho_\nu$ for sufficiently large constant C^*. Lemma 3.3.7 follows. □

3.4 Reference Models and Utility

We note that the vast literature on ICA has little to say about the interpretation of the output of ICA algorithms when input datasets are not well-described by an ICA model. For example, the isotropic fourth moments of such datasets may have many more than $2p$ local optima, so the recovered mixing matrix need not be unique up to rotation, and it is unclear that the output of an ICA algorithm run on such data provides any meaningful information about the database. We therefore prove the utility of our mechanism with respect to some generative ICA *reference model*. If there exists such a model which describes our original database well, then the model also describes the database's noisy fourth moment tensor well.

Definition 3.4.1 ((α, β)-reference model). *We say that a p-dimensional real random row vector v is an* (α, β)-reference model *with orthonormal mixing matrix $A \in \mathbb{R}^{p\times p}$ for database $X \in \mathbb{R}^{n\times p}$ if:*

1. *$(v - \mu_v)\Sigma_v^{-1/2} = sA$ for independent, isotropic, non-Gaussian, p-dimension real random vector s,*
2. *$\|\Sigma_X - \Sigma_v\|_{op} \leq \alpha$, and*
3. *$\|M_X - M_v\|_{op} \leq \beta$.*

For fixed (α, β), there may be an entire family of reference models v; we will guarantee utility with respect to the mixing matrix underlying *any* particular choice of (α, β)-reference model. Note that if a database is sampled from an ICA-generative model

v, then the strong law of large numbers guarantees almost sure convergence of the sample second and fourth moments to the model moments. The error norms in this definition are spectral norms, which is a weaker condition than the Frobenius norm and coheres with much of the existing literature on the accuracy of sampled moments [86, 93, 95, 1, 39].

The eventual output of ICA run on samples of a random vector generated from an ICA model is a set of vectors that give the columns of the underlying unmixing matrix A^{-1}, the columns of which are the local optima of the isotropic fourth moment of the samples. To accomplish this, many ICA algorithms approximate a *contrast function* [23], such as the fourth moment tensor, using the samples given, and then they iteratively find a set of local optima to this contrast function using some orthogonalization scheme to ensure that the local optima are approximately orthogonal. The orthogonalization component of this process presents formidable technical challenges. Practitioners often simply assume that the orthogonalization scheme is effective and does not incur too much error.

In order to give guarantees that are agnostic to a particular choice of fourth moment tensor-based ICA algorithms, we do not consider the role of orthogonalization, but instead we give guarantees for each approximate local optima of our noisy fourth moment tensor relative to a close row of the unmixing matrix associated with some reference model. Definition 3.4.2 formally states our notion of an approximate local optimum of a noisy tensor, recalling the notation from Section 3.2.1.

Definition 3.4.2 ((ϵ_1, ϵ_2)-approximate local optimum). *For the quartic form f associated with $M \in \mathbb{R}^{p\times p\times p\times p}$, we say $u \in \mathbb{S}^{p-1}$ is an (ϵ_1, ϵ_2)-approximate local optimum of M if*

$$\langle Df_u, u\rangle \geq \|Df_u\|_{\ell_2} - \epsilon_1 \quad \textit{and} \quad \lambda_1(D^2 f_u) \leq 9f(u)/p + \epsilon_2.$$

This definition is motivated by the optimization problem of maximizing $f(u,u,u,u)$ over unit u. Writing out the Lagrangian $\mathcal{L} = f(u,u,u,u) - \lambda(\sum_i u_i^2 - 1)$ and computing the first order conditions implies that $\langle Df_u, u\rangle = \|Df_u\|_{\ell_2}$ (i.e., the derivative at a point u is parallel to u at all stationary points). For u to be an approximate local optimum, we require the first order condition to be satisfied approximately, and the second order condition will guarantee that u is a not a saddle point, but rather a local maximum or minimum. Note that even in the *exact* case, the first order condition $\langle Df_u, u\rangle = \|Df_u\|_{\ell_2}$ insufficiently characterizes approximate local optima. In the generative model $x_i = s_i$ where s_i is a uniform over $[0,1]$, for example, $u = (e_1+e_2)/\sqrt{2}$ satisfies the first order conditions but is not a local optimum. Lemma 3.4.4 from [94] shows that in the exact case, the second order condition is enough to characterize approximate local optima.

Our utility theorem states that every approximate local optimum of the mechanism's output is close to some column of the unmixing matrix of the reference ICA model, both in the isotropic setting and in the original setting, and furthermore that every column of the unmixing matrix is an approximate local optimum of the mechanism's output. In other words, given a reference model v with mixing matrix A that is close to database X, a vector is an approximate optimum of our mechanism's private fourth moment tensor for X *if and only if* it is an approximate column of A^{-1}.

Theorem 3.4.3 (Utility). *Let $(\mu_X + \Gamma_1, \Sigma_X + \Gamma_2, M_X + \Gamma_4)$ be the output of the private ICA mechanism on database X with noise parameters $\sigma_1, \sigma_2, \sigma_4$, and let random vector v be an (α,β)-reference model for X with mixing matrix $A \in \mathbb{R}^{p\times p}$. If $\beta \le c'/p$ and $\sigma_4 \le c''/(p\sqrt{p+\log 1/\delta})$ for appropriate universal constants $c', c'' > 0$ and $\delta \in (0,1)$, then with probability at least $1-\delta$:*

- *For every $(\epsilon, 0)$-approximate local optimum $u \in \mathbb{S}^{p-1}$ of $M_X + \Gamma_4$, there exists a column $a^\top$ of A^{-1} such that for some choice of $\epsilon' = O(\epsilon + \beta + \sigma_4\sqrt{p+\log 1/\delta})$:*

  $|\langle u, a\rangle| \ge 1 - \epsilon'$, *and*

$$\|u(\Sigma_X + \Gamma_2)^{1/2} - a\Sigma_v^{1/2}\|_{\ell_2} \leq \lambda_1(\Sigma_v^{1/2})(\alpha + O(\sigma_2\sqrt{p}) + \epsilon' + 2\sqrt{\epsilon'}).$$

- *Every column of* A^{-1} *is an* (ϵ_1, ϵ_2)*-approximate local optimum of* $M_X + \Gamma_4$ *for some choice of*

  $\epsilon_1 = O(\beta + \sigma_4\sqrt{p + \log 1/\delta})$ *and* $\epsilon_2 = O(\beta + \sigma_4\sqrt{p + \log 1/\delta})$.

Proof. Throughout the proof, let f and g be the quartic forms associated with $M_X + \Gamma_4$ and M_v, respectively. Our proof uses two new lemmas that follow. Lemma 3.4.5 shows that a symmetric fourth order tensor with Gaussian entries is likely to have small spectral norm. Lemma 3.4.6 shows that symmetric fourth order tensors that are close in spectral norm have close local optima. The proof that every column of A^{-1} is an approximate local optimum of $M_X + \Gamma_4$ is straightforward with these results. To show the converse, we relate the approximate local optima to the mixing matrix of the underlying reference model using a result from [94]. Finally, a result from [68] (also used in the proof of Lemma 3.3.8) allows us to convert these results to the coordinates of the original, non-isotropic data.

By Lemma 3.4.5 and triangle inequality, we have that with probability all but δ,

$$\|M_v - (M_X + \Gamma_4)\|_{op} \leq \beta + c\sigma_4\sqrt{p + \log 1/\delta}.$$

Then if we let u be an $(\epsilon, 0)$-approximate local optimum of $M_X + \Gamma_4$, Lemma 3.4.6 shows that u is an approximate local optimum of M_v with:

$$\langle Dg_u, u\rangle \geq \|Dg_u\|_{\ell_2} - \epsilon - 8(\beta + c\sigma_4\sqrt{p + \log 1/\delta}),$$
$$\lambda_1(D^2 g_u) \leq \frac{9g(u)}{p} + (12 + 9/p)(\beta + c\sigma_4\sqrt{p + \log 1/\delta}).$$

The following result from [94], immediately extended from 2- to p-component subspaces, relates approximate local optima of ICA-generative fourth moments to columns of the unmixing matrix:

Lemma 3.4.4 (from [94]). *Let* g *be the quartic form associated with the isotropic fourth moment of a random variable generated from an ICA model. Let* u *be a unit*

vector such that $\langle Dg_u, u\rangle \geq (1-\epsilon')\|Dg_u\|_{\ell_2}$ *and* $\lambda_1(D^2 g_u) \leq \frac{12g(u)}{p}$. *Then* g *has a local optimum* a *(which is a column of the model's unmixing matrix) with* $|\langle u, a\rangle| \geq 1 - 16\epsilon'$.

In the isotropic setting, the fourth moment in any direction is at least 1, so $\|Dg_u\|_{\ell_2} \geq 4g(u) \geq 4$. Therefore, we may simply turn our additive error into the required multiplicative error in the first order condition at the gain of a small constant. The second order hypothesis is satisfied with our assumptions bounding β and σ_4. Thus, we have that for any $(\epsilon, 0)$-approximate local optimum $u \in \mathbb{S}^{p-1}$ of $M_X + \Gamma_4$, there exists a column $a^\top$ of the unmixing matrix A^{-1} for the (α, β)-reference model v for X and an absolute constant c such that:

$$|\langle u, a\rangle| \geq 1 - c(\epsilon + \beta + \sigma_4\sqrt{p + \log 1/\delta}).$$

Therefore, each approximate local optimum of $M_X + \Gamma_4$ is close in angle to some column of A^{-1}.

The result from [68] cited in the proof of Lemma 3.3.8 directly implies the following, which allows us to put our utility guarantee in terms of the raw (non-isotropic) data:

Theorem (from [68]). *Let* H *and* ΔH *be* $n \times n$ *Hermitian matrices such that* H *is positive definite:*

$$\|(H + \Delta H)^{1/2} - H^{1/2}\|_{op} \leq \|H\|_{op}^{1/2}\|\Delta H\|_{op}$$

Since $|\langle u, a\rangle| \geq 1 - \epsilon'$ for $\epsilon' = c(\epsilon + \beta + \sigma_4\sqrt{p + \log 1/\delta})$, we have $u - a = \epsilon' + r$ with $\|r\|_{\ell_2} \leq \sqrt{2\epsilon' - \epsilon'^2}$. Then we can get our final bound with the triangle inequality:

$$\begin{aligned}\|u(\Sigma_X + \Gamma_2)^{1/2} - a\Sigma_v^{1/2}\|_{\ell_2} &\leq \|u(\Sigma_X + \Gamma_2)^{1/2} - u\Sigma_v^{1/2}\|_{\ell_2} + \|u\Sigma_v^{1/2} - a\Sigma_v^{1/2}\|_{\ell_2} \\ &\leq \lambda_1(\Sigma_v^{1/2})\left(\alpha + c''''\sigma_2\sqrt{p} + \epsilon' + 2\sqrt{\epsilon'}\right). \qquad \square\end{aligned}$$

In Lemma 3.4.5, we show that adding Gaussian noise to the isotropic sample fourth moment tensor is unlikely to introduce too much error in the spectral norm using a tensor version of a standard ϵ-net argument for bounding spectral norm (see for example [96]):

Lemma 3.4.5 (Random tensor). *Let Γ be a random symmetric fourth order tensor whose (distinct) entries are drawn from $\mathcal{N}(0, \sigma^2)$. Then $\|\Gamma\|_{op} \leq c\sigma\sqrt{p + \log 1/\delta}$ with probability at least $1 - \delta$ for $\delta \in (0, 1)$ and sufficiently large absolute constant $c > 0$.*

Proof. Observe that a symmetric random tensor with distinct entries drawn from $\mathcal{N}(0, \sigma^2)$ can be written as $\Gamma = \Gamma_1 + \cdots + \Gamma_{4!}$, where the entries in each Γ_i are independent of each other but Γ_i and Γ_j are not independent of each other. If for Γ_i with some zero entries and all other entries independent from $\mathcal{N}(0, \sigma^2)$, we have $\|\Gamma_i\|_{op} \leq t$ with all but probability $\delta/4!$, then by union bound, $\|\Gamma\|_{op} \leq t$ with all but δ probability. It therefore suffices to analyze a fully independent tensor.

Let $N_\epsilon \subset \mathbb{S}^{p-1}$ denote an ϵ-net over the unit sphere (ie each point on the sphere is no more than distance ϵ from some point in N_ϵ). Lemma 5.2 of [96] gives a constructive upper bound of ϵ-net of size $(1 + 2/\epsilon)^p$ over the unit sphere. For $u \in \mathbb{S}^{p-1}$, there exists $u' \in N_\epsilon$ such that $\|u - u'\|_{\ell_2} \leq \epsilon$. Then

$$|\Gamma(u, v, w, x) - \Gamma(u', v, w, x)| \leq \epsilon \|\Gamma\|_{op}.$$

This holds for all four arguments, so by the triangle inequality there exist $u', v', w', x' \in N_\epsilon$ satisfying

$$|\Gamma(u, v, w, x) - \Gamma(u', v', w', x')| \leq 4\epsilon \|\Gamma\|_{op}.$$

Then it follows that

$$\max_{y \in N_\epsilon} \Gamma(y, y, y, y) \leq \|\Gamma\|_{op} \leq \frac{1}{1 - 4\epsilon} \max_{y \in N_\epsilon} \Gamma(y, y, y, y).$$

Thus it suffices to bound the value of Γ on a ϵ-net where $\epsilon = 1/8$, which will give a bound on the entire sphere up to constant factor. For fixed $u, v, w, x \in N_\epsilon$, we can see that $\Gamma(u, v, w, x)$ is a Gaussian random variable with mean 0 and standard deviation at most σ by writing it as a sum of independent sub-Gaussian random variables:

$$\Gamma(u, v, w, x) = \sum_{i,j,k,l} \Gamma_{ijkl} u_i v_j w_k x_l.$$

Applying a standard Gaussian tail bound, for $t \geq 1$ we have

$$\Pr\left[|\sum_{i,j,k,l} \Gamma_{ijkl} u_i v_j w_k x_l| \geq t\right] \leq \frac{\sigma}{t\sqrt{2\pi}} \cdot \exp\left(-\frac{t^2}{2\sigma^2}\right).$$

Picking $t = c\sigma\sqrt{p + \log 1/\delta}$ for sufficiently large absolute constant $c > 0$ and union bounding over the 1/8-net in four coordinates we have

$$\Pr\left[\|\Gamma\|_{op} \geq c\sigma\sqrt{p + \log 1/\delta}\right]$$
$$\leq (1+16)^{4p} \cdot \frac{1}{c\sqrt{2\pi} \cdot \sqrt{p + \log 1/\delta}} \cdot \exp\left(-c^2/2 \cdot (\sqrt{p + \log 1/\delta})^2\right) \leq \delta. \ \square$$

Lemma 3.4.5 above shows that $M_X + \Gamma_4$ and M_X are close in spectral norm. Since M_X and M_v are close in spectral norm for a good reference model v, our remaining strategy is to transfer the approximate local optima conditions of Definiton 3.4.2 between tensors close in spectral norm with some loss of accuracy, which we do in Lemma 3.4.6 below. Our calculations are very similar to [94], though in the proof we fill in and clarify some omissions.

Lemma 3.4.6. *Let M, N be symmetric tensors in $\mathbb{R}^{p\times p\times p\times p}$ with $\|M - N\|_{op} \leq \beta$; let f and g be the associated quartic forms. If $u \in \mathbb{S}^{p-1}$ is an $(\epsilon, 0)$-approximate local optimum of M, then u is an $(\epsilon + 8\beta, (12 + 9/p)\beta)$-approximate local optimum of N.*

Proof. To establish the desired first order condition $\langle Dg_u, u\rangle \geq \|Dg_u\|_{\ell_2} - \epsilon - 8\beta$, we first use the spectral closeness of M and N to bound the difference in projections of u onto Df_u and Dg_u:

$$|\langle Df_u, u\rangle - \langle Dg_u, u\rangle| = 4|(M - N)(u, u, u, u)| \leq 4\|M - N\|_{op} \leq 4\beta.$$

Since by assumption we have $\langle Df_u, u\rangle \geq \|Df_u\|_{\ell_2} - \epsilon$, next consider $\|Df_u\|_{\ell_2}$ in terms

of $\|Dg_u\|_{\ell_2}$:

$$\begin{aligned}
|\|Df_u\|_{\ell_2} - \|Dg_u\|_{\ell_2}| &\leq \|Df_u - Dg_u\|_{\ell_2} \\
&\leq 4\|(M-N)(u,u,u,\cdot)\|_{\ell_2} \\
&\leq 4\langle (M-N)(u,u,u,\cdot), (M-N)(u,u,u,\cdot)/\|(M-N)(u,u,u,\cdot)\|_{\ell_2}\rangle \\
&\leq 4 \max_{v\in\mathbb{S}^{p-1}} |(M-N)(u,u,u,v)| \\
&\leq 4\beta.
\end{aligned}$$

The second to last inequality is clear when we observe that the second argument in the inner product is unit, and the last inequality follows from the hypothesis on the operator norm of $M - N$. Now we may combine these inequalities to get:

$$\langle Dg_u, u\rangle \geq \langle Df_u, u\rangle - 4\beta \geq \|Df_u\|_{\ell_2} - \epsilon - 4\beta \geq \|Dg_u\|_{\ell_2} - \epsilon - 8\beta.$$

To bound $\lambda_1(D^2 g_u)$, first note that λ_1 restricted orthogonal to u defines a spectral norm over matrices:

$$\begin{aligned}
\|D^2 f_u - D^2 g_u\|_{op} &= 12 \max_{v\in\mathbb{S}^{p-1}\cap u^\perp} |M(u,u,v,v) - N(u,u,v,v)| \\
&\leq 12\beta.
\end{aligned}$$

By the triangle inequality, $|\|D^2 f_u\|_{op} - \|D^2 g_u\|_{op}| = |\lambda_1(D^2 f_u) - \lambda_1(D^2 g_u)| \leq 12\beta$, and by assumption, $\lambda_1(D^2 f_u) \leq 9f(u)/p$ and $|f(u) - g(u)| \leq \beta$. Together we have:

$$\lambda_1(D^2 g_u) - 12\beta \leq \lambda_1(D^2 f_u) \leq \frac{9f(u)}{p} \leq \frac{9(g(u)+\beta)}{p}. \qquad \square$$

3.5 Utility for a Provable ICA Algorithm

Here we specialize our utility guarantees to an algorithm of [2], which constructs the isotropic fourth moment tensor and fixes two of its parameters as Gaussian random vectors to give a matrix. Observe that the matrix Quadruples(η, η') defined in Section 4.2 of [2] is the excess kurtosis tensor projected to a matrix $(M_v - \mathrm{E}\,[z^{\otimes 4}])(\eta, \eta', \cdot, \cdot)$

for standard p-dimensional Gaussian z. The spectrum of this matrix corresponds to the independent components. Given the following lemma, their (poly-time) ICA algorithm can be equivalently reproduced as Algorithm 2:

Lemma 3.5.1 (Lemma 4.2 of [2]). *Let $v \in \mathbb{R}^p$ be an isotropic random vector with $v = sA$ for $s \in \mathbb{R}^p$ fully independent with $\mathrm{E}[s_i^2] = 1$ and $A \in \mathbb{R}^{p \times p}$ unitary. Let $z \in \mathbb{R}^p$ be a standard Gaussian, and fix $\eta, \eta' \in \mathbb{R}^p$. Then,*

$$(M_v - \mathrm{E}[z^{\otimes 4}])(\eta, \eta', \cdot, \cdot) = A \cdot \mathrm{diag}(A^\top \eta) \cdot \mathrm{diag}(A^\top \eta') \cdot \mathrm{diag}(\mathrm{E}[s_1^4] - 3, \ldots, \mathrm{E}[s_2^4] - 3) \cdot A^\top.$$

Algorithm 2 Algorithm 2 from [2]

Input: Database $X \in \mathbb{R}^{n \times p}$

1: Compute the isotropic fourth moment tensor M_X.
2: Generate two random vectors θ, θ' uniformly at random from the unit sphere.
3: Compute the eigenvectors $\{v_1, \ldots, v_p\}$ of $(M_X - \mathrm{E}[z^{\otimes 4}])(\theta, \theta', \cdot, \cdot)$.
4: **return** $\{v_1, \ldots, v_p\}$.

Utility of our mechanism used in conjunction with the ICA algorithm of [2] is as follows:

Theorem 3.5.2 (ICA Utility for [2] with Private Tensors). *Fix $X \in \mathbb{R}^{n \times p}$ with an (α, β)-reference model for X. Let $\phi \geq 0$ be such that for each s_i of the reference model, we have $|\mathrm{E}[s_i^4] - 3| \geq \phi$. If we run Algorithm 2 using $M_X + \Gamma_4$ from Algorithm 1 in place of X in Step 1 and it outputs eigenvectors $\{v_1, \ldots, v_p\}$, then with probability greater than $3/4$ there exists a permutation $\tau : [p] \to [p]$ of eigenvectors such that for each column $a_i^\top$ of A^{-1}, $i \in [p]$:*

$$\|v_{\tau(i)} - a_i\|_{\ell_2} \leq c \frac{p^{1/2}}{\phi/p^5 - \|(M_X + \Gamma_4) - M_v\|_{op}} \|(M_X + \Gamma_4) - M_v\|_{op}.$$

Proof. Again let z denote the p-dimensional standard Gaussian, and fix two arbitrary unit vectors $\eta, \eta' \in \mathbb{S}^{p-1}$. Then we have:

$$\begin{aligned}\|(M_X + \Gamma_4 - \mathrm{E}\,[z^{\otimes 4}])(\eta, \eta', \cdot, \cdot) - (M_v - \mathrm{E}\,[z^{\otimes 4}])(\eta, \eta', \cdot, \cdot)\|_{op} \\ = \|(M_X + \Gamma_4 - M_v)(\eta, \eta', \cdot, \cdot)\|_{op} \leq \|M_X + \Gamma_4 - M_v\|_{op}\end{aligned}$$

We now consider the spectrum of $(M_v - \mathrm{E}\,[z^{\otimes 4}])(\theta, \theta', \cdot, \cdot)$ for $\theta, \theta' \in \mathbb{S}^{p-1}$. From Lemma 3.5.1, we have:

$$\begin{aligned}K &:= (M_v - \mathrm{E}\,[z^{\otimes 4}])(\theta, \theta', \cdot, \cdot) \\ &= A \cdot \mathrm{diag}(A^\top \theta) \cdot \mathrm{diag}(A^\top \theta') \cdot \mathrm{diag}(\mathrm{E}\,[s_1^4] - 3, \ldots, \mathrm{E}\,[s_2^4] - 3) \cdot A^\top\end{aligned}$$

Then by the rotational invariance of standard gaussian distributions, we see that the eigenvalues of K are given by $\lambda_i := \theta_i \theta_i' (\mathrm{E}\,[s_i^4] - 3)$ for $i = 1, \ldots, p$. We apply the following lemma from [2] to these λ_i:

Lemma 3.5.3 (Lemma C5 of [2]). *Fix $\{c_1, \ldots, c_n\} \subset \mathbb{R}^n$. Then for standard independent Gaussians $\{z_1, \ldots, z_n\}$ and $\delta \in (0, 1)$, we have $\min_i |z_i c_i| \geq \frac{\delta}{\sqrt{e} k^{2.5}} \min_i |c_i|$ with probability at least $1 - \delta$.*

We apply this lemma twice to the eigenvalues λ_i of K and get that with probability greater than $1 - 2\delta$:

$$|\lambda_i| \geq \frac{\delta^2 \phi}{4p^5} \qquad \text{and} \qquad |\lambda_i - \lambda_{i+1}| \geq \frac{\delta^2 \phi}{4p^5}$$

for each $i = 1, \ldots, p$. Then the theorem follows immediately from the following lemma:

Lemma 3.5.4 (Lemma C4 of [2]). *Let $A \in \mathbb{R}^{p \times p}$ be a Hermitian matrix with eigenvalues λ_i and eigenvectors v_i such that $|\lambda_i| \geq \Delta$ and $|\lambda_i - \lambda_{i+1}| \geq \Delta$ for $i = 1, \ldots, p$. Let $\hat{v}_i$ for $i = 1, \ldots, p$ denote the SVD of A recovered by the process used in Step 3 of Algorithm 2, and let $E \in \mathbb{R}^{p \times p}$ be a Hermitian matrix with $\|E\|_{op} \leq \Delta$. Then for $i = 1, \ldots, p$, we have:*

$$\|v_i - \hat{v}_i\|_{\ell_2} \leq \frac{2\sqrt{p}\|E\|_{op}}{\Delta - \|E\|_{op}}.$$

CHAPTER IV

ECONOMIC MARKETS FOR DIFFERENTIAL PRIVACY

4.1 Overview

Recall the original definition of differential privacy for a mechanism that analyzes n data records of type $\mathcal{D}$:

Definition (Exogenous privacy). *For $\epsilon \geq 0$, a mechanism $\mathcal{M} : \mathcal{D}^n \to Range(\mathcal{M})$ is ϵ-differentially private if for any neighboring $X, Y \in \mathcal{D}^n$ (i.e., databases differing on one record) and for any $S \subseteq Range(\mathcal{M})$,*

$$\Pr[\mathcal{M}(X) \in S] \leq \exp(\epsilon) \cdot \Pr[\mathcal{M}(Y) \in S].$$

Differentially private mechanisms protect individuals' data while also permitting meaningful statistical analysis. There is by now a rich body of research that establishes how to conduct a wide variety of statistical analysis goals while maintaining differential privacy. The privacy parameter ϵ implies a tradeoff between the strength of the privacy guarantee and the utility (accuracy) of the statistical analysis allowed by the ϵ privacy guarantee. However, the literature is largely agnostic to the choice of ϵ.

This chapter presents the work of [59], which continues a new line of differential privacy research seeking to develop *endogenous* private mechanisms that internalize this privacy/utility tradeoff. Of particular focus are mechanisms that set ϵ according to the preferences of the *data contributors* and *analyst*. The data contributors are the individuals who each contribute one record to the database, and these parties are concerned with the privacy of their data. In contrast, the analyst wants access to statistically useful data. The mechanism itself is responsible for collecting the data and ensuring privacy. In the real world, the analyst may be the party responsible

for collecting the data, and a promise to access data through the privacy mechanism can be seen as part of the analyst's privacy contract with the data contributors. Alternatively, a third-party database curator may collect data and run the mechanism. The models and mechanisms presented in this chapter consider only the interests of the data contributors and analyst, and not who operates the mechanism.

This overview first outlines new considerations in modeling and analyzing mechanisms that set their own level of privacy. It ends with a description of a specific realization of a general-purpose mechanism that simulates a market to endogenously find and privately implement an optimal privacy/utility tradeoff based on data contributor and analyst preferences.

Modeling Endogenous Privacy Mechanisms

The closely related mechanisms of Ghosh and Roth [34] do four things: 1) solicit privacy/accuracy preferences from the data contributors and analyst, 2) determine an appropriate value of ϵ, 3) charge the analyst some payment in exchange for a noisy statistic on the data, and 4) distribute this payment among data contributors to compensate for the ϵ privacy loss. Mechanisms in this chapter fit a straightforward generalization of this framework. As noted in [34] and others, if privacy preferences and private data are correlated, then privacy must be guaranteed for both. Unfortunately, Ghosh and Roth show that if there is no a priori bound on the costs data contributors experience for privacy loss, then any mechanism that protects the privacy of data contributors' privacy preferences cannot adequately compensate data contributors for their privacy loss while satisfying other natural market properties.

In Section 4.2, we strengthen this negative result. Even if data contributors have positive value for a privacy guarantee and must pay the analyst for such a guarantee, and even with further relaxations of the desired market properties, mechanisms are unable to effectively simulate a market for privacy (Theorem 4.2.3). We find that

the core of this negative result is not the requirement that privacy be guaranteed with respect to privacy preferences as implied in [34], but rather that the differential privacy guarantee itself is inconsistent with this setting in which a market chooses the level of privacy endogenously.

This work's first main contribution is a new *definition* of privacy for this setting. The usual definition treats ϵ as a data-independent parameter, but our mechanisms choose ϵ as a function of data. In [34], a mechanism must ensure that its outputs on *any* two neighboring databases must be ϵ close for the *smallest* ϵ the mechanism may output on *any* input database, even if the databases in question only yield much higher values of ϵ. The standard differential privacy definition is parametrized by a single fixed ϵ, and this precludes any meaningful privacy implication of an ϵ that the mechanism chooses as a function of its input data. The new notion of *endogenous differential privacy* rectifies this:

Definition (Endogenous privacy, informal). *A mechanism* $\mathcal{M} : \mathcal{D}^n \to \mathit{Range}(\mathcal{M})$ *is* endogenously differentially private *if for any neighboring* $X, Y \in \mathcal{D}^n$*, for any* ϵ *in the privacy support of* $\mathcal{M}(X)$*, and for any* $S \subseteq \mathit{Range}(\mathcal{M})$*,*

$$\Pr[\mathcal{M}(X) \in S] \leq \exp(\epsilon) \cdot \Pr[\mathcal{M}(Y) \in S].$$

When we relax the worst-case privacy requirement of [34] to this endogenous privacy requirement, the proofs of the negative results do not survive, and so we turn to realizing positive results that apply the new privacy definition.[1]

Markets for Privacy

The second main contribution of this work is a positive result in our framework in the form of a class of mechanisms whose privacy guarantees provide positive value to the data contributors. These mechanisms follow four steps: 1) receive the private

[1] Definitions 4.2.1 and 4.3.1 respectively formalize the privacy requirement of [34] and the new endogenous privacy requirement for the privacy markets formalized by Model 3.

data and solicit privacy/accuracy preferences, 2) choose a market-appropriate privacy level, 3) transfer payments, and 4) run a standard differentially private mechanism at the previously determined level of privacy. These mechanisms provide privacy for the private data as well as data contributors' preferences for privacy, circumventing the negative results of [34] using the new endogenous privacy definition.

When data contributors value a mechanism's privacy guarantee, privacy should be seen as a *public good* because the same guarantee is enjoyed by all data contributors although they may be charged different amounts for it. The new challenge is to discourage data contributors from understating their individual desires for privacy, letting others pay for the privacy enjoyed by all. The cumulative works of [97, 22, 37, 38] provide an elegant solution to this "free-rider problem" as it exists more generally in neoclassical economies, achieving the market goal of a *Pareto efficient* (see Definition 4.4.4) level of production while incentivizing consumers to report their true preferences.

The class of mechanisms presented in Section 4.4 adapts this solution to the free-rider problem to our privacy market framework. A market-simulating mechanism $\mathcal{M}$ in our class computes an efficient level of privacy, and it charges each data contributor a payment that aligns individual utility with social utility. The market mechanism then runs some standard differentially private mechanism $\mathcal{M}_q$ to approximate the desired query (or class of queries) on the database using the market-determined ϵ. These mechanisms are endogenously differentially private, and they satisfy several desired market properties that are simultaneously impossible under the old privacy requirement.

4.2 Negative Results for Privacy Markets

We first specify our model for a mechanism that chooses a privacy guarantee as a function of its inputs, which include privacy and accuracy preferences for the data

contributors and analyst. Such a mechanism computes a statistic at the prescribed level of privacy, and it enforces monetary transfers between data contributors and analyst.

We refer to the set of (ϵ, δ) output by mechanism $\mathcal{M}$ in Model 3 as its privacy support. We denote the joint random variables representing the distribution of (public and private) outputs by $\mathcal{M}$ running on inputs $\mathbf{d} \in \mathcal{D}^n, \mathbf{v} \in \mathcal{V}^n, c \in \mathcal{C}$ as $(y, \epsilon, \delta, \mathbf{p}, \hat{P}, \hat{R}) \leftarrow \mathcal{M}(\mathbf{d}, \mathbf{v}, c)$, and similarly for the marginal distributions. When the outputs are not specified, $\mathcal{M}(\mathbf{d}, \mathbf{v}, c)$ denotes the distribution of public outputs $(\hat{R}, \hat{P})$.

Model 3 Privacy Market

1: Upon initialization, there exist general types $\mathcal{D}, \mathcal{R}, \mathcal{Y}$, preference function families $\mathcal{V}, \mathcal{C} \subseteq \{\mathcal{Y} \to \mathbb{R}\}$, and a mechanism $\mathcal{M} : \mathcal{D}^n \times \mathcal{V}^n \times \mathcal{C} \to \mathcal{Y} \times \mathbb{R}^+ \times \mathbb{R}^+ \times \mathbb{R}^n \times \mathbb{R} \times \mathcal{R}$. Each data contributor $i \in [n]$ has some data $d_i \in \mathcal{D}$ and true preference $v_i^* \in \mathcal{V}$ for the privacy properties of $\mathcal{M}$, and an analyst has a preference $c \in \mathcal{C}$.
2: $\mathcal{M}$ receives the verifiable data, and data contributors and the analyst report their preferences to $\mathcal{M}$.
3: $\mathcal{M}$ privately outputs endogenous privacy parameter $y \in \mathcal{Y}$ and corresponding privacy guarantees $\epsilon, \delta \in \mathbb{R}^+$.
4: $\mathcal{M}$ makes monetary transfers $\mathbf{p} \in \mathbb{R}^n$ and $\hat{P} \in \mathbb{R}$ from and to the data contributors and analyst, respectively.
5: $\mathcal{M}$ publishes statistic $\hat{R} \in \mathcal{R}$.
6: Each data contributor $i \in [n]$ realizes utility $v_i^*(y) - p_i$.

Model 3 represents a straightforward syntactic generalization of the models of [34], in which each data record is a bit $d_i \in \mathcal{D} = \{0, 1\}$, the published statistic is a noisy average of these bits $\hat{R} \in \mathcal{R} = \mathbb{R}$, and $\mathcal{Y} = \mathbb{R}^+$ with y directly specifying a differential privacy requirement $\epsilon = y, \delta = 0$ for the mechanism. The generalized model is compatible with their positive results in the *insensitive value model*, which requires mechanisms to be private only with respect to private data. We restrict our focus to mechanisms that guarantee privacy also with respect to privacy preferences, noting the settings described in other works [65, 30, 80], in which privacy preferences are likely to reveal information about private data. To address this concern, [34] also proposes a *sensitive value model*, which requires mechanisms to be private with respect

to privacy preferences as well as private data:

Definition 4.2.1 (Privacy, [34]). *A mechanism $\mathcal{M}$ in Model 3 is differentially private if for any ϵ in the privacy support of $\mathcal{M}$, for any $(\mathbf{d}, \mathbf{v}) \sim (\mathbf{d}', \mathbf{v}')$ differing on one row and $c \in \mathcal{C}$, and for any $E \subseteq \mathbb{R} \times \mathcal{R}$,*

$$\Pr[\mathcal{M}(\mathbf{d}, \mathbf{v}, c) \in E] \leq \exp(\epsilon) \cdot \Pr[\mathcal{M}(\mathbf{d}', \mathbf{v}', c) \in E].$$

However, they give a strong negative result about the existence of meaningful mechanisms in this model. Stronger negative results generalizing their model as in Model 3 are given in Theorems 4.2.2 and 4.2.3. The proofs of these results use essentially the same techniques as [34] and are given at the end of the chapter in Section 4.5.3, where we also reproduce the original statement and proof of the negative result of [34].

Theorem 4.2.2 (Negative results for costly privacy loss). *Consider any mechanism $\mathcal{M}$ in Model 3 with imperfect privacy, and assume data contributors' preferences are arbitrarily negative.*

1. [34] *If $\mathcal{M}$ adequately compensates data contributors and is budget-balanced, then for any finite $B > 0$, we have that $\mathcal{M}$ running on any fixed inputs always charges the analyst $-\hat{P} > B$.*

2. *If $\mathcal{M}$ adequately compensates data contributors and is likely budget-balanced, then for any finite $B > 0$, we have that $\mathcal{M}$ running on any fixed inputs charges the analyst $-\hat{P} \leq B$ with probability at most $\exp(n \cdot \epsilon_{\text{inf}})/2$.*

If we assume that data contributors have *positive* value for a mechanism's privacy guarantee, departing from [34] and others, we can circumvent the [34] result easily: simply fix some $\epsilon > 0$ exogenously, make no monetary transfers, and compute $\hat{R}$ using any standard ϵ-differentially private mechanism. Such a mechanism provides data contributors with nonnegative utility and loses no money. However, the following

theorem shows that these trivial moneyless mechanisms are essentially the only mechanisms that satisfy the desired properties:

Theorem 4.2.3 (Negative results for valuable privacy). *Consider any private mechanism $\mathcal{M}$ in Model 3, and assume data contributors' preferences are positive and arbitrarily small.*

1. *If $\mathcal{M}$ fairly charges data contributors and is budget-balanced, then for any fixed $R > 0$, we have that $\mathcal{M}$ running on any fixed inputs always pays the analyst $\hat{P} < R$.*

2. *If $\mathcal{M}$ fairly charges data contributors and is likely budget-balanced, then for any fixed $R > 0$, we have that $\mathcal{M}$ running on any fixed inputs pays the analyst $\hat{P} \geq R$ with probability at most $\exp(n \cdot \epsilon_{\text{inf}})/2$.*

To motivate the following section, we note that the working privacy definition states that the output of mechanisms on neighboring databases must be ϵ-close for any ϵ output by the mechanism on any inputs, and in particular, for the smallest such $\epsilon_{\text{inf}} := \inf\{\epsilon \leftarrow \mathcal{M}(\mathbf{d}, \mathbf{v}, c) \ : \ \mathbf{d} \in \mathcal{D}^n, \mathbf{v} \in \mathcal{V}^n, c \in \mathcal{C}\}$. As privacy preferences become arbitrarily negative or arbitrarily small, reasonable mechanisms should output very small ϵ. In particular, if $\epsilon_{\text{inf}} = o(1/n)$, then the theorems roughly imply that if an analyst has any fixed (arbitrarily large) budget or (arbitrarily small) target revenue, then on any inputs, $\mathcal{M}$ will exceed the budget or fail to supply the target revenue, respectively, with probability almost 1/2. The proofs of Section 4.5.3 reveal that the worst-case privacy guarantee of Definition 4.2.1 is responsible for these unsatisfying results. Noticing the disconnect between the input-dependent ϵ chosen by a privacy market and the global, input-independent ϵ_{inf} enforcing a privacy guarantee, we modify the standard exogenous differential privacy definition.

4.3 Endogenous Privacy

Although data contributors realize utility as a function of the privacy level determined endogenously by the market, Definition 4.2.1 guarantees ϵ_{inf} privacy for any pair of neighboring databases. Because we want the privacy level selected by a mechanism to be a function of its inputs, we must endogenize the privacy guarantee itself in order to justify the utility realized by this guarantee.

The usual definition of differential privacy captures the idea that an individual cares about the difference between two output distributions: that of the mechanism run on the true database, and that of the mechanism run on the same database with his row changed. Requiring privacy for all ϵ in the support of the mechanism on *any* inputs goes far beyond the true concerns of a data contributor, whose reality is associated with some particular database. We propose a new definition that endogenizes the privacy guarantee by requiring that the output distribution of the mechanism on any set of reference inputs is close to the output distribution of the mechanism on any neighboring set of inputs, and *only the ϵ supported by mechanism running on the reference inputs* stipulate the closeness of these distributions.

Definition 4.3.1 (Endogenous differential privacy). *A mechanism $\mathcal{M}$ in Model 3 is endogenously differentially private if for any $(\mathbf{d}, \mathbf{v}) \sim (\mathbf{d}', \mathbf{v}')$ differing on one row and $c \in \mathcal{C}$, for any (ϵ, δ) in the privacy support of $\mathcal{M}(\mathbf{d}, \mathbf{v}, c)$, and for any $E \subseteq \mathbb{R} \times \mathcal{R}$,*

$$\Pr[\mathcal{M}(\mathbf{d}, \mathbf{v}, c) \in E] \leq \exp(\epsilon) \cdot \Pr[\mathcal{M}(\mathbf{d}', \mathbf{v}', c) \in E] + \delta, \text{ and}$$
$$\Pr[\mathcal{M}(\mathbf{d}', \mathbf{v}', c) \in E] \leq \exp(\epsilon) \cdot \Pr[\mathcal{M}(\mathbf{d}, \mathbf{v}, c) \in E] + \delta.$$

Utility for endogenous privacy. We privately output an endogenous privacy parameter $y \in \mathcal{Y}$ in addition to the explicit guarantees $\epsilon, \delta \in \mathbb{R}^+$ to separate the tasks of proving privacy and of modeling data contributors' utility for (the privacy properties of) a mechanism. Data contributors' utility may not be characterized solely by a

mechanism's provable (ϵ, δ) guarantee. For simple mechanisms that establish privacy with optimally-calibrated Laplace noise, the ϵ-privacy guarantee is tight. However, [78, 20, 80] argue that in general ϵ alone can only provide an *upper bound* on true privacy cost of a mechanism. Such a guarantee may not provide a tight bound on the information leaked by the mechanism, for example, when output distributions of a mechanism on neighboring databases are only ϵ apart for an extremely unlikely set of events and closer otherwise, or when the analyzed upper bound on ϵ may itself be loose. By allowing y to be of a general form, our mechanisms can potentially release more specific information about their privacy policies that may allow tighter analysis of privacy loss. Furthermore, data contributors' utilities for a mechanism outputting a general privacy parameter need not be limited to the mechanism's privacy properties. For example, y may include the public outputs of the mechanism, allowing the new framework to model outcome-dependent utility [78, 79, 20, 80], which depends on public outputs as well as differential privacy guarantees.

4.4 A Class of Endogenous Privacy Markets

Theorem 4.2.3 shows that a mechanism that fairly charges data contributors for privacy and is likely budget balanced cannot guarantee worst-case exogenous differential privacy if the analyst hopes to extract revenue from the data contributors with reasonably high probability on some inputs. In this section, we present a class of mechanisms that are endogenously private and satisfy all of these properties. We begin by formalizing these desired economic properties. We then discuss a particular challenge in the positive value setting and a market-based solution approach, we provide a simple warm-up mechanism in Section 4.4.3, and we present our class of general-purpose mechanisms in Section 4.4.4.

4.4.1 Market Properties

The principal mechanism design goal of both [34] and the current work is to elicit data contributors' true privacy preferences v_i^*. In order to reasonably assume that data contributors report these preferences truthfully, we would like our mechanisms to be *incentive compatible*, meaning that each individual maximizes his expected utility by reporting his true privacy type:

Definition 4.4.1 (Incentive compatibility). *A mechanism $\mathcal{M}$ in Model 3 is incentive compatible if for any $i \in [n]$, privacy preference $v_i^* \in \mathcal{V}$, and any inputs $\mathbf{d} \in \mathcal{D}^n, \mathbf{v}_{-i} \in \mathcal{V}^{n-1}, c \in \mathcal{C}$,*

$$v_i^* \in \arg\max_{v_i} \mathrm{E}[v_i^*(y) - p_i],$$

where $(y, p_i) \leftarrow \mathcal{M}(\mathbf{d}, \mathbf{v}_{-i} \| v_i, c)$ and the expectation is over the randomness of $\mathcal{M}$.

As in [34], our mechanisms should incentivize truthful reporting of privacy preferences, but we assume that true private data is already held somewhere and is verifiable, e.g., by a trusted database curator. In these settings, even if an individual expects a privacy market not to benefit him, it may not be possible for him to retract his data. Nonetheless, if a mechanism provides non-negative utility to every data contributor, it is *individually rational* for everyone to opt in:

Definition 4.4.2 (Individual rationality). *A mechanism $\mathcal{M}$ in Model 3 is individually rational if for any $i \in [n]$ and inputs $\mathbf{d} \in \mathcal{D}^n, \mathbf{v} \in \mathcal{V}^n, c \in \mathcal{C}$,*

$$\mathrm{E}[v_i(y) - p_i] \geq 0,$$

where $(y, p_i) \leftarrow \mathcal{M}(\mathbf{d}, \mathbf{v}, c)$ and the expectation is over the randomness of $\mathcal{M}$.

Since the verifiable data setting means that opting out may be impossible, we permit our mechanisms to violate individual rationality in the case of artificially extreme inputs.

We say a mechanism has a *balanced budget* when $\hat{P} = \sum p_i$. For our negative results, we relax this and require only that a mechanism takes in at least as much money as it pays out with probability at least 1/2:

Definition 4.4.3 (Likely balanced budget). *A mechanism $\mathcal{M}$ in Model 3 is likely budget-balanced if for any inputs $\mathbf{d} \in \mathcal{D}^n, \mathbf{v} \in \mathcal{V}^n, c \in \mathcal{C}$,*

$$\Pr[\textstyle\sum p_i \geq \hat{P}] \geq 1/2,$$

where $(\mathbf{p}, \hat{P}) \leftarrow \mathcal{M}(\mathbf{d}, \mathbf{v}, c)$ and the expectation is over the randomness of $\mathcal{M}$.

Our positive results are likely budget-balanced and satisfy the additional requirement that they avoid a deficit in expectation for any fixed inputs. An *expected balanced budget*, $\mathrm{E}[\sum p_i - \hat{P}] \geq 0$, can be thought of as a cyclically balanced budget, in that the surpluses will offset the deficits over time. The left tail of $\sum p_i - \hat{P}$ should also be tightly bounded so the mechanism is unlikely to ever run a large deficit.

The mechanisms discussed in [34] that protect the privacy of data contributors' private bits but not of their privacy valuations seek to either minimize analyst payment subject to some minimum accuracy requirement or maximize accuracy subject to some maximum budget. Such mechanisms do not take into account an analyst's desired tradeoff between money and accuracy. Mechanisms in the generalized framework solicit some $c \in \mathcal{C} \subseteq \{\mathcal{Y} \to \mathbb{R}\}$ from the analyst that describes this tradeoff. A privacy level $y \in \mathcal{Y}$ typically corresponds to the noisiness of the statistic $\hat{R}$, so $c(y)$ represents the opportunity cost to the analyst of $\hat{R}$ generated by the mechanism running on y compared to the noiseless statistic. Given the preferences of data contributors and the analyst, our mechanisms seek to find a *Pareto efficient* (or *Pareto optimal*) level of privacy, meaning some $y \in \mathcal{Y}$ for which no data contributor can be made strictly better off without making another strictly worse off, subject to collecting enough total funds for the analyst. Preference families $\mathcal{V}$ and $\mathcal{C}$ should be chosen so that for any $\mathbf{v} \in \mathcal{V}^n, c \in \mathcal{C}$, there exists some Pareto efficient $y \in \mathcal{Y}$.

Definition 4.4.4 (Pareto efficiency). *Privacy level $y \in \mathcal{Y}$ is Pareto efficient for $\mathbf{v} \in \mathcal{V}^n, c \in \mathcal{C}$ if there exist payments $\mathbf{p} \in \mathbb{R}^n$ such that $\sum p_i \geq c(y)$, and for any $y', \mathbf{p}'$ such that $\sum p'_i \geq c(y')$ and $v_i(y') - p'_i > v_i(y) - p_i$ for some $i \in [n]$, there exists some $j \in [n]$ with $v_j(y') - p'_j < v_j(y) - p_j$.*

In [34], a mechanism releasing statistic $\hat{s} \approx \sum b_i$ is called α-*accurate* if $\Pr[|\hat{s} - \sum b_i| > \alpha n] \leq 1/3$. When we generalize the negative result of [34], we replace their accuracy requirement with a requirement of nontrivial privacy support (see Section 4.5.3), so this work does not require a formal definition of accuracy. The statistic $\hat{R} \in \mathcal{R}$ output by a mechanism in our framework is typically an approximation of some generally-typed query $q : \mathcal{D}^n \to \mathcal{R}$, so the accuracy of $\hat{R}$ should be query-specific. Utility guarantees for standard differentially privacy mechanisms may inform the accuracy goals of mechanisms in our framework.

4.4.2 Public Goods and Free-Riders

In [34], data contributors with high privacy valuations v_i^* must be paid more to compensate them for costlier privacy losses. To discourage data contributors from overstating their costs to extract greater payments, the [34] mechanisms simulate a second price auction for a given privacy guarantee, and this approach guarantees incentive compatibility. The mechanism determines some $v_{\max}$ and ϵ, and all data contributors with reported $v_i < v_{\max}$ receive a payment $-v_{\max}(\epsilon)$ that exceeds their cost from ϵ privacy loss. Others get nothing and their data is not used in the summary statistic. However, since the privacy preferences are used to determine this cutoff, this solution approach does not protect privacy of these preferences, and we consider protecting privacy of preference inputs to be a primary concern.

We model a setting in which data contributors receive positive utility for the privacy of a mechanism (compared to a baseline of no privacy guarantee at all), so we can *charge* privacy-sensitive data contributors more to offset the cost associated

with strengthening the privacy guarantee to accommodate them. Interestingly, this significantly changes the incentive structure. Our preference functions when privacy has positive value model privacy as a *public good*, because the same privacy guarantee is enjoyed by all data contributors, even though they may be charged different amounts for it. The new challenge is that data contributors may try to avoid higher charges by understating their sensitivity, letting others pay for the privacy enjoyed by all. This "free-rider problem" is solved in a much more general public goods setting in the cumulative works of [97, 22, 37, 38].

In the setting studied by [22, 37, 38], *consumers* communicate to some central body, called the *government*, their valuation $v_i(\cdot)$ of a certain public good. The government chooses the level of public good that optimizes social utility, and it levies taxes designed to align individual consumers' utilities with social utility in order to avoid free-riding. Specifically, the government pays a *producer* $c(y)$ to produce $y \geq 0$ units of the good for the level y maximizing consumer surplus, $\sum v_i(y) - c(y)$. Each consumer i receives utility $v_i(y)$ for the public good and is charged the amount he diminishes others' surplus: $p_i = c(y) - \sum_{j \neq i} v_j(y) + \max_{y_{-i}}(\sum_{j \neq i} v_j(y_{-i}) - \frac{n-1}{n} c(y_{-i}))$. With these allocation and tax rules, consumers are incentivized to communicate their true preferences, and sufficient funds are raised to produce a Pareto efficient level of the public good.[2]

[2]To verify incentive compatibility, note that $\max_{y_{-i}}(\sum_{j \neq i} v_j(y) - \frac{n-1}{n} c(y_{-i}))$ is independent of v_i, so to maximize his utility, i should report $\arg\max_{v_i} v_i^*(y) - (c(y(\mathbf{v}_{-i} \| v_i, c)) - \sum_{j \neq i} v_j(y(\mathbf{v}_{-i} \| v_i))$. Because $y(\mathbf{v}_{-i} \| v_i^*) = \arg\max_y v_i^*(y) + \sum_{j \neq i} v_j(y) - c(y)$, this quantity is indeed maximized when $v_i = v_i^*$.

A sufficient condition for Pareto efficiency is the Samuelson condition [88], that the sum of the marginal benefit of a public good over all consumers equals its marginal cost. With this allocation rule, the level of y maximizing consumer surplus is y such that $\sum \frac{d}{dy} v_i(y) = \frac{d}{dy} c(y)$. Assuming incentive compatibility, this is equivalent to the condition that the sum of the marginal benefit of y is equal to the marginal cost, so the allocation rule is Pareto efficient.

It can be easily verified that the sum of payments is at least $c(y)$. Since it may be strictly greater, the payments collected are not guaranteed to be Pareto efficient (even though the privacy level y is). This is a problem addressed in [38] through different tax and allocation rules, but these modifications complicate the privacy utility model in our setting so we opt for these simpler rules.

4.4.3 Warm-Up Privacy Market

In our setting, privacy is a public good. The mechanism, data contributors, and analyst respectively play the roles of government, consumers, and producer in the public goods market due to [22, 37, 38]. Our approach is to solicit preferences for privacy from the data contributors, determine an efficient level of privacy and appropriate payments using this public goods market, and then compute a statistic of the input data with noise calibrated for the target privacy level.

We fix the sets of privacy parameters, preference functions, and cost functions as:

$$\mathcal{Y} := \mathbb{R}^+, \qquad \mathcal{V} := \{y \mapsto v_i \ln(y+1) : v_i \in \mathbb{R}^+\}, \qquad \mathcal{C} := \{y \mapsto cy : c \in \mathbb{R}^+\}.$$

This choice of privacy parameters and cost functions follows [22, 38]. Logarithmic privacy preferences $\mathcal{V}$ help us establish privacy, which we discuss in Section 4.5.1 when proving privacy for the more general class of markets in Section 4.4.4. We identify $\mathcal{V}$ and $\mathcal{C}$ with $\mathbb{R}^+$ in the natural way. For the warm-up, we fix $\mathcal{D} := \{0,1\}$ and $\mathcal{R} := \mathbb{R}$, and then Mechanism 4 publishes a noisy sum $\hat{R} \approx \sum_{i\in[n]} d_i$, following [34].

Mechanism 4 Warm-Up Privacy Market

Inputs: Database $\mathbf{d} \in \{0,1\}^n$, preferences $\mathbf{v} \in (\mathbb{R}^+)^n$, cost $c \in \mathbb{R}^+$.

1: $\bar{v}_i \leftarrow \min(v_i, c \ln n)$ for all $i \in [n]$.
2: $y \leftarrow (\frac{\sum_i \bar{v}_i}{c} - 1)^+$.
3: $\epsilon \leftarrow 3 \ln n / \sqrt{y}$.
4: $\delta \leftarrow \exp(-2\sqrt{y - \ln n})$.
5: $p_i \leftarrow cy - \sum_{j\neq i} \bar{v}_j \ln(y+1) + \max_{y_{-i}\geq 0} \left(\sum_{j\neq i} \bar{v}_j \ln(y_{-i}+1) - \frac{n-1}{n} c y_{-i} \right)$ for each $i \in [n]$.
6: $\hat{P} \leftarrow c(y+\gamma)$ with γ drawn from $\mathrm{Lap}(\sqrt{y + \ln n})$.
7: $\hat{R} \leftarrow \sum d_i + \gamma'$ with γ' drawn from $\mathrm{Lap}(\sqrt{y}/\ln n)$.
8: Privately output y, ϵ, δ, collect p_i from each $i \in [n]$, pay analyst $\hat{P}$, and publish $\hat{R}$.

Endogenous privacy follows from Lemma 4.5.2 for the general class of privacy markets. Accuracy of $\hat{R}$ depends on the noise $\mathrm{Lap}(\sqrt{y}/\ln n) = \mathrm{Lap}(3/\epsilon)$ in the usual way. Assuming that v_i, c are constant with respect to n, then $\epsilon = \Theta(1/\sqrt{n})$ and δ is negligible, and no v_i will be truncated in Step 1, so the y computed in Step 2 is

Pareto optimal. Incentive compatibility also follows from [97, 22, 37, 38], and the conditions of Lemma 4.5.4 for individual rationality hold as well. The mechanism is budget-balanced in expectation, with the probability of a deficit $\geq t$ decreasing exponentially with $t/\Theta(\sqrt{n})$.

4.4.4 General Class of Privacy Markets

The general class of markets uses an exogenous differentially private mechanism as a black-box subroutine for the analysis task, so we leave the database $\mathbf{d} \in \mathcal{D}^n$ and published statistic $\hat{R} \in \mathcal{R}$ of general types. We assume that the analysis goal is characterized by some query (or class of queries) $q : \mathcal{D}^n \to \mathcal{R}$. We assume the existence of some mechanism $\mathcal{M}_q : \mathcal{D}^n \to \mathcal{R}$ that is parametrized by ϵ_q, δ_q such that it is (ϵ_q, δ_q)-differentially private (in the standard sense) when instantiated on any ϵ_q, δ_q in some legal set of privacy parameters that includes arbitrarily small $\epsilon_q \geq 0$. For accuracy, we should have $\mathcal{M}_q(\mathbf{d}) \approx q(\mathbf{d})$, where the approximation is a statistically meaningful metric on $\mathcal{R}$, with accuracy depending reasonably on the privacy parameters ϵ_q, δ_q.

We again fix the sets of privacy parameters, preference functions, and cost functions:

$$\mathcal{Y} := \mathbb{R}^+, \qquad \mathcal{V} := \{y \mapsto v_i \ln(y+1) : v_i \in \mathbb{R}^+\}, \qquad \mathcal{C} := \{y \mapsto cy : c \in \mathbb{R}^+\}.$$

At the end of Section 4.5.1, we show how to extend our results to other preference families $\mathcal{V}$.

Mechanism 5 first truncates the privacy valuations (according to parameter Δ). It then computes the Pareto efficient privacy level y and incentive-compatible charges to the data contributors, using the allocation and tax rules from [22, 37, 38]. The mechanism adds noise to the analyst's payment (according to a function f of the privacy level y), and the statistic $\hat{R}$ is computed by $\mathcal{M}_q$ with privacy parameters $\epsilon_q = \Delta / f(y - \Delta)$ and fixed δ_q. We will see that truncation and noise suffice for proving endogenous privacy. Lemma 4.5.5 provides guidance for setting parameters; $\Delta = \ln n$ and $f(y) = \sqrt{y + \Delta}$ as in the warm-up mechanism are reasonable choices.

Mechanism 5 Privacy Market

Parameters: Truncation rule $\Delta \in \mathbb{R}^+$ and differentiable noise function $f : \mathbb{R} \to \mathbb{R}^+$; mechanism $\mathcal{M}_q : \mathcal{D}^n \to \mathcal{R}$ and fixed δ_q.

Inputs: Database $\mathbf{d} \in \mathcal{D}^n$, preferences $\mathbf{v} \in (\mathbb{R}^+)^n$, cost $c \in \mathbb{R}^+$.

1: $\bar{v}_i \leftarrow \min(v_i, c \cdot \Delta)$ for all $i \in [n]$.
2: $y \leftarrow (\frac{\sum_i \bar{v}_i}{c} - 1)^+$.
3: $\epsilon \leftarrow 3\Delta / f(y - \Delta)$.
4: $\delta \leftarrow \delta_q + 1/\exp(1/f'(y - \Delta))$.
5: $p_i \leftarrow cy - \sum_{j \neq i} \bar{v}_j \ln(y+1) + \max_{y_{-i} \geq 0} \left(\sum_{j \neq i} \bar{v}_j \ln(y_{-i} + 1) - \frac{n-1}{n} c y_{-i} \right)$ for each $i \in [n]$.
6: $\hat{P} \leftarrow c(y + \gamma)$ with γ drawn from $\text{Lap}(f(y))$.
7: $\hat{R} \leftarrow \mathcal{M}_q(\mathbf{d})$ with privacy parameters $\epsilon_q = \frac{\Delta}{f(y-\Delta)}$ and δ_q.
8: Privately output y, ϵ, δ, collect p_i from each $i \in [n]$, pay analyst $\hat{P}$, and publish $\hat{R}$.

The main theorem below is an immediate corollary of the lemmas and discussion in Sections 4.5.1 and 4.5.2:

Theorem 4.4.5. *Fix any $\Delta \in \mathbb{R}^+$, any function $f : \mathbb{R} \to \mathbb{R}^+$ that is increasing, concave, and differentiable with $f'(0) \leq 1$, and any mechanism $\mathcal{M}_q$ that is differentially private in the standard sense for arbitrarily small ϵ_q and fixed δ_q. Then Mechanism 5 instantiated with $\Delta, f, \mathcal{M}_q$ is endogenously differentially private, it is incentive compatible with respect to privacy valuations, it is individually rational when $c \leq \sum \bar{v}_i / e$, it is budget-balanced in expectation, and it selects a Pareto efficient level of privacy when $\max v_i \leq c \cdot \Delta$.*

4.5 Proofs of Positive and Negative Results

4.5.1 Endogenous Privacy of Mechanism 5

After fixing $c(y) = cy$ for $c, y \in \mathbb{R}^+$ following [22, 38], the main goal is to release $\hat{P} \approx cy$ in an endogenously differentially private manner for some appropriate choices of ϵ and δ. Classical differential privacy techniques would suggest first bounding sensitivity Δ of y, and then $\hat{P} = c(y + \text{Lap}(\Delta/\epsilon))$ is ϵ-differentially private for some fixed (exogenous) ϵ. We begin the privacy proof by bounding the sensitivity of y, and then we derive functions ϵ, δ that are positive and decreasing in y and for which we

can prove endogenous privacy.

For consistency with the perspective of y as a public good, willingness-to-pay functions $v_i(y)$ should be nonnegative, increasing, and concave. With $c(y) = cy$, the unique consumer surplus-maximizing level of privacy will have $\sum \frac{d}{dy} v_i(y) = c$ unless $y = 0$. When $v_i(y) = v_i \ln(y+1)$, the uniquely optimal privacy level is given by $\sum v_i/c - 1$. We control the sensitivity by first truncating the preferences to $\bar{v}_i := \min(v_i, c \cdot \Delta)$, denoting the non-negative surplus-maximizing level of privacy as:

$$y(\mathbf{v}, c) := \left(\frac{\sum \bar{v}_i}{c} - 1 \right)^+. \qquad \text{(Step 2)}$$

Then sensitivity of y is immediate from $|(\sum \bar{v}_i/c - 1)^+ - (\sum \bar{v}'_i/c - 1)^+| \leq |(v_1 - v'_1)/c|$, assuming without loss of generality that neighboring $\mathbf{v}, \mathbf{v}'$ differ on the first row:

Fact 4.5.1 (Sensitivity of y). *For any $c \in \mathbb{R}^+$ and neighboring $\mathbf{v} \sim \mathbf{v}'$, we have $|y(\mathbf{v}, c) - y(\mathbf{v}', c)| \leq \Delta$.*

We may worry that truncating v_i will generate sample bias, as argued in [34]. Indeed, if data contributors have negative utilities for privacy loss, a mechanism operating on truncated costs will not be able to adequately compensate data contributors, and if these privacy-sensitive data contributors are able to opt out of the mechanism, this may bias the data. Alternatively, we may worry that truncation might break truthfulness. In the next section, however, Lemma 4.5.3 shows that truncation preserves the incentive compatibility argument of [38], and Lemma 4.5.4 shows that our positive-value mechanism is individually rational for all data contributors as long as analyst cost is not too high relative to data contributor valuations.

It remains to determine the functions $\epsilon(y)$ and $\delta(y)$ for which $\hat{P} \approx cy$ is endogenously differentially private, given the sensitivity of y. The standard deviation of the Laplace noise added to y in Step 6 should be small enough that the analyst is not over- or under-compensated by too much, but it should be an appropriate function of y that permits provable privacy for $\epsilon(y), \delta(y)$ decreasing as $y \to \infty$. Our mechanism leaves

this noise parameter as a general positive function $f(y)$, and we derive functions $\epsilon(y)$ and $\delta(y)$ depending on Δ and f for which we can guarantee endogenous differential privacy in Lemma 4.5.2. Note that in our mechanism, noise is *heteroskedastic* in that the variance of the noise added to y is not uniform across all values of y. This deviates significantly from the usual privacy scenario, and it requires $\delta > 0$ even though the noise for privacy is Laplace.

Lemma 4.5.2 (Privately publishing y). *Define $\epsilon(y) := 2\Delta/f(y-\Delta)$ and $\delta(y) := 1/\exp(1/f'(y-\Delta))$ for increasing, differentiable, concave $f : \mathbb{R}^+ \to \mathbb{R}^+$ with $f'(0) \leq 1$, fixed $\Delta \in \mathbb{R}^+$, and any $y \in \mathbb{R}^+$. For any $c \in \mathbb{R}^+$ and neighboring $\mathbf{v} \sim \mathbf{v}'$, let $y = y(\mathbf{v}, c), y' = (\mathbf{v}', c)$, and let γ and γ' denote random variables with distributions Lap$(f(y))$ and Lap$(f(y'))$, respectively. Then for any $T \subseteq \mathbb{R}$,*

$$\Pr[y + \gamma \in T] \leq \exp(\epsilon(y)) \cdot \Pr[y' + \gamma' \in T] + \delta(y), \text{ and} \tag{4.5.1}$$

$$\Pr[y' + \gamma' \in T] \leq \exp(\epsilon(y)) \cdot \Pr[y + \gamma \in T] + \delta(y). \tag{4.5.2}$$

Proof. We show (4.5.1) for $y < y'$ and then for $y' < y$. Then (4.5.2) follows by symmetry. First note that if $y < y'$, then $\Pr[y + \gamma = t] \leq \Pr[y' + \gamma' = t]$ for t sufficiently far from y', and otherwise their ratios differ maximally at $t = y$. Then it is enough to show $\Pr[y + \gamma = y]/\Pr[y' + \gamma' = y] \leq \epsilon(y)$:

$$\begin{aligned}
\frac{\Pr[y + \gamma = y]}{\Pr[y' + \gamma' = y]} &= \frac{\frac{1}{2f(y)} \cdot \exp(\frac{-0}{f(y)})}{\frac{1}{2f(y')} \cdot \exp(\frac{-(y'-y)}{f(y')})} \\
&\leq \frac{f(y')}{f(y)} \exp(\Delta/f(y')) \\
&\leq \exp(\ln \frac{f(y')}{f(y)} + \frac{\Delta}{f(y')}) \\
&\leq \exp(\frac{\Delta}{f(y)}(f'(y) + 1)).
\end{aligned}$$

The final expression is at most $\exp(\epsilon(y))$ as long as $f'(y) \leq f'(0) \leq 1$, which is true by assumption.

Now consider $y' < y$. Set $t^* = y + f(y)/f'(y-\Delta)$ so that $\int_{t^*}^{\infty} \Pr[y+\gamma = t]dt = \exp(-1/f'(y-\Delta))/2$. Since $\frac{\Pr[y+\gamma=t]}{\Pr[y'+\gamma'=t]}$ increases with $t > y$ and these probabilities decrease with t, it is enough to show that $\Pr[y+\gamma=t^*]/\Pr[y'+\gamma'=t^*] \leq \epsilon(y)$:

$$\begin{aligned}
\frac{\Pr[y+\gamma=t^*]}{\Pr[y'+\gamma'=t^*]} &\leq \frac{\Pr[\gamma = f(y)/f'(y-\Delta)]}{\Pr[\gamma' = f(y)/f'(y-\Delta)+\Delta]} \\
&\leq \frac{f(y')}{f(y)} \exp\left(\frac{f(y)/f'(y-\Delta)+\Delta}{f(y')} - \frac{f(y)/f'(y-\Delta)}{f(y)}\right) \\
&\leq \exp\left(\frac{f(y')+\Delta f'(y')}{f(y')}/f'(y-\Delta) + \frac{\Delta}{f(y')} - 1/f'(y-\Delta)\right) \\
&\leq \exp\left(\frac{\Delta}{f(y')} \cdot \left(f'(y')/f'(y-\Delta)+1\right)\right) \\
&\leq \exp(\frac{2\Delta}{f(y')}),
\end{aligned}$$

where the final inequality follows from f increasing and concave. □

Note that $y(\mathbf{v}, c)$ is the *unique* y in the privacy support of $\mathcal{M}(\mathbf{d}, \mathbf{v}, c)$ for any fixed inputs. Therefore, Lemma 4.5.2 establishes endogenous differential privacy (Definition 4.3.1) of $\hat{P} = c(y + \text{Lap}(f(y)))$ for the ϵ, δ in the lemma statement. Endogenous differential privacy of the overall mechanism (Theorem 4.4.5) is an immediate corollary, assuming $\mathcal{M}_q$ is differentially private in the standard sense and using the general composition theorem for differential privacy.

Privacy for Non-Logarithmic Preferences

Mechanism 5 relies on the assumption each data contributor's utility for the level of y provided by the mechanism is represented by some $v_i(y) = v_i \ln(y+1)$. This choice of logarithmic utility functions was the convenient one, since it allows us to easily bound the sensitivity of y using a simple truncation rule. However, many other non-negative, increasing, concave functions of y may be appropriate models of the utility to data contributors of y.

Consider the case that each data contributor has valuation function $v_i(y) = v_i y^{1/a}$ for $a > 1$. As before, we first truncate the v_i so that $\bar{v}_i := \max(v_i, v_{\max})$ for some $v_{\max}$

to be determined later to adequately control the sensitivity of the consumer surplus maximizing level of privacy $y(\mathbf{v}, c) := \arg\max_{y \geq 0} \sum \bar{v}_i(y) - cy$. Then we have:

$$
\begin{aligned}
y(\mathbf{v}, c) &= \left(\frac{\sum \bar{v}_i'}{ac} \right)^{\frac{a}{a-1}} \\
|y(\mathbf{v}_{-i} \| v_i, c) - y(\mathbf{v}, c)| &= (ac)^{\frac{a-1}{a}} \cdot |(v_i' + \sum_{j \neq i} v_j)^{\frac{a}{a-1}} - (v_i + \sum_{j \neq i} v_j)^{\frac{a}{a-1}}| \\
&\leq (ac)^{\frac{a-1}{a}} \cdot \left((v_{\max} + \sum_{j \neq i} v_j)^{\frac{a}{a-1}} - (\sum_{j \neq i} v_j)^{\frac{a}{a-1}} \right) \\
&\leq (ac)^{\frac{a-1}{a}} \cdot \left((n v_{\max})^{\frac{a}{a-1}} - ((n-1) v_{\max})^{\frac{a}{a-1}} \right) \\
&\leq \left(\frac{(n-1) v_{\max}}{ac} \right)^{\frac{a}{a-1}} \cdot \left((1 + \frac{1}{n-1})^{\frac{a}{a-1}} - 1 \right) \\
&\leq \left(\frac{(n-1) v_{\max}}{ac} \right)^{\frac{a}{a-1}} \cdot \left(\frac{\frac{a}{a-1}(1 + 1/n)^{\frac{1}{a-1}}}{n} \right)
\end{aligned}
$$

In the case that $a \geq 2$, we have $|y(\mathbf{v}_{-i} \| v_i, c) - y(\mathbf{v}, c)| \leq (n v_{\max}/c)^{\frac{a}{a-1}}/n$. Then if we set $v_{\max} = c\Delta^{\frac{a-1}{a}}/n^{1/a}$, the sensitivity of y is Δ for some fixed Δ as before, and privacy follows as in Lemma 4.5.2. Incentive compatibility also follows as in Lemma 4.5.3 below. Individual rationality, however, does not appear to hold for agents with low privacy sensitivity. In particular, when $v_i = 0$, i will always be charged $p_i > 0$ whenever $\sum_{j \neq i} v_j > 0$. With the exception of individual rationality, the other properties of Lemma 4.5.5 hold with $\Theta(n)$ replaced with $\Theta(n^{a/(a-1)})$. It remains an open problem to identify further classes of valuation functions for which our mechanism or variants of it satisfy all desired properties for endogenous privacy markets.

4.5.2 Market Properties of Mechanism 5

Recall notation for the optimal level of privacy, and define another function for individual taxes computed by the mechanism on inputs $\mathbf{v} \in (\mathbb{R}^+)^n, c \in \mathbb{R}^+$, recalling

that $\bar{v}_i := \min(v_i, c \cdot \Delta)$:

$$y(\mathbf{v}, c) := \left(\frac{\sum \bar{v}_i}{c} - 1 \right)^+$$

$$p_i(\mathbf{v}, c) := cy(\mathbf{v}, c) - \sum_{j \neq i} \bar{v}_j \ln(y(\mathbf{v}, c) + 1) + \max_{y_{-i}} \left(\sum_{j \neq i} \bar{v}_j \ln(y_{-i} + 1) - \tfrac{n-1}{n} c y_{-i} \right).$$

Lemma 4.5.3. *Mechanism 5 is incentive compatible.*

Proof. Fix any $i \in [n]$ and $v_i^* \in \mathbb{R}^+$, and denote the utility of i as

$$\begin{aligned} U_i(\mathbf{v}, c) &:= v_i^*(y(\mathbf{v}, c)) - p_i(\mathbf{v}, c) \\ &= (v_i^* + \sum_{j \neq i} \bar{v}_j) \ln(y(\mathbf{v}, c) + 1) - cy(\mathbf{v}, c) - \max_{y_{-i} \geq 0} \left(\sum_{j \neq i} \bar{v}_j \ln(y_{-i} + 1) - \tfrac{n-1}{n} c y_{-i} \right). \end{aligned}$$

For incentive compatibility, we need to show that for any $\mathbf{v}_{-i} \in (\mathbb{R}^+)^{n-1}$ and $c \in \mathbb{R}^+$, we have $v_i^* \in \arg\max_{v_i} U_i(\mathbf{v}_{-i} \| v_i, c)$.

Observing that $\max_{y_{-i} \geq 0} (\sum_{j \neq i} \bar{v}_j \ln(y_{-i} + 1) - \frac{n-1}{n} c y_{-i})$ has no dependence on v_i, we see that $U_i(\mathbf{v}, c)$ increases with y until $y = (v_i^* + \sum_{j \neq i} \bar{v}_j)/c - 1$. Therefore, by declaring $v_i = v_i^*$, $y(\mathbf{v}, c)$ coincides with i's optimal value of y if $v_i^* \leq c \cdot \Delta$, and it it maximizes i's utility subject to truncation otherwise. □

We prove individual rationality for a restricted set of preference inputs:

Lemma 4.5.4. *Mechanism 5 is individually rational on inputs* $\mathbf{v} \in (\mathbb{R}^+)^n, c \leq \sum \bar{v}_i / e$.

Proof. First note that $v_i \ln(y(\mathbf{v}, c) + 1) = v_i \ln \frac{\sum \bar{v}_i}{c} \geq \bar{v}_i$, so it is enough to show that $p_i(\mathbf{v}, c) \leq \bar{v}_i$. Bound p_i as follows:

$$\begin{aligned} p_i(\mathbf{v}, c) &= cy(\mathbf{v}, c) - \sum_{j \neq i} \bar{v}_j \ln(y(\mathbf{v}, c) + 1) + \max_{y_{-i}} \left(\sum_{j \neq i} \bar{v}_j \ln(y_{-i} + 1) - \frac{n-1}{n} c y_{-i} \right) \\ &= (\sum \bar{v}_i - c) - \sum_{j \neq i} \bar{v}_j \ln \frac{\sum \bar{v}_i}{c} + \sum_{j \neq i} \bar{v}_j \left(\left(\ln \frac{\sum_{j \neq i} \bar{v}_j}{\frac{n-1}{n} c} \right) - 1 \right) + \frac{n-1}{n} c \\ &= \bar{v}_i - \frac{c}{n} - \sum_{j \neq i} \bar{v}_j (1 + \ln \frac{\sum \bar{v}_i}{c} - \ln \frac{\sum_{j \neq i} \bar{v}_j}{\frac{n-1}{n} c}) \\ &= \bar{v}_i - \frac{c}{n} - \sum_{j \neq i} \bar{v}_j \ln \frac{e^{\frac{n-1}{n}} \sum \bar{v}_i}{\sum_{j \neq i} \bar{v}_j}. \end{aligned}$$

Since the $\bar{v}_i$ are nonnegative, it is enough to show that $e\frac{n-1}{n} \geq 1$, which clearly holds for any $n \geq 2$. □

Note that the conditions for individual rationality hold whenever $c \leq n$ and $\sum \min(v_i, c \cdot \Delta)/n \geq e$. The mechanism could easily be modified to enforce $c \leq n$ and in many scenarios it may be a reasonable to assume a distribution on v_i satisfying the latter requirement.[3] Note that these qualifications affect the positive-value negative result (Theorem 4.2.3) since a mechanism running on this restricted set of inputs cannot output a privacy level with arbitrarily small value to the data contributors. However, we can modify our negative results for the restricted set of inputs as follows. Note that $\mathbf{v} = (0, \ldots, 0, ce)$ satisfies $c \leq \sum \bar{v}_i/e$ for $\Delta \geq 1$. Mechanism 5 outputs $y = \sum \bar{v}_i/c - 1 = e - 1$ on $\mathbf{v}, c$, and $\sum v_i(e-1) = ce \ln(e - 1 + 1) = ce$. Then with standard differential privacy, individually rational mechanisms running on the restricted set of inputs can pay the analyst at most ce with almost $1/2$ probability.[4] Endogenous privacy of course escapes this bound.

Parameters for Accuracy and Efficiency. The internally chosen consumer surplus maximizing privacy level y is noiseless, so its Pareto efficiency follows immediately by the arguments of [38] whenever truncation is avoided, i.e., when each $v_i \leq \Delta \cdot c$. The chosen level of privacy differs from the Pareto efficient level by $\sum(v_i - \bar{v}_i)/c$, which grows with the total amount truncated. If we expect constant v_i and c, we

[3] If qualified individual rationality is undesired, one might consider applying the propose-test-release strategy of [26] and aborting as a first step if the conditions of Lemma 4.5.4 are not met. However, note that $\sum \bar{v}_i/c$ has sensitivity Δ. Adding noise $\text{Lap}(\Delta/\epsilon)$ for differential privacy (although there would be some modifications to make this endogenously private) would overwhelm the threshold e when $\Delta = \omega(1)$ as in the usual case, so this strategy seems unlikely to work directly. We leave the issue of unqualified individual rationality as a question for future work.

[4] Recall that Theorem 4.2.3 concerns mechanisms with arbitrarily small value to the data contributors, i.e. for any $P > 0$ there exists some $\mathbf{v} \in \mathcal{V}^n$ with $\sum v_i(\epsilon_i) < P$ for any $\boldsymbol{\epsilon}$ in the support of $\mathcal{M}$ on $\mathbf{v}$. If instead we only require $\mathbf{v}$ with $\sum v_i(\epsilon_i) < P_0$ for some fixed rather than arbitrarily small P_0, then such a mechanism that satisfies standard differential privacy and individual rationality can never pay the analyst more than P_0 and maintain a balanced budget, and it can pay the analyst more than P_0 with probability at most $\exp(\sum \epsilon_{i,\text{inf}})/2$ while maintaining $\Pr[\hat{P} > \sum p_i] \leq 1/2$. This argument applies to the above case for $P_0 = ce$.

should set $\Delta = \omega(1)$ to avoid truncation.

Accuracy of the released statistic $\hat{R}$ is inherited directly from $\mathcal{M}_q$ with the chosen privacy parameters, and the overall mechanism is endogenously private. For v_i, c constant, we will expect y linear in n, so we should have $\Delta = o(f(n))$ to ensure that ϵ decreases with n, and $f'(n) = o(1/\ln n)$ will ensure $\delta(y) \leq 1/\text{poly}(n)$ for small enough δ_q. Since $y = 0$ results in zero privacy utility, we should have $\epsilon(0) = \infty$ with $\epsilon(y)$ finite for $y > 0$, which we get with $f(-\Delta) = 0$ and strictly increasing. The budget balances in expectation since $\sum p_i \geq cy$ and $\hat{P} = c(y + \text{Lap}(f(y)))$. To ensure that the mechanism is not likely to run too great a deficit, we should have $f(n) = o(n)$. In summary:

Lemma 4.5.5. *Let $\Delta = \omega(1)$, let f increasing, differentiable, and concave with $f(\Theta(n)) = \omega(\Delta)$, $f(\Theta(n)) = o(n)$, $f(-\Delta) = 0$, $f'(0) \leq 1$, and $f'(\Theta(n)) = o(1/\ln n)$, and let $\mathcal{M}_q$ be accurate for $\epsilon_q = \Theta(\Delta/f(\Theta(n)))$ and $\delta_q = 1/\text{poly}(n)$. If $v_i, c_i = \Theta(1)$, then Mechanism 5 is accurate, incentive compatible, individually rational, Pareto efficient, budget-balanced in expectation with deficit $\geq t$ with probability $\leq \exp(-t/f(\Theta(n)))/2$, and endogenously private with $\epsilon = o(1)$ and $\delta = 1/\text{poly}(n)$.*

4.5.3 Negative Results for Exogenous Privacy in Endogenous Markets

Before proving Theorems 4.2.2 and 4.2.3 from Section 4.2, we first present the original theorem statement and proof of the negative result of [34] for comparison. Recall the exogenous privacy definition used by [34] (Definition 4.2.1) and the definitions of other properties from Section 4.4.1. We give the proof using the syntax of Model 3, noting that the framework of [34] allows mechanisms to output a separate guarantee ϵ_i to each data contributor $i \in [n]$, which we discuss after presenting their original result.

Theorem (Theorem 5.1 from [34]). *If data contributors' preferences for privacy may be arbitrarily negative, then no individually rational direct revelation*[5] *mechanism*

[5] *A* direct revelation *mechanism is one that request players true types. This term sometimes refers*

$\mathcal{M}$ can protect the privacy of data contributor preferences and promise nontrivial accuracy (unless the analyst always pays an infinite amount).

Proof [34]. If $\mathcal{M}$ is nontrivially accurate and private with respect to private data, then $\sum \epsilon_i \geq \ln 4/3$ for the ϵ_i output by $\mathcal{M}$ on *any* set of inputs. (See [34] for accuracy definition and calculation.)

Assume for simplicity that $\mathcal{V} \leftrightarrow \mathbb{R}^+$ with $v_i(\epsilon_i) = -v_i \cdot \epsilon_i$. Consider the $\epsilon_i, \hat{P}, p_i$ output by $\mathcal{M}$ on arbitrary inputs $\mathbf{v} \in \mathcal{V}^n, \mathbf{d} \in \mathcal{D}^n, c \in \mathcal{C}$ with $v_{\min} := \min v_j$. By individual rationality (and balanced budget):

$$-\hat{P} = \sum -p_i \geq \sum -v_i(\epsilon_i) \geq v_{\min} \sum \epsilon_i \geq v_{\min} \ln 4/3, \tag{4.5.3}$$

and so $\Pr[-\hat{P} < v_{\min} \ln 4/3] = 0$ for $\hat{P} \leftarrow \mathcal{M}(\mathbf{v}, \mathbf{d}, c)$. Let $P := v_{\min} \ln 4/3$. Then for *any* inputs $\mathbf{v}', \mathbf{d}'$, we note that $\mathbf{v}, \mathbf{d}$ can be obtained by a sequence of single row changes for rows $i = 1, \ldots, n$, and so by (worst-case, per-row) privacy with respect to data contributors' private data and privacy preferences, we have:

$$\Pr[-\hat{P} < P \mid \hat{P} \leftarrow \mathcal{M}(\mathbf{v}', \mathbf{d}', c)] \leq \exp(\textstyle\sum \epsilon_i) \cdot \Pr[-\hat{P} < P \mid \hat{P} \leftarrow \mathcal{M}(\mathbf{v}, \mathbf{d}, c)] = 0. \tag{4.5.4}$$

Since $\mathbf{v}$ could have been chosen with arbitrarily large $v_{\min}$, it follows that $\mathcal{M}$ can never charge an analyst less than *any* finite payment P. □

When mechanisms need not be private with respect to preferences as in the insensitive value model of [34], it may be useful to output a separate guarantee ϵ_i to each data contributor $i \in [n]$. However, it is unclear how a mechanism that is private with respect to preferences could fulfill heterogeneous guarantees ϵ_i, and so our new negative results focus on mechanisms in Model 3 that output a single privacy guarantee $\epsilon = y \in \mathcal{Y} := \mathbb{R}^+$ (with $\delta = 0$). Furthermore, heterogeneous privacy guarantees ϵ_i are

to mechanisms that incentivize players to report their true types, but note that incentive compatibility *is not assumed in this proof (or the new proofs that follow).*

not needed for the above negative result. In fact, the theorem holds for any arbitrarily large lower bound in Inequality (4.5.3). Since accuracy and arbitrarily large privacy loss costs are only required in this step, we can replace them with the following two requirements:

> $\mathcal{M}$ *has imperfect privacy*, i.e., it is differentially private (Definition 4.2.1) with $0 < \epsilon_{\text{inf}} := \inf\{\epsilon \leftarrow \mathcal{M}(\mathbf{v}, \mathbf{d}, c), \mathbf{v} \in \mathcal{V}^n, \mathbf{d} \in \mathcal{D}^n, c \in \mathcal{C}\}$.
>
> $\mathcal{V}$ *are arbitrarily negative*, i.e., every $v \in \mathcal{V}$ is monotonically nonincreasing in ϵ, and for any $B, \epsilon > 0$, there exists some $v \in \mathcal{V}$ such that $-v(\epsilon) > B$.

Observe that in order for $\mathcal{M}$ to guarantee perfect privacy $\epsilon = 0$, it would have to ignore all of its inputs, so necessarily it could provide no accuracy whatsoever. The requirements of Theorem 4.2.2, (1) are therefore weaker than those of Theorem 5.1 from [34], so the new result is stronger. [6]

We also note that the proof in [34] implicitly assumes that $\mathcal{M}$ must be budget-balanced. We make this requirement explicit, and Theorem 4.2.2, (2) shows that when we relax the requirement somewhat, the result stays mostly in tact, with a probabilistic dependency on ϵ_{inf}. Note that as privacy preferences become arbitrarily negative, reasonable mechanisms should output very small ϵ. In particular, if $\epsilon_{\text{inf}} = o(1/n)$, then the theorem roughly implies that if an analyst has any fixed (arbitrarily large) budget, we expect $\mathcal{M}$ on any fixed input to charge the analyst more than his budget with probability almost 1/2.

Theorem (4.2.2). *Consider any mechanism $\mathcal{M}$ in Model 3 with imperfect privacy, and assume data contributors' preferences are arbitrarily negative.*

[6] [80] also strengthened the [34] negative result, although in different ways. They show that a mild incentive compatibility property is enough to preclude very minimal accuracy. They model the cost of privacy loss as a function of all of the inputs and outputs of the mechanism, and this function is bounded by (data-dependent) privacy characteristics of the mechanism. However, their mechanisms do not choose a privacy level internally, so their results are not applicable to the problem studied in this work.

1. *If $\mathcal{M}$ is individually rational and budget-balanced, then for any finite $B > 0$, we have that $\mathcal{M}$ running on any fixed inputs always charges the analyst $-\hat{P} > B$.*

2. *If $\mathcal{M}$ is individually rational and likely budget-balanced, then for any finite $B > 0$, we have that $\mathcal{M}$ running on any fixed inputs charges the analyst $-\hat{P} \leq B$ with probability at most $\exp(n \cdot \epsilon_{\text{inf}})/2$.*

Proof. Fix any $B > 0$ and $\epsilon \in (0, \epsilon_{\text{inf}})$, and choose $v \in \mathcal{V}$ such that $-v(\epsilon) > B$. Then choose any inputs $\mathbf{v} \in \mathcal{V}^n, \mathbf{d} \in \mathcal{D}^n, c \in \mathcal{C}$ such that $v_1 = v$, and consider the p_i output by $\mathcal{M}$ on $\mathbf{v}, \mathbf{d}, c$. By monotonicity and individual rationality:

$$\sum -p_i \geq \sum -v_i(\epsilon_{\text{inf}}) \geq \sum -v_i(\epsilon) > B,$$

and so $\Pr[-\hat{P} \leq B] \leq \Pr[-\hat{P} < -\sum p_i]$ for $(\hat{P}, \mathbf{p}) \leftarrow \mathcal{M}(\mathbf{v}, \mathbf{d}, c)$. Then for *any* inputs $\mathbf{v}', \mathbf{d}'$, we note that $\mathbf{v}, \mathbf{d}$ can be obtained by a sequence of single row changes for rows $i = 1, \ldots, n$, and so by privacy, we have:

$$\begin{aligned}\Pr[-\hat{P} \leq B \mid \hat{P} \leftarrow \mathcal{M}(\mathbf{v}', \mathbf{d}', c)] &\leq \exp(\textstyle\sum \epsilon_{\text{inf}}) \cdot \Pr[-\hat{P} \leq B \mid \hat{P} \leftarrow \mathcal{M}(\mathbf{v}, \mathbf{d}, c)] \\ &\leq \exp(n \cdot \epsilon_{\text{inf}}) \cdot \Pr[-\hat{P} < -\textstyle\sum p_i \mid (\hat{P}, \mathbf{p}) \leftarrow \mathcal{M}(\mathbf{v}, \mathbf{d}, c)].\end{aligned}$$

If $\mathcal{M}$ must be strictly budget-balanced, the last term above is zero, and it follows that $\mathcal{M}$ can never charge the analyst at most any finite payment B, completing (1). If $\mathcal{M}$ is likely budget-balanced, this final probability is bounded by $1/2$ and the second conclusion (2) follows. □

When a privacy guarantee has positive value to data contributors, note that any standard moneyless differentially private mechanism is trivially individually rational, and such a mechanism is accurate and differentially private in the usual sense for some exogenous ϵ. Strictly speaking, this assumption of nonnegative utility for privacy circumvents the impossibility result of [34], albeit in a way that does not make any progress towards determining a sensible value of ϵ endogenously. However, we are able

to provide results analogous to Theorem 4.2.2 in assuming positive privacy preferences that roughly show that *only* these trivial mechanisms that do not make any monetary transfers are possible. Now instead of assuming imperfect privacy and arbitrarily negative preferences, Theorem 4.2.3 requires that:

> $\mathcal{V}$ *are arbitrarily small,* i.e. for any $R > 0$, there exist some $\mathbf{d} \in \mathcal{D}^n, \mathbf{v} \in \mathcal{V}^n, c \in \mathcal{C}$ such that $\sum v_i(\epsilon) < R$ for any ϵ in the privacy support of $\mathcal{M}(\mathbf{d}, \mathbf{v}, c)$.

Following the argument from Theorem 4.2.2, the second part of this theorem roughly says that for sufficiently small ϵ_{inf}, the analyst will fall short of any arbitrarily small target revenue R with probability almost $1/2$. The proof follows the structure of the previous proof almost exactly and is given only for completeness.

Theorem (4.2.3). *Consider any private mechanism $\mathcal{M}$ in Model 3, and assume data contributors' preferences are positive and arbitrarily small.*

1. *If $\mathcal{M}$ is individually rational and budget-balanced, then for any fixed $R > 0$, we have that $\mathcal{M}$ running on any fixed inputs always pays the analyst $\hat{P} < R$.*

2. *If $\mathcal{M}$ is individually rational and likely budget-balanced, then for any fixed $R > 0$, we have that $\mathcal{M}$ running on any fixed inputs pays the analyst $\hat{P} \geq R$ with probability at most* $\exp(n \cdot \epsilon_{\text{inf}})/2$.

Proof. Fix any $R > 0$, and let $\mathbf{d} \in \mathcal{D}^n, \mathbf{v} \in \mathcal{V}^n, c \in \mathcal{C}$ be such that $\sum v_i(\epsilon) < R$ for any ϵ in the privacy support of $\mathcal{M}(\mathbf{d}, \mathbf{v}, c)$. By monotonicity and individual rationality:

$$\sum p_i \leq \sum v_i(\epsilon) < R,$$

and so $\Pr[\hat{P} \geq R] \leq \Pr[\hat{P} \geq \sum p_i]$ for $(\hat{P}, \mathbf{p}) \leftarrow \mathcal{M}(\mathbf{d}, \mathbf{v}, c)$. Then for any inputs $\mathbf{v}', \mathbf{d}'$, we note that $\mathbf{v}, \mathbf{d}$ can be obtained by a sequence of single row changes for rows $i = 1, \ldots, n$, and so by privacy, we have:

$$\begin{aligned}\Pr[\hat{P} \geq R \mid \hat{P} \leftarrow \mathcal{M}(\mathbf{v}', \mathbf{d}', c)] &\leq \exp(\textstyle\sum \epsilon_{\text{inf}}) \cdot \Pr[\hat{P} \geq R \mid \hat{P} \leftarrow \mathcal{M}(\mathbf{v}, \mathbf{d}, c)] \\ &\leq \exp(n \cdot \epsilon_{\text{inf}}) \cdot \Pr[\hat{P} \geq \textstyle\sum p_i \mid (\hat{P}, \mathbf{p}) \leftarrow \mathcal{M}(\mathbf{v}, \mathbf{d}, c)].\end{aligned}$$

If $\mathcal{M}$ must be strictly budget-balanced, the last term above is zero, and it follows that $\mathcal{M}$ can never pay the analyst at least any minimum payment R, completing (1). If $\mathcal{M}$ is likely budget-balanced, this final probability is bounded by $1/2$ and the second conclusion (2) follows. □

Underlying the first part of each of the above theorems is the zero probability event that a strictly budget-balanced mechanism pays the analyst more than it can charge data contributors for some fixed inputs and corresponding privacy guarantees (or charges the analyst less than required to compensate data contributors for privacy loss). By privacy, this event must remain impossible for any set of inputs. When we relax the balanced budget assumption, this arbitrary bound on analyst charge/payment carries over to any other set of inputs through the strongest privacy guarantees ϵ_{inf} *if the standard notion of input-independent differential privacy is used.* The probability bound in Theorem 4.2.3, for example, arises from collapsing the bounds in probabilities of the event that the mechanism pays the analyst $\hat{P} \geq R$ on neighboring pairs of databases in the chain of databases $\mathbf{v}^{(0)} = \mathbf{v}', \ldots, \mathbf{v}^{(j)} = (v_1, \ldots, v_j, v'_{j+1}, \ldots, v'_n), \ldots, \mathbf{v}^{(n)} = \mathbf{v}$ where $\mathbf{v}$ is such that $\sum v_i(\epsilon) < R$ and $\mathbf{v}'$ is arbitrary. With endogenous privacy, the reference databases and not ϵ_{inf} determine the probability differences across neighboring databases. Let $\epsilon^{(j)}$ denote the minimum privacy parameter in the support of $\mathcal{M}(\mathbf{v}^{(j)})$. Then endogenous privacy yields the bound:

$$\begin{aligned}
\Pr_{[}\hat{P} \geq R \mid \hat{P} \leftarrow \mathcal{M}(\mathbf{v}')] &\leq \exp(\epsilon^{(1)}) \Pr[\hat{P} \geq R \mid \hat{P} \leftarrow \mathcal{M}(\mathbf{v}^{(1)})] \\
&\ldots \\
&\leq \exp(\textstyle\sum \epsilon^{(i)}) \Pr[\hat{P} \geq R \mid \hat{P} \leftarrow \mathcal{M}(\mathbf{v})] \\
&\leq \exp(\textstyle\sum \epsilon^{(i)})/2.
\end{aligned}$$

For $\mathbf{v}$ small enough for $\sum v_i(\epsilon) < R$, we expect the $\epsilon^{(i)}$ for large i to be large, making this bound loose if not trivial. We also get a similar bound for every permutation π

on $[n]$ identifying a different chain of hybrid databases between $\mathbf{v}'$ and $\mathbf{v}$, but in all of these cases, the $\epsilon^{(i)}$ for large i should be large.

Future Work On Privacy Markets

The new class of endogenously private mechanisms is based on a special case of the market for Pareto efficient allocation of goods described in [38]. Their framework is actually much more general, allowing for multiple public goods with different production prices. This generality could be readily exploited to create markets for privacy with multiple analysts, possibly with different levels of ϵ for different databases or different queries.

A downside of applying the public goods allocation problem to the privacy setting is that the former implicitly assumes a prevalence of producers who pressure each other to not overstate their production costs. Without this assumption, the mechanism is incentive compatible for consumers (data contributors) but *not* for the producer (analyst). Generalizations allowing for multiple analysts that compete for access to the data may partially resolve this concern. Nonetheless, future works should consider the worst-case effects of an analyst who strategically misreports his accuracy costs, or, ideally, the mechanism can be modified to be incentive compatible on both sides.

Laplace mechanism and generalizations guarantee ϵ-differential privacy for all i by adding noise to the true query answer. This fact influences our perspective of privacy as a public good, and so we focus on mechanisms providing a single privacy guarantee for all data contributors. Outputting a single noisy query answer while guaranteeing heterogeneous ϵ_i is an interesting question for future research.

While the view of privacy as a positively-valued good may be appropriate for settings in which data contributors have already divulged their private data to an entity that they can expect will try to profit from it, it remains a very interesting open question whether the endogenous differential privacy relaxation alone is enough

to circumvent the negative result of [34] when data contributors have disutility for imperfect privacy. The techniques used in Section 4.4 rely heavily on the view of privacy as a public good, but we may hope that other techniques could yield incentive compatible endogenous privacy markets when loss of privacy is costly as is assumed in most of the prior literature.

CHAPTER V

CONCLUSION AND FUTURE DIRECTIONS

The connection between differential privacy and game theory was first explored in [70], who developed the powerful differentially private *exponential mechanism* for setting the price in a digital goods auction. In the same way that the exponential mechanism protects the privacy of buyers' bids, it also bounds their incentives to lie about their values for the good, making approximate incentive compatibility a convenient and powerful side effect of privacy.

[81] surveys several subsequent works exploring this connection between privacy and game theory. A number of these works also use the fact that privacy can be used as a tool to limit how much a player can gain by misreporting his type, and truthfulness as an approximately dominant strategy is an almost automatic consequence of differential privacy. However, [78] argued against the idea of approximate truthfulness that results from differentially private mechanisms, noting that even if a player cannot gain *much* by falsely reporting his type, it may not be safe to assume he will report truthfully as a default, particularly when it is easy to see that another strategy is better (if only slightly). Since differential privacy is so dependent on random noise, settling on economically sensible notions of approximate equilibria and approximate truthfulness is important for future work at the intersection of differential privacy and game theory.

In ongoing work [60], we consider the issue of private computation of *correlated equilibrium*. The notion of correlated equilibria generalizes the notion of Nash equilibria in that players need not choose their strategies independently. It is easiest to think of a correlated equilibrium as a traffic light: each player has some probability of crossing an intersection, but these probabilities are not independent. We show that

this is not possible for general games, in which equilibria may be highly sensitive to a single player's type. Another recent work proposed a private recommender mechanism that selects an approximate correlated equilibrium in large games where the outcome observed by a single player is insensitive to the reported type of any other player [50]. Rather than impose some a priori bound on how much an individual can affect the utility of another player, we investigate more generally the qualities of games whose sets of correlated equilibria are highly sensitive to a single player's type, and we seek to develop private equilibrium computation methods for games whose equilibria are not too sensitive to individual players. Better understanding the sensitivity of equilibria to small changes in a game will likely yield many nice privacy results, but this question is also of interest to game theorists independent of privacy goals.

Although we cannot hope to be able to privately compute equilibria for general games, computational game theory is rich with privacy applications. Differential privacy as a *feature* of methods for equilibrium computation is a fascinating and wide open area for future research.

REFERENCES

[1] Adamczak, R., Litvak, A., Pajor, A., and Tomczak-Jaegermann, N., "Quantitative estimates of the convergence of the empirical covariance matrix in logconcave ensembles," *J. Amer. Math. Soc.*, vol. 233, pp. 535–561, 2011.

[2] Anandkumar, A., Foster, D. P., Hsu, D., Kakade, S. M., and Liu, Y., "A spectral algorithm for latent Dirichlet allocation," *arXiv preprint arXiv:1204.6703*, 2012.

[3] Anandkumar, A., Ge, R., Hsu, D., Kakade, S. M., and Telgarsky, M., "Tensor decompositions for learning latent variable models," *arXiv preprint arXiv:1210.7559*, 2012.

[4] Anshelevich, E., Dasgupta, A., Tardos, É., and Wexler, T., "Near-optimal network design with selfish agents," *Theory of Computing*, vol. 4, no. 1, pp. 77–109, 2008.

[5] Arora, S., Ge, R., Moitra, A., and Sachdeva, S., "Provable ICA with unknown gaussian noise, and implications for gaussian mixtures and autoencoders," *arXiv:1206.5349*, 2012.

[6] Balcan, M.-F., Blum, A., and Mansour, Y., "Improved equilibria via public service advertising," in *Proceedings of the twentieth Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, (Philadelphia, PA, USA), pp. 728–737, Society for Industrial and Applied Mathematics, 2009.

[7] Balcan, M.-F., Blum, A., and Mansour, Y., "Circumventing the price of anarchy: Leading dynamics to good behavior," *SIAM Journal on Computing*, vol. 42, no. 1, pp. 230–264, 2013.

[8] Balcan, M.-F., Krehbiel, S., Piliouras, G., and Shin, J., "Near-optimality in covering games by exposing global information," *ACM Trans. Econ. Comput.*, vol. 2, pp. 13:1–13:22, Oct. 2014.

[9] Bar-Ness, J. W., Carlin, Y., and Steinberger, M. L., "Bootstrapping adaptive interference cancelers - some practical limitations," in *Proc. the Globecom Conference*, pp. 1251–1255, 1982.

[10] Bendlin, R., Krehbiel, S., and Peikert, C., "How to share a lattice trapdoor: Threshold protocols for signatures and (h)ibe," in *Applied Cryptography and Network Security* (Jacobson, M., Locasto, M., Mohassel, P., and Safavi-Naini, R., eds.), vol. 7954 of *Lecture Notes in Computer Science*, pp. 218–236, Springer Berlin Heidelberg, 2013.

[11] Bengio, Y., "Learning deep architectures for AI," *Foundations and trends in Machine Learning*, vol. 2, no. 1, pp. 1–127, 2009.

[12] BenOr, M., Goldwasser, S., and Wigderson, A., "Completeness theorems for non-cryptographic fault-tolerant distributed computation," *20th STOC*, pp. 1–10, 1988.

[13] Biswal, B. and Ulmer, J., "Blind source separation of multiple signal sources of fMRI data sets using independent component analysis," *J. of Computer Assisted Tomography*, vol. 23, pp. 265–271, 1999.

[14] Blum, A., Ligett, K., and Roth, A., "A learning theory approach to non-interactive database privacy," in *Proceedings of the 40th annual ACM symposium on Theory of computing*, pp. 609–618, ACM, 2008.

[15] Buchbinder, N., Lewin-Eytan, L., Naor, J. S., and Orda, A., "Non-cooperative cost sharing games via subsidies," *Theory of Computing Systems*, vol. 47, pp. 15–37, July 2010.

[16] Campos-Naóez, E., Garcia, A., and Li, C., "A game-theoretic approach to efficient power management in sensor networks," *Operation Research*, vol. 56, no. 3, pp. 552–561, 2008.

[17] Cardinal, J. and Hoefer, M., "Non-cooperative facility location and covering games," *Theoretical Computer Science*, vol. 411, no. 16-18, pp. 1855–1876, 2010.

[18] Cardinal, J. and Hoefer, M., "Selfish service installation in networks," in *Proceedings of the 2rd international conference on Internet and network economics* (Spirakis, P., Mavronicolas, M., and Kontogiannis, S., eds.), vol. 4286 of *WINE'06*, pp. 174–185, Springer Berlin Heidelberg, 2006.

[19] Chaudhuri, K., Sarwate, A., and Sinha, K., "Near-optimal differentially private principal components," in *Advances in Neural Information Processing Systems 25*, pp. 998–1006, 2012.

[20] Chen, Y., Chong, S., Kash, I. A., Moran, T., and Vadhan, S., "Truthful mechanisms for agents that value privacy," in *Proceedings of the Fourteenth ACM Conference on Electronic Commerce*, EC '13, (New York, NY, USA), pp. 215–232, ACM, 2013.

[21] Chvatal, V., "A greedy heuristic for the set-covering problem," *Mathematics of Operations Research*, vol. 4, no. 3, pp. 233–235, 1979.

[22] Clarke, E. H., "Multipart pricing of public goods," *Public Choice*, vol. 11, no. 1, pp. 17–33, 1971.

[23] Comon, P., "Independent component analysis, a new concept?," *Signal Process.*, vol. 36, pp. 287–314, Apr. 1994.

[24] Demaine, E. D. and Zadimoghaddam, M., "Constant price of anarchy in network creation games via public service advertising," in *Algorithms and Models for the Web-Graph (WAW)*, pp. 122–131, 2010.

[25] Dwork, C., "Differential privacy," *Automata, languages and programming*, pp. 1–12, 2006.

[26] Dwork, C. and Lei, J., "Differential privacy and robust statistics," in *Proceedings of the 41st annual ACM symposium on Theory of computing*, pp. 371–380, ACM, 2009.

[27] Dwork, C., McSherry, F., Nissim, K., and Smith, A., "Calibrating noise to sensitivity in private data analysis," in *Theory of Cryptography* (Halevi, S. and Rabin, T., eds.), vol. 3876 of *Lecture Notes in Computer Science*, pp. 265–284, Springer Berlin Heidelberg, 2006.

[28] Escoffier, B., Gourves, L., and Monnot, J., "On the impact of local taxes in a set cover game," in *Structural Information and Communication Complexity (SIROCCO)*, vol. 6058, pp. 2–13, Springer-Verlag New York Inc, 2010.

[29] Fabrikant, A., Luthra, A., Maneva, E., Papadimitriou, C., and Shenker, S., "On a network creation game," in *Proceedings of the twenty-second annual symposium on Principles of distributed computing (PODC)*, pp. 347–351, ACM, 2003.

[30] Fleischer, L. and Lyu, Y.-H., "Approximately optimal auctions for selling privacy when costs are correlated with data," in *ACM Conference on Electronic Commerce*, pp. 568–585, 2012.

[31] Fox, M., Piliouras, G., and Shamma, J., "Medium and long-run properties of linguistic community evolution," in *9th International Conference on the Evolution of Language (Evolang IX)*, pp. 110–118, March 2012.

[32] Frieze, A. M., Jerrum, M., and Kannan, R., "Learning linear transformations," in *FOCS*, pp. 359–368, 1996.

[33] Ghosh, A. and Ligett, K., "Privacy and coordination: computing on databases with endogenous participation," in *Proceedings of the Fourteenth ACM Conference on Electronic Commerce*, EC '13, (New York, NY, USA), pp. 543–560, ACM, 2013.

[34] Ghosh, A. and Roth, A., "Selling privacy at auction," in *Proceedings of the 12th ACM Conference on Electronic Commerce*, EC '11, (New York, NY, USA), pp. 199–208, ACM, 2011.

[35] Goldreich, O., Micali, S., and Wigderson, A., "How to play any mental game," in *Proceedings of the Nineteenth Annual ACM Symposium on Theory of Computing*, STOC '87, (New York, NY, USA), pp. 218–229, ACM, 1987.

[36] Goyal, N., Vempala, S., and Xiao, Y., "Fourier PCA," in *STOC*, 2014.

[37] Groves, T., *The Allocation of Resources Under Uncertainty: The Informational and Incentive Roles of Prices and Demands in a Team*. Technical report (University of California, Berkeley. Center for Research in Management Science), University of California, 1970.

[38] Groves, T. and Ledyard, J. O., "Optimal allocation of public goods: a solution to the "free rider" problem," *Econometrica*, vol. 45, pp. 783–809, May 1977.

[39] Guedon, O. and Rudelson, M., "L_p moments of random vectors via majorizing measures," *Advances in Mathematics*, vol. 208, pp. 798–823, 2007.

[40] Hardt, M. and Roth, A., "Beating randomized response on incoherent matrices," in *Proceedings of the 44th symposium on Theory of Computing*, STOC '12, (New York, NY, USA), pp. 1255–1268, ACM, 2012.

[41] Hardt, M. and Roth, A., "Beyond worst-case analysis in private singular vector computation," in *Proceedings of the 45th annual ACM symposium on Symposium on theory of computing*, pp. 331–340, ACM, 2013.

[42] Harks, T. and Peis, B., "Resource buying games," in *European Symposium on Algorithms (ESA)*, pp. 563–574, 2012.

[43] Hoefer, M., "Competitive cost sharing with economies of scale," *Algorithmica*, vol. 60, no. 4, pp. 743–765, 2011.

[44] Hyvarinen, A., Karhunen, J., and Oja, E., *Independent Component Analysis*. John Wiley and Sons, 2001.

[45] Immorlica, N., Markakis, E., and Piliouras, G., "Coalition formation and price of anarchy in Cournot oligopolies," in *Proceedings of the 6th international conference on Internet and network economics*, WINE'10, (Berlin, Heidelberg), pp. 270–281, Springer-Verlag, 2010.

[46] Jin, Y., Ok, J., Yi, Y., and Shin, J., "On the impact of global information on diffusion of innovations over social networks," in *IEEE International Workshop on Network Science for Communication Networks (NETSCICOM)*, pp. 3267–3272, 2013.

[47] Jung, T., Humphries, C., Lee, T.-W., Makeig, S., McKeown, M., Iragui, V., and Sejnowski, T., "Extended ICA removes artifacts from electroencephalographic recordings," *Advances in Neural Information Processing Systems*, vol. 10, 1998.

[48] Jutten, C. and Herault, J., "Blind separation of sources, part I: An adaptive algorithm based on neuromimetic architecture," *Signal Processing*, vol. 24, no. 1, pp. 1 – 10, 1991.

[49] Jutten, C. and Taleb, A., "Source separation: From dusk till dawn," in *Proc. Int. Workshop on Independent Component Analysis and Blind Signal Separation, ICA2000*, pp. 15–26, 2000.

[50] Kearns, M., Pai, M. M., Roth, A., and Ullman, J., "Mechanism design in large games: Incentives and privacy," *American Economic Review*, vol. 104, no. 5, pp. 431–35, 2014.

[51] Kempe, D., Kleinberg, J., and Tardos, É., "Influential nodes in a diffusion model for social networks," in *ICALP*, pp. 1127–1138, Springer Verlag, 2005.

[52] Kempe, D., Kleinberg, J., and Tardos, É., "Maximizing the spread of influence through a social network," in *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '03, (New York, NY, USA), pp. 137–146, ACM, 2003.

[53] Khan, A., Onoue, T., Hashiodani, K., Y. Fukumizu, Y., and Yamauchi, H., "Signal and noise separation in medical diagnostic system based on independent component analysis," *IEEE APCCAS*, pp. 812–815, 2010.

[54] Kiviluoto, K. and Oja, E., "Independent component analysis for parallel financial time series," *ICONIP*, vol. 2, pp. 895–898, 1998.

[55] Kleinberg, R., Piliouras, G., and Tardos, É., "Load balancing without regret in the bulletin board model," *Distributed Computing*, vol. 24, no. 1, pp. 21–29, 2011.

[56] Kleinberg, R., Piliouras, G., and Tardos, É., "Multiplicative updates outperform generic no-regret learning in congestion games: extended abstract," in *Proceedings of the 41st annual ACM symposium on Theory of computing*, STOC '09, (New York, NY, USA), pp. 533–542, ACM, 2009.

[57] Kleinberg, R. D., Ligett, K., Piliouras, G., and Tardos, É., "Beyond the Nash equilibrium barrier," in *Innovations in Computer Science (ICS)*, pp. 125–140, 2011.

[58] Koutsoupias, E. and Papadimitriou, C., "Worst-case equilibria," in *Symposium on Theoretical Aspects of Computer Science (STACS)*, pp. 404–413, Springer-Verlag, 1999.

[59] Krehbiel, S., "Markets for database privacy," *In submission*, 2015.

[60] Krehbiel, S., Mehta, R., and Vazirani, V., "Limits of privacy in correlated games," *Ongoing work*, 2015.

[61] Krehbiel, S., Peikert, C., and Xiao, Y., "Differentially private independent component analysis," *In submission*, 2015.

[62] Le, Q. V., Karpenko, A., Ngiam, J., and Ng, A., "Ica with reconstruction cost for efficient overcomplete feature learning," in *Advances in Neural Information Processing Systems*, pp. 1017–1025, 2011.

[63] Lee, H., Ekanadham, C., and Ng, A., "Sparse deep belief net model for visual area v2," in *Advances in neural information processing systems*, pp. 873–880, 2007.

[64] Ligett, K. and Piliouras, G., "Beating the best Nash without regret," *SIGecom Exch.*, vol. 10, pp. 23–26, Mar. 2011.

[65] Ligett, K. and Roth, A., "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, (Berlin, Heidelberg), pp. 378–391, Springer-Verlag, 2012.

[66] Machado, R. and Tekinay, S., "Diffusion-based approach to deploying wireless sensors to satisfy coverage, connectivity and reliability," in *Mobile and Ubiquitous Systems: Networking Services, 2007. MobiQuitous 2007. Fourth Annual International Conference on*, pp. 1–8, 2007.

[67] Malaroiu, S., Kiviluoto, K., and Oja, E., "Time series prediction with independent component analysis," *Int. Conf. on Advanced Investment Technology*, 2000.

[68] Mathias, R., "A bound for the matrix square root with application to eigenvector perturbation," *SIAM Journal of Matrix Analysis Applications*, vol. 18, no. 4, pp. 861–867, 1997.

[69] McKeown, M., Makeig, S., Brown, S., Jung, T.-P., Kindermann, S., Bell, A., Iragui, V., and Sejnowski, T., "Blind separation of functional magnetic resonance imaging (fMRI) data," *Human Brain Mapping*, vol. 6, pp. 368–372, 1998.

[70] McSherry, F. and Talwar, K., "Mechanism design via differential privacy," *Foundations of Computer Science, IEEE Annual Symposium on*, vol. 0, pp. 94–103, 2007.

[71] Monderer, D. and Shapley, L., "Potential games," *Games and Economic Behavior*, vol. 14, pp. 124–143, 1996.

[72] Nadav, U. and Piliouras, G., "No regret learning in oligopolies: Cournot vs. Bertrand," in *Proceedings of the third international conference on Algorithmic game theory*, SAGT'10, (Berlin, Heidelberg), pp. 300–311, Springer-Verlag, 2010.

[73] Narayanan, A. and Shmatikov, V., "Robust de-anonymization of large sparse datasets," in *Proceedings of the 2008 IEEE Symposium on Security and Privacy*, SP '08, (Washington, DC, USA), pp. 111–125, IEEE Computer Society, 2008.

[74] Nash, J., "Non-cooperative games," *Annals of Mathematics*, vol. 54, no. 2, pp. 286–295, 1951.

[75] Ngiam, J., Chen, Z., Chia, D., Koh, P. W., Le, Q. V., and Ng, A., "Tiled convolutional neural networks," in *Advances in Neural Information Processing Systems*, pp. 1279–1287, 2010.

[76] Nguyen, P. Q. and Regev, O., "Learning a parallelepiped: Cryptanalysis of GGH and NTRU signatures," *J. Cryptology*, vol. 22, no. 2, pp. 139–160, 2009.

[77] Nisan, N., Roughgarden, T., Tardos, É., and Vazirani, V. V., *Algorithmic Game Theory*. New York, NY, USA: Cambridge University Press, 2007.

[78] Nissim, K., Orlandi, C., and Smorodinsky, R., "Privacy-aware mechanism design," in *Proceedings of the 13th ACM Conference on Electronic Commerce*, EC '12, (New York, NY, USA), pp. 774–789, ACM, 2012.

[79] Nissim, K., Smorodinsky, R., and Tennenholtz, M., "Approximately optimal mechanism design via differential privacy," in *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, ITCS '12, (New York, NY, USA), pp. 203–213, ACM, 2012.

[80] Nissim, K., Vadhan, S., and Xiao, D., "Redrawing the boundaries on purchasing data from privacy-sensitive individuals," in *Proceedings of the 5th Conference on Innovations in Theoretical Computer Science*, ITCS '14, (New York, NY, USA), pp. 411–422, ACM, 2014.

[81] Pai, M. M. and Roth, A., "Privacy and mechanism design," *SIGecom Exchanges*, vol. 12, no. 1, pp. 8–29, 2013.

[82] Piliouras, G. and Shamma, J. S., "Optimization despite chaos: Convex relaxations to complex limit sets via Poincaré recurrence," in *Twenty-Fifth Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pp. 861–873, 2014.

[83] Piliouras, G., Valla, T., and Végh, L. A., "LP-based covering games with low price of anarchy," in *Proceedings of the 8th international conference on Internet and Network Economics*, WINE'12, (Berlin, Heidelberg), pp. 184–197, Springer-Verlag, 2012.

[84] Porrill, J., Stone, J., Berwick, J., J.Mayhew, and Coffey, P., "Analysis of optical imaging data using weak models and ICA," *Advances in Independent Component Analysis*, pp. 217–233, 2000.

[85] Raz, R. and Safra, S., "A sub-constant error-probability low-degree test, and a sub-constant error-probability PCP characterization of NP," in *ACM Symposium on Theory of Computing (STOC)*, pp. 475–484, 1997.

[86] Rudelson, M., "Random vectors in the isotropic position," *Journal of Functional Analysis*, vol. 164, no. 1, pp. 60 – 72, 1999.

[87] Sadagopan, N., Singh, M., and Krishnamachari, B., "Decentralized utility-based sensor network design," *Mob. Netw. Appl.*, vol. 11, pp. 341–350, June 2006.

[88] Samuelson, P. A., "The pure theory of public expenditure," *The Review of Economics and Statistics*, vol. 36, pp. 387–389, Nov. 1954.

[89] Schmid, S. and Wattenhofer, R., "Algorithmic models for sensor networks," in *In 14th International Workshop on Parallel and Distributed Real-Time Systems (WPDRTS)*, pp. 51–54, 2006.

[90] Shamma(ed), J., *Cooperative Control of Distributed Multiagent Systems*. Wiley, 2008.

[91] Sharma, Y. and Williamson., D. P., "Stackelberg thresholds in network routing games or the value of altruism," in *In ACM Conference on Electronic Commerce (EC)*, pp. 93–102, 2007.

[92] Smith, A., "Efficient, differentially private point estimators," *arXiv preprint arXiv:0809.4794*, 2008.

[93] Srivastava, N. and Vershynin, R., "Covariance estimates for distributions with $2 + \epsilon$ moments," *submitted*, 2011.

[94] Vempala, S. and Xiao, Y., "Structure from local optima: Learning subspace juntas via higher order PCA," *(submitted)*, 2012.

[95] Vershynin, R., "How close is the sample covariance to the actual covariance matrix?," *Journal of Theoretical Probability*, vol. to appear, 2010.

[96] Vershynin, R., "Introduction to the non-asymptotic analsis of random matrices," in *Compressed Sensing: Theory and Applications* (Eldar, Y. and Kutyniok, G., eds.), pp. 210–268, Cambridge University Press, 2012.

[97] Vickrey, W., "Counterspeculation, auctions, and competitive sealed tenders," *Journal of Finance*, vol. 16, pp. 8–37, 03 1961.

[98] Vigário, R., "Extraction of ocular artifacts from EEG using independent component analysis," *Electroenceph. Clin. Neurophysiol.*, vol. 103, pp. 395–404, 1997.

[99] Vigário, R., Jousmäki, V., Hämäläinen, M., Hari, R., and Oja, E., "Independent component analysis for identification of artifacts in magnetoencephalographic recordings," *Advances in Neural Information Processing Systems*, vol. 10, pp. 229–235, 1998.

[100] Vigário, R., Särel, J., Jousmäki, V., Hämäläinen, M., and Oja, E., "Independent component approach to the analysis of EEG and MEG recordings," *IEE Trans. Biomedical Engineering*, vol. 47, pp. 589–593, 2000.

[101] Xiao, D., "Is privacy compatible with truthfulness?," in *In Proc. ITCS 2013*, pp. 67–86, 2013.

[102] Yan, H., Chen, H., Xia, Y., Lai, Y., and Zhou, D., "Independent component analysis for human epileptic spikes extraction," *Neural Interface and Control*, pp. 93–95, 2005.

[103] Yao, A. C., "Protocols for secure computations," in *Proceedings of the 23rd Annual Symposium on Foundations of Computer Science*, SFCS '82, (Washington, DC, USA), pp. 160–164, IEEE Computer Society, 1982.

[104] Zalyubovskiy, V., Erzin, A., Astrakov, S., and Choo, H., "Energy-efficient area coverage by sensors with adjustable ranges," *Sensors*, vol. 9, no. 4, pp. 2446–2460, 2009.

[105] Zennaro, M., Bagula, A., Gascon, D., and Noveleta, A. B., "Long distance wireless sensor networks: simulation vs reality," in *Proceedings of the 4th ACM Workshop on Networked Systems for Developing Regions*, NSDR '10, (New York, NY, USA), pp. 12:1–12:2, ACM, 2010.

[106] Zhou, S., Ligett, K., and Wasserman, L., "Differential privacy with compression," *arXiv:0901.1365*, 2009.