New Data Markets Deriving from the Internet of Things: A Societal Perspective on the Design of New Service Models

Roy Dong

Electrical Engineering and Computer Sciences
University of California at Berkeley

Technical Report No. UCB/EECS-2017-52
http://www2.eecs.berkeley.edu/Pubs/TechRpts/2017/EECS-2017-52.html

May 11, 2017

Acknowledgement

(By order of appearance:)

To my parents, rct, jmb, mdc, and krdc.

New Data Markets Deriving from the Internet of Things:

A Societal Perspective on the Design of New Service Models

by

Roy Dong

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Engineering – Electrical Engineering and Computer Sciences

and the Designated Emphasis

in

Communication, Computation, and Statistics

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor S. Shankar Sastry, Chair
Professor Alexandre M. Bayen
Professor Joan Walker

Spring 2017

The dissertation of Roy Dong, titled New Data Markets Deriving from the Internet of Things: A Societal Perspective on the Design of New Service Models, is approved:

Chair _____ Date _____

_____ Date _____

_____ Date _____

University of California, Berkeley

New Data Markets Deriving from the Internet of Things:
A Societal Perspective on the Design of New Service Models

Abstract

New Data Markets Deriving from the Internet of Things:
A Societal Perspective on the Design of New Service Models

by

Roy Dong

Doctor of Philosophy in Engineering – Electrical Engineering and Computer Sciences
and the Designated Emphasis in
Communication, Computation, and Statistics

University of California, Berkeley

Professor S. Shankar Sastry, Chair

The Internet of Things (IoT) is a term that represents a huge technological trend that is taking place: almost every device is being imbued with the intelligence of a microprocessor and an Internet connection. We view IoT as a phenomena in which new service models will emerge. Central to these service models will be the provided data and the conversations surrounding it. In this document, we outline our research in the formulation and analysis of these new service models. This work is focused on the role of data and the value of information in IoT.

First, we present our work on disaggregation algorithms, which take aggregate measurements at a higher level of abstraction to infer component measurements at a lower level of abstraction. This is inspired by many IoT settings where aggregation frequently happens along the data pipeline due to energy and bandwidth constraints, as well as limitations on sensor placement. Additionally, we present our work on blind system identification, which provide a method to identify the dynamics of observed systems when both the internal states and the inputs are not observed, as is typical in many IoT settings. For example, smart meters observe the aggregate energy consumption of a building, but do not directly observe the individual device's energy consumptions, the transient energy consumption dynamics of devices, or how devices are being utilized inside a building. Disaggregation and blind system identification allow IoT system managers to infer models of these components of IoT systems even when sensor measurements do not provide them.

Second, we present our work in quantifying, analyzing, and incentivizing privacy in IoT systems. Motivated in part by the efficacy of disaggregation algorithms, we consider the privacy facet of IoT technologies. We discuss the literature on quantifying privacy, and also discuss a new metric inspired by classical information theoretic and statistical frameworks, which we call inferential privacy. We translate some of the existing information theoretic and statistical literature into privacy guarantees in this new framework. Then, we discuss

different design paradigms for privacy, which range from passive privacy analysis to optimal privacy-by-design. These taxonomies are complemented by detailed examples in transportation networks, smart grid control, and air quality regulation.

Third, we discuss the value of information, data markets, and new service models in IoT. We consider a model of a data buyer who deal with strategic data sources: the data buyer must balance its objective of having a low-error statistical estimator with the cost of issuing incentives to effort-averse data sources. We extend these models to competitive settings, where multiple data buyers are incentivizing the same data sources, and analyze the existence of equilibria in such settings, as well as properties of such equilibria. We also consider how IoT systems allow for a new mode of actuation: causal imputation. In many smart infrastructure applications, we no longer have control commands that directly affect the dynamics. For example, the turn signal on a car does not have any direct effect on the dynamics of the car. As another example, many demand response programs are moving towards preferential pricing, rebates, and other incentive schemes to curtail peak energy consumption, rather than a direct load control. Both of these situations can be modeled with our optimal causal imputation framework. These control actions are structurally different from previously studied modes of actuation and we formulate the problem of optimal causal imputation, and provide algorithms for calculating these solutions under certain assumptions.

In this document, we provide a technical analysis of the IoT systems and the statistical properties of their data, a behavioral analysis of the human actors who respond to IoT systems and participate by the revelation of their data (or lack thereof), and a game theoretic analysis of the data analytics companies who drive competitive data markets with market power. This work is an effort to build a larger picture of the IoT as an emerging data market, and motivates much of the theoretical frameworks we have developed and plan to develop in future work.

(By order of appearance:)
To my parents, rct, jmb, mdc, and krdc.

Contents

Acknowledgments

I'd like to take one of the earliest pages in my dissertation to thank all the people who have, directly or indirectly, helped shape the academic I've grown into throughout the course of my Ph.D. career.

I'd like to thank my advisors throughout my higher education. I wish to thank my undergraduate advisor, Xiaobo Tan, for inaugurating me into both the worlds of control theory and research. Without the opportunities you gave me, I am not sure I would have been able to attend graduate school, and would have a drastically different life right now. I wish to thank my Ph.D. advisor, S. Shankar Sastry, for providing the resources and freedom to see the cutting edge research and mathematical frameworks in many different fields, as well as pushing me to be a researcher who searches for theoretically sound formulations of real-world problems. This influence has caused me to look at mathematics as a field inspired by practicality, where the properties of physical motion and time itself led to calculus, and where the properties of mass and size led to measure theory, and so on, and so on.

I'd like to thank the other professors I've had the pleasure of interacting in graduate school, as well. You have served as my instructors, my mentors, and my role models. In particular, I'd like to raise a glass to Claire Tomlin, Murat Arcak, Alexandre Bayen, and Ruzena Bajcsy for their role in teaching me how to formulate important problems and analyze them rigorously.

I'd like to thank the collaborators and friends I've interacted with in graduate school. I'd like to thank Allen Yang and Henrik Ohlsson for mentoring a young graduate student in his first research project, with patience for all his early mistakes in matrix multiplication. I'd like to thank the older graduate students from my early academic career for telling me what's up: cheers Sam Burden, Sam Coogan, and Humberto Gonzalez. I'd like to thank Alvaro Cardenas and Saurabh Amin for initiating me into privacy research. I'd like to thank members of my cohort for all the highly educational, always productive, and never procrastinating conversations: it was always technically research, probably, Austin Buchan and Aaron Bestick. I'd like to thank Dan Calderone and Walid Krichene for taking my lines and curving them upward in our convex reading group, which has served as my mental paradigm for everything a group of friends learning math should be. I must thank Maximilian Balandat for forcing me to be pedantic at all the right times in my math courses. I'd also like to thank Jupiter Zhu for being there in my most uncomprehending moments of mathematics: you have had a huge influence in building my intuitions with your mantra of 'proceed fearlessly'. I hear your voice every time I enter unknown territory. I'd like to thank the younger grad students I have worked with: Ruoxi Jia, Tyler Westenbroek, Dexter Scobee, Eric Mazumdar, and Oladapo Afolabi. You are endlessly creative and have impressed me many times. Also, just general props to my friends: Dorsa Sadigh, Kamil Nar, Jaime Fernandez-Fisac, Vincenc Rubios Royo, and so on, and so on. I'd like to extend my greatest thanks to Lillian Ratliff, who has been a constant source of ideas, support, and friendship throughout my graduate school career. There are countless moments I am glad I

had someone there: from being called a ‘nerd!’ at a formal dinner by a board member of a local utility company to the surreal grant review for no one during the government shutdown.

No recent thesis from my research group would be complete without acknowledging the amazing work of the people who make our group run: Carolyn Winter, Aimee Tabor, Larry Rohrbough, and Jessica Gamble. You make our office a smoothly sailing ship in which we can safely and calmly conduct research.

Lastly, Katherine Driggs-Campbell, you make me who I am.

Chapter 1

Introduction

The Internet of Things (IoT) is a term that represents a huge technological trend that is taking place: almost every device is being imbued with the intelligence of a microprocessor and an Internet connection. The interconnection in IoT promises an infrastructure that can drastically change how consumers live their day-to-day lives, with huge gains in efficiency, value, and possibility due to the shared knowledge and autonomy allowed. In profound ways, as the technology develops, the modalities of existence people experience will grow and shift.

However, the scale and scope of IoT raises new problems for engineers to consider. These problems are significantly different from ones previously explored in the design of comparatively isolated systems, and require a new theoretical underpinning to analyze IoT with models that capture all salient facets of these new technologies. This document contains a handful of theoretical frameworks, and their applications, as a first step into this new research frontier.

First, we consider the problem of large amounts of data. For example, in the energy sector, advanced metering infrastructures collect energy consumption data for a large number of consumers at relatively high frequencies. This glut of data isn't useful for most operational purposes, such as phase-alignment, and is often aggregated for control purposes. Additionally, these smart meter readings are usually at a household level, and themselves represent an aggregate of several devices inside an energy consumer's home.

Furthermore, if these devices are thought of as dynamical systems, these smart meter readings only capture the 'output' of the system: both the internal state dynamics and the driving inputs remain unobserved. This generally is a trend with IoT sensors: they capture a process but not the 'inputs' driving it. As another example, smart phone data tracks the location of users, which can be thought of as a random process whose distribution depends on the user's itinerary, which is often not a direct 'input' into Google Maps.

In Chapter 2, we address these two problems.

We define the disaggregation problem as the recovery of component signals y_i from observations of an aggregate $\sum_i y_i$. We focus on an application in energy disaggregation, and outline assumptions motivated by this context. Under these assumptions, we can phrase the disaggregation problem as a hybrid optimal control problem. Using adaptive filtering

techniques, we provide an algorithm that can tractably find the optimal solutions to the disaggregation problem.

Additionally, we provide a blind system identification method to simultaneously identify the inputs and dynamics of systems when only output observations are available. This, in and of itself, is an ill-posed problem, so we find prior knowledge that can be supported by IoT sensors. For example, if we are tracking occupancy in a building, IoT sensors can detect when a door is opened. This serves as regularizing knowledge for blind system identification: we know the discrete time points when occupancy of a sector can change.

Chapter 2 outlines some new estimation problems due to the scale and nature of IoT sensor measurements.

A reader who is sensitive about the abundance of collected and transmitted technology may be unnerved by some of the results in Chapter 2, and rightfully so. In Chapter 3, we outline some of the work in preserving privacy in new IoT systems. Fundamentally, we suppose IoT sensors are collecting data for some estimation or control tasks that are not directly in line with privacy violations of the user. In other words, privacy risks are a byproduct of a system designed to do something else.

In Chapter 3, we outline methods to quantify the privacy of a system. We review the literature on quantitative measures for privacy, and discuss some of our contributions, predominantly in *inferential privacy*. We discuss different design paradigms for the incorporation of privacy into IoT systems as a design level: from a passive privacy audit of a system, to an optimal design perspective that creates sensors that provide measurements that optimally benefit the control objective while minimizing privacy threats to users. Finally, we implement these privacy design concepts with examples in ground transportation, smart grid control, and air quality regulation.

As IoT technologies permeate our socioeconomic spheres more and more, users will become more privacy-aware and conscious of their data flows. A common saying heard these days in tech circles is: 'If you are not paying for a service, you the product, not the customer.' This applies to services such as Gmail or Facebook. Individual users are increasingly aware that their data is, in many ways, the currency of new technologies. Web services such as PrivacyFix allow users to calculate how much revenue is generated by Google from their data alone, and change their privacy settings.

In Chapter 4, we model users who are strategic about their data sharing, and how the actions of data buyers and advertisers must change when users become more strategic. The existing literature has some methods that can handle strategic data sources, and we extend these results to consider how these methods work when multiple data buyers are competing to create better estimators.

Additionally, IoT allows new means and modes of providing feedback into our systems. This can come in the form of economic rebates for products that change the energy consumption profiles inside a house, or warning lights inside a vehicle outfitted with intelligent sensors, or cell-phone messages suggesting faster routes. None of these are direct control actions in a control-theoretic sense, but can be thought of as actions that have an effect on the distribution of system behaviors after the fact.

Furthermore, these are often imputations on endogenous variables inside a system. For example, the price of an eco-friendly fridge is determined by supply and demand in the market, but a rebate can, in a sense, manually force the market price to something else, at some cost to the entity issuing the rebates. We argue that this phenomena can be modeled as *causal imputations*, and discuss a framework for finding the optimal imputations under some system performance objective and a cost of imputation. This is also discussed in Chapter 4.

Finally, in Chapter 5, we provide some closing remarks on the results presented here, and our vision of future work in IoT-motivated research.

1.1 Notation and conventions

Finally, to conclude the introduction, we provide an overview of the notation used throughout the rest of this document.

Let $\mathbb{N}$ refer to the set of natural numbers, and $\mathbb{R}$ refer to the set of real numbers. For any $D \in \mathbb{N}$, let $[D] = \{1, 2, \ldots, D\}$.

Whenever random variables are discussed, they will be with respect to a general probability space $(\Omega, \mathcal{A}, P)$, where $\mathcal{A}$ is a σ-algebra on Ω and $P : \mathcal{A} \to [0, 1]$ is a probability measure. For any measurable space $(S, \mathcal{S})$, if $X : \Omega \to S$ is a random variable and P_X is a probability measure on $(S, \mathcal{S})$, then $X \sim P_X$ is the assertion that $P_X(E) = P(\{\omega : X(\omega) \in E\})$ for all $E \in \mathcal{S}$. Similarly, we define the following notation for indicator functions:

$$\mathbb{I}\{x \in A\} = \begin{cases} 1 & \text{if } x \in A \\ 0 & \text{otherwise} \end{cases}$$

For matrices, $\text{rank}(X)$ denotes the rank of X and $\text{Tr}(X)$ denotes the trace. A^T denotes the transpose of A.

Any other notation will be defined in the appropriate context. Additionally, due to the limitations of the cardinality of the English and Greek alphabets, all variable definitions will be scoped to the section or subsection in which they are defined[1] unless otherwise noted.

[1]Sections in this thesis have headings of the form 0.0, and subsections have headings of the form 0.0.0.

Chapter 2

Disaggregation and Inference in the Internet of Things

The Internet of Things (IoT) facilitates the communication and control of physical objects that previously operated in isolation. As discussed in Chapter 1, this requires the measurement, collection, transfer, and storage of unprecedented amount of data from physical systems.

Furthermore, this IoT revolution is taking place across all scales, ranging from the behavior of a single household or an ad-hoc network of a handful of personal devices, to the behavior of the entire energy grid or ground transportation infrastructure. The algorithms used by IoT devices work at different levels of abstraction, based which facets of these new data streams are salient to the operation of the respective devices. For example, your toaster and refrigerator may share their individual usage patterns, whereas smart meters operating in an advanced metering infrastructure (AMI) will communicate to each other using only the total energy consumption of a household. I Note as we increase the level of abstraction, our ontologies necessarily aggregate individual entities into collections of entities. In this Chapter, we consider algorithms designed to take data at one level of abstraction, e.g. the energy consumption profile of a building, and estimate data at a lower level of abstraction, e.g. the usage patterns of individual devices within the building. We will refer to such methods as *disaggregation* algorithms, and will give a more formal exposition of disaggregation in Section 2.1.

Throughout this Chapter we will focus on the problem of energy disaggregation for concreteness of the concepts presented. However, we note that the concept of disaggregation shows up in many other situations. For example, in transportation networks, researchers have been working on methods to estimate flows along different routes, based on the aggregate flows along roadways and side information [Wu+15].

There are many ways to interpret the results presented in this chapter, with some interpretations sunnier than others.

One illumination notes the great impact that disaggregation algorithms could have on the efficiency and utility of these IoT systems [Arm+13]. In this light, we can think of ag-

gregation as a method of compression, and good disaggregation algorithms allow for much more efficient methods of sensing and transmitting data. In the energy sector, the cost of instrumenting all energy-consuming devices with sensors for AMI communication prohibits large-scale deployment, and the transmission of frequent samples from several devices is also very expensive. In this sense, the fact that a smart meter records only the aggregate energy consumption of a household is a form of compressed sensing [Don06]. Then, a much lower dimensional data stream is transmitted across the network, reducing bandwidth consumption.

Once this aggregate data is transmitted to the utility company, it can use the disaggregated data to improve estimation and control of the the energy grid. Detailed information about consumption patterns allow the construction of better predictive models for forecasting future loads, which allows for load schedules that incorporate more clean and inefficient energy sources. At a household level, this disaggregated data is useful for diagnosing inefficient operating points, such as heating, ventilation, and air conditioning (HVAC) settings, as well as malfunctioning or faulty devices. Even more forward looking, the utility company can use this information to design more effective programs to incentivize better consumption profiles. In fact, simply providing users with their disaggregated energy consumption data can induce more energy efficient behaviors [EM+10].

Another perspective is much more cautious about the promise of disaggregation techniques. The device usage patterns inside our home are intertwined with our lifestyles and behaviors, so the possibility of disaggregation is seen as a threat to the privacy of users in modern IoT systems. Thus, Theorem 1 can be seen as conditions in which an adversary can take smart meter readings and infer the behaviors of a user inside the home. We will develop this perspective further in Chapter 3, where we will prove fundamental limits on the disaggregation problem, and generalize this to create a new privacy metric.

The rest of this Chapter is organized as follows. First, in Section 2.1, we present our work in disaggregation. We frame the general disaggregation problem, discuss related work, and provide a tractable algorithm that gives theoretical guarantees of optimal performance under mild conditions. In Section 2.2, we present our work in blind system identification. Blind system identification is widely useful in cyber-physical systems, where physical processes often follow unknown dynamics driven by unknown inputs, but there may be strong structural priors and large amounts of data that can be leveraged. We formulate the blind system identification problem, provide a solution via lifting, discuss theoretical results in the noise-free and noisy case, and show the performance on a small example.

2.1 Disaggregation

In this Section, we outline a formalization of the basic components that are present in any disaggregation problem.

We have a finite set of individual entities. We will say there are $D \in \mathbb{N}$ entities indexed by $[D]$. For each entity $i \in [D]$, there is an associated abstract input space $\mathcal{U}_i$ and output space

$\mathcal{Y}_i$. The input $u = (u_1, u_2, \ldots, u_D) \in \prod_{i\in[D]} \mathcal{U}_i$ can be interpreted either in a frequentist framework, where u is some unknown, fixed parameter, or in a Bayesian framework, where $u \sim P_u$ for some probability measure P_u. In either framework, the distribution of the output y_i is determined by u_i, i.e. $y_i|u_i \sim P_{y_i|u_i}(\cdot|u_i)$ for some probability kernel[1] $P_{y_i|u_i}$. Additionally, y_i is conditionally independent of u_j for $j \neq i$ given u_i.

These outputs are aggregated by some function h into $y = h(y_1, y_2, \ldots, y_D)$, where y is the observable. The goal of disaggregation is to estimate $\{u_i\}_{i\in[D]}$ from y and the available side information. Depending on the application, criteria for the 'best' estimate may differ.

We will instantiate this framework on the context of energy disaggregation below.

2.1.1 Related work

In this Subsection, we discuss literature related to the problem of disaggregation. Much existing work falls into the general category of energy disaggregation; however, almost all approaches are not agnostic to the application domain and cannot be generally applied to all disaggregation problems.

The problem of energy disaggregation, and the existing hardware for disaggregation, has been studied extensively in the literature (see [Ber+09; Ber+10], for example). The goal of the current disaggregation literature is to present methods for improving energy monitoring at the consumer level without having to place sensors at device level, but rather use existing sensors at the whole building level.

Disaggregation, in essence, is a single-channel source separation problem. The problem of recovering the components of an aggregate signal is an inverse problem and as such is, in general, ill-posed. Most disaggregation algorithms are batch algorithms and produce an estimate of the disaggregated signals given a batch of aggregate recordings. There have been a number of survey papers summarizing the existing methods (e.g. see [ZR11], [KJ11]). In an effort to be as self-contained as possible, we try to provide a broad overview of the existing methods and then explain how the disaggregation method presented in this paper differs from existing solutions.

The literature can be divided into two main approaches, namely, supervised and unsupervised. Supervised disaggregation methods require a disaggregated data set for training. This data set could be obtained by, for example, monitoring typical appliances using plug sensors. Supervised methods assume that the variations between signatures for the same type of appliances is less than that between signatures of different types of appliances. Hence, the disaggregated data set does not need to be from the building that the supervised algorithm is designed for. However, the disaggregated data set must be collected prior to deployment, and come from appliances of a similar type to those in the target building. Supervised methods are typically discriminative.

[1] Recall that a probability kernel is a function such that $P_{y_i|u_i}(\cdot|v)$ is a probability measure for each $v \in \mathcal{U}_i$ and $P_{y_i|u_i}(E|\cdot)$ is a measurable function for each measurable set $E \subset \mathcal{Y}_i$.

Unsupervised methods, on the other hand, do not require a disaggregated data set to be collected. They do, however, require hand tuning of parameters, which can make it hard for the methods to be generalized in practice. It should be said that also supervised methods have tuning parameters, but these can often be tuned using the training data.

The existing supervised methods include sparse coding [KN10], change detection and clustering based approaches [DK99; Rah+12] and pattern recognition [FZ99]. The sparse coding approach tries to reconstruct the aggregate signal by selecting as few signatures as possible from a library of typical signatures. Similarly, in our proposed framework we construct a library of dynamical models and reconstruct the aggregate signal by using as few as possible of these models.

The existing unsupervised methods include factorial hidden Markov models (HMMs), difference hidden Markov models and variants [Kim+11; KJ12; JW12; Par+12; Pat12] and temporal motif mining [Sha+12]. Most unsupervised methods model the on/off sequences of appliances using some variation of HMMs. These methods do not directly make use of the signature of a device and assume that the power consumption is piecewise constant.

All methods we are aware of lack the use of the dynamics of the devices. While the existing supervised methods often do use device signatures, these methods are discriminative and an ideal method would be able to generate a power consumption signal from a given consumer usage profile. Both HMMs and linear dynamical models are generative as opposed to discriminative, making them more advantageous for modeling complex system behavior. In the unsupervised domain, HMMs are used; however, they are not estimated using data and they do not model the signature of a device.

In our work, we formulate hypotheses on the on/off state of the devices over the time horizon for which we have data. The on/off state corresponds to whether the input is activated or not. Using filter banks and the dynamical models we have for device behavior, we evaluate which is the most likely hypothesis on the inputs. We provide an algorithm for this process. Under mild assumptions on the noise characteristics we are able to provide guarantees for when the our algorithm results in an optimal solution. The filter bank framework is similar to HMM frameworks in the sense that both methods essentially formulate hypotheses on which devices are on at each time instant. However, in contrast to HMMs, in the filter bank framework we incorporate the use of dynamical models to capture the transients of the devices, which helps identify them.

2.1.2 Energy disaggregation

Here, we present our algorithm and results for energy disaggregation. The following text is an extension of the work presented in [Don+13a; Don+13b].

First, let us consider the problem of energy disaggregation in the framework of general disaggregation.

Problem Statement 1. (Energy disaggregation)

There are $D \in \mathbb{N}$ devices, and we work with a time horizon of $T \in \mathbb{N}$ discrete time steps. For $i \in [D]$, $u_i[t] \in \mathbb{R}$ is the usage pattern of device i at time t, for $t \in \{0, 1, \ldots, T\}$, and, similarly, $y_i[t] \in \mathbb{R}$ is the energy consumption of device i at time t.

We can represent u_i as a vector, i.e. $u_i \in \mathbb{R}^{T+1}$, or we can represent u_i as a function, i.e. $u_i : \{0, 1, \ldots, T\} \to \mathbb{R}$. Since the two are equivalent, we will often vacillate between the two representations for convenience. Similarly, we will treat y_i as both a vector and a function, and u as a function mapping from $\{0, 1, \ldots, T\} \to \mathbb{R}^D$.

The energy consumption signals follow the distribution $y_i|u_i \sim P_{y_i|u_i}(\cdot|u_i)$. Additionally, the energy consumption of device i, y_i, is conditionally independent of the usage patterns of other devices u_j for $j \neq i$, given u_i. The aggregate energy consumption is given by $y = \sum_{i\in[D]} y_i$, which is known.

Formally, we can define the frequentist energy disaggregation problem as the tuple[2]*:*

$$(\{P_{y_i|u_i}\}_{i\in[D]}, y) \tag{2.1.1}$$

The maximum likelihood (ML) *estimate given by:*

$$\widehat{u_{\text{ML}}} = \arg\max_u \int \cdots \int \left(\prod_{i=1}^{D-1} P_{y_i|u_i}(y_i|u_i) \right) P_{y_D|u_D} \left(y - \sum_{i=1}^{D-1} y_i \middle| u_D \right) \, dy_1 \ldots dy_{D-1} \tag{2.1.2}$$

Additionally, we can place this problem in a Bayesian framework. First, assume that $u \sim P_u$. Then, we can define the Bayesian energy disaggregation problem as the tuple:

$$(P_u, \{P_{y_i|u_i}\}_{i\in[D]}, y) \tag{2.1.3}$$

The maximum a posteriori (MAP) *estimate of u is given by:*

$$\widehat{u_{\text{MAP}}} = \arg\max_u P_u(u) \int \cdots \int \left(\prod_{i=1}^{D-1} P_{y_i|u_i}(y_i|u_i) \right) P_{y_D|u_D} \left(y - \sum_{i=1}^{D-1} y_i \middle| u_D \right) \, dy_1 \ldots dy_{D-1} \tag{2.1.4}$$

Throughout the following text, we will consider the Bayesian case for a unified presentation, but results will easily extend to the frequentist case. Additionally, we'll suppose there is a unique element in the arg max, although this assumption is for simplicity of notation and not necessary for the development of the subsequent text[3]. In the development of the algorithm pseudocode, we will again tread carefully regarding the potential of minimizing sets with more than one element.

In this problem formulation, we are given P_u and $P_{y_i|u_i}$. In practice, these will have to be learned from a training set of individual device's energy consumption data. We will discuss how to learn these models in Section 2.2.

[2]Note that D is implicitly contained in the size of the set in the first element of the tuple, and T is implicitly contained in the domain of the probability measures.

[3]For full rigor, we can simply replace any statement of the form '$x = \arg\max_{x\in C} f(x)$' with 'pick any $x \in \arg\max_{x\in C} f(x)$'.

Switching times

An observation that helps us solve the problem of energy disaggregation is the following: when data is collected at high frequencies, the inputs to devices are often piecewise constant. For example, the Reference Energy Disaggregation Dataset (REDD) [KJ11] contains data sampled at rates upwards of 1/3 Hz. In contrast, heating, ventilation, and cooling (HVAC) systems switch states at much slower rates, and energy consumers change lighting settings at slower rates as well.

With this observation in mind, we can define the switching times of a given input.

Definition 1. (Switching times)
For some fixed $v : \{0, 1, \ldots, T\} \to \mathbb{R}^D$, we can define the switching times *of v, denoted $T_{switch}(v) \subset \{0, 1, \ldots, T\}$, as the unique set such that:*

- *$v[t-1] \neq v[t]$ for all $t \in T_{switch}(v)$.*
- *$v[t-1] = v[t]$ for all $t \notin T_{switch}(v)$.*

Here, we adopt the convention where $v[-1] = 0$.

We will often use the notation $T_{switch}(v) = \{t_1, t_2, \ldots, t_N\}$ with the understanding that N depends on v and $t_1 < t_2 < \cdots < t_N$. Switching times will also be referred to as a segmentation. *Each interval $\{t_n, t_n + 1, \ldots, t_{n+1} - 1\}$ for $n \in \{0, 1, \ldots, N\}$ is defined as a* segment, *with the convention $t_0 = 0$ and $t_{N+1} = T+1$. Additionally, we adopt the convention that $N = 0$ when $T_{switch} = \emptyset$.*

We can similarly define the switching times of the random variable $u \sim P_u$, where $T_{switch}(u)$ is now a random set-valued element.

The energy disaggregation problem can be thought of as the process of finding the switching times of $\widehat{u_{\text{MAP}}}$, and subsequently estimating the actual value of $\widehat{u_{\text{MAP}}}$. If we consider every possible switching time, we would solve the energy disaggregation problem exactly. This is formally stated in Proposition 1.

First, we will introduce the following notation for $z \in \mathbb{R}^T$ and $v : \{0, 1, \ldots, T\} \to \mathbb{R}^D$.

$$P_{y|u}(z|v) = \int \cdots \int \left(\prod_{i=1}^{D-1} P_{y_i|u_i}(z_i|v_i) \right) P_{y_D|u_D} \left(y - \sum_{i=1}^{D-1} z_i \middle| v_D \right) \, dz_1 \ldots dz_{D-1} \tag{2.1.5}$$

Note that we can write:

$$\widehat{u_{\text{MAP}}} = \arg\max_{\widehat{u}} \left(P_u(\widehat{u}) P_{y|u}(y|\widehat{u}) \right) \tag{2.1.6}$$

Definition 2. *For any set $T_{switch} \subset \{0, 1, \ldots, T\}$, we define $p(T_{switch}|y)$ as follows:*

$$p(T_{switch}|y) = \max_{\widehat{u}} \left(P_u(\widehat{u}) P_{y|u}(y|\widehat{u}) \right) \tag{2.1.7}$$

Here, the maximization is taken across all $\widehat{u}$ such that $T_{switch}(\widehat{u}) = T_{switch}$.

Note that $p(T_{switch}|y)$ is the posterior probability associated with the estimate $\widehat{u}$ that maximizes the posterior probability subject to the switching times constraint, $T_{switch}(\widehat{u}) = T_{switch}$. A slight nuance to note is that this function p does *not* yield the posterior probability of T_{switch} given y, which would require marginalizing across $\widehat{u}$ rather than maximizing.

Proposition 1. *$T_{switch}(\widehat{u_{\text{MAP}}})$ satisfies the following condition.*

$$T_{switch}(\widehat{u_{\text{MAP}}}) = \arg\max_{T_{switch}} p(T_{switch}|y) \tag{2.1.8}$$

Additionally, if $T_{switch}(\widehat{u_{\text{MAP}}})$ is known, then:

$$\widehat{u_{\text{MAP}}} = \arg\max_{\widehat{u}} \left(P_u(\widehat{u})P_{y|u}(y|\widehat{u})\right) \tag{2.1.9}$$

Here, the maximization is taken across all $\widehat{u}$ such that $T_{switch}(\widehat{u}) = T_{switch}(\widehat{u_{\text{MAP}}})$.

Similarly:

$$P_u(\widehat{u_{\text{MAP}}})P_{y|u}(y|\widehat{u_{\text{MAP}}}) = \max_{T_{switch}} p(T_{switch}|y) \tag{2.1.10}$$

Proposition 1 implies that the problem of energy disaggregation can be broken up into two parts. First, we identify the set of switching times $T^*_{switch} = \arg\max_{T_{switch}} p(T_{switch}|y)$. Then, we calculate the MAP estimate $\widehat{u_{\text{MAP}}}$ by maximizing the posterior probability $P_u(\widehat{u})P_{y|u}(y|\widehat{u})$ across the set of $\widehat{u}$ such that $T_{switch}(\widehat{u}) = T^*_{switch}$. Note that this problem is still combinatorial, as there are 2^{T+1} different possible sets of switching times.

Next, we will argue that the optimal $\widehat{u}$ for a fixed segmentation is a light calculation, given some assumptions motivated by the energy disaggregation application. Then, we will present an algorithm which allows us to selectively consider possible switching times, while still ensuring the recovery of $\widehat{u_{\text{MAP}}}$.

Energy disaggregation via filter banks (EDFB)

In this section, we will first introduce conditions on which the estimation of $\widehat{u}$ is computationally inexpensive for a fixed segmentation. Essentially, the next assumption will ensure that the estimation problem in each segment decouples.

We quickly introduce the notation:

$$y_{s_1:s_2} = (y[s_1], y[s_1+1], \ldots, y[s_2]) \tag{2.1.11}$$

Assumption 1. (Conditional independence of segments)

Conditioned on $T_{switch}(u) = \{t_1, t_2, \ldots, t_N\}$, the following random variables are independent:

$$\{u[t_1], u[t_2], \ldots, u[t_N]\} \tag{2.1.12}$$

In other words, we can express P_u in the following way:

$$P_u(v) = \prod_{n=0}^{N} P_u(v[t_n]; T_{switch}(v), n) \tag{2.1.13}$$

Similarly, conditioned on $T_{switch}(u) = \{t_1, t_2, \ldots, t_N\}$*, the following are independent:*

$$\{y_{t_0:t_1-1}, y_{t_1:t_2-1}, \ldots, y_{t_N:t_{N+1}-1}\} \tag{2.1.14}$$

That is, we can express $P_{y|u}$ *as follows:*

$$P_{y|u}(z|v) = \prod_{n=0}^{N} P_{y|u}(z_{t_n:t_{n+1}-1}|v[t_n]; T_{switch}(v), n) \tag{2.1.15}$$

Finally, we assume these systems are causal in the following sense.

For any $t \in \{0, 1, \ldots, T\}$*,* $T_{switch} = \{t_1, t_2, \ldots, t_N\} \subset \{0, 1, \ldots, t-1\}$*, and* $y \in \mathbb{R}^t$*, let* $t_{N+1} = t$ *and let* $T'_{switch} = T_{switch} \cup t_{N+1}$*. Then:*

$$\prod_{n=0}^{N} P_u(\widehat{u}[t_n]; T_{switch}, n) \int P_{y|u}((y_{t_n:t_{n+1}-1}, z)|\widehat{u}[t_n]; T_{switch}, n)\ dz = \tag{2.1.16}$$

$$\prod_{n=0}^{N} P_u(\widehat{u}[t_n]; T'_{switch}, n) P_{y|u}(y_{t_n:t_{n+1}-1}|\widehat{u}[t_n]; T'_{switch}, n) \tag{2.1.17}$$

This condition states that the probability distribution of $u_{0:t}$ *and* $y_{0:t}$ *are independent of the input for* $u_{t+1:T}$*.*

Our notation here has $P_u(v; T_{switch}, n)$ denote the likelihood of an input value $v \in \mathbb{R}^D$ at the nth segment of the segmentation $T_{switch} \subset \{0, 1, \ldots, T\}$. Similarly, $P_{y|u}(z|v; T_{switch}, n)$ denotes the likelihood of a sequence of observations $(z[t_n], z[t_n + 1], \ldots, z[t_{n+1} - 1])$ conditioned on the input being equal to $v \in \mathbb{R}^D$ in the nth segment of the segmentation T_{switch}. Additionally, note that the assumption on P_u is specific to the Bayesian framework, and can be ignored in a frequentist framework.

In the context of energy disaggregation, these assumptions have a nice interpretation.

The condition on P_u states that, intuitively, given that devices switch at time t, how devices were being used before the switch and after the switch are independent. This may not be true in practice, since many device usages are coupled. For example, a toaster and a water kettle may be likely to activate at nearby times for a morning breakfast. However, we note that this assumption is significantly weaker than the assumption that device states are governed by a Markov chain, which is an underlying assumption of state-of-the-art hidden Markov model (HMM) techniques [Kim+11; Fro+11; KJ12; JW12; Par+12].

The condition on $P_{y|u}$ states that the observation at time t only depends on how devices are being used in that segment. The energy consumption of devices generally undergo a transient as the device switches state, and then approach a steady-state energy consumption. Intuitively, what this assumption states is that time between device switches is long enough that devices reach steady-state in each segment. As mentioned before, this assumption is motivated by the fact the time scale of these transients is much shorter than the time scale of our usage patterns, e.g. lighting reaches steady-state far more quickly than we flip

the light switches in normal use. We note that this assumption is likely necessary for good performance: if this assumption were violated, then the transient signatures of devices would collide and interfere with each other.

We note a quick consequence of Assumption 1 and Proposition 1.

Proposition 2. *Let* $T_{switch}(\widehat{u_{\text{MAP}}}) = \{t_1, t_2, \ldots, t_N\}$. *Under Assumption 1:*

$$\widehat{u_{\text{MAP}}} = \arg\max_{\widehat{u}} \left(\sum_{n=0}^{N} \left(\ln(P_u(\widehat{u}[t_n]; T_{switch}(\widehat{u_{\text{MAP}}}), n)) + \ln(P_{y|u}(y_{t_n:t_n-1} | \widehat{u}[t_n]; T_{switch}(\widehat{u_{\text{MAP}}}), n)) \right) \right) \tag{2.1.18}$$

If $T_{switch}(\widehat{u_{\text{MAP}}})$ *is given, then these optimizations are decoupled, so:*

$$\widehat{u_{\text{MAP}}}[t_n] = \arg\max_{\widehat{u}} \left(\ln(P_u(\widehat{u}; T_{switch}(\widehat{u_{\text{MAP}}}), n)) + \ln(P_{y|u}(y_{t_n:t_n-1} | \widehat{u}; T_{switch}(\widehat{u_{\text{MAP}}}), n)) \right) \tag{2.1.19}$$

Proposition 2 states that, if we are given the true switching time $T_{switch}(\widehat{u_{\text{MAP}}})$, then calculating $\widehat{u_{\text{MAP}}}$ is very tractable. In fact, if we are given a collection of potential switching times, calculating $\widehat{u_{\text{MAP}}}$ is still tractable, provided $T_{switch}(\widehat{u_{\text{MAP}}})$ is in the collection and the collection is not too large.

Next, we will develop an algorithm to achieve this: it will selectively consider candidate subsets of $\{0, 1, \ldots, T\}$ such that it considers significantly less than 2^{T+1} candidates, but still will consider $T_{switch}(\widehat{u_{\text{MAP}}})$.

We draw on results in the adaptive filtering literature, discussed in Section 2.1.1. In our particular case, we use a filter bank approach to handle the energy disaggregation problem. A filter bank is a collection of filters, and the adaptive element of a filter bank is in the insertion and deletion of filters, as well as the selection of the optimal filter.

As previously mentioned, there are 2^{T+1} different possible segmentations of $\{0, 1, \ldots, T\}$ so iteratively considering each one is intractable. The process of finding the best segmentation can be seen as exploring a binary tree. This is visualized in Figure 2.1.2.

Consider how nodes in this tree correspond to segmentations. At depth t, if $t \in T_{switch}$, then we take the 1 branch. If $t \notin T_{switch}$, then we take the 0 branch. In this fashion, each of the 2^{T+1} leaf nodes can be related to a unique segmentation. Additionally, if we pick a node at depth t, we can associate it with a switching time by just taking the 0 branch repeatedly to the leaf, i.e. assuming that no switches occur after time t. Thus, *every* node on this tree can be associated with a switching time.

Before we can introduce our algorithm, we'll define one intermediary function which calculates the maximum posterior probability when only given the measurements $y_{0:t}$. For any $t \in \{0, 1, \ldots, T\}$, $T_{switch} = \{t_1, t_2, \ldots, t_N\} \subset \{0, 1, \ldots, t-1\}$, and $y \in \mathbb{R}^{T+1}$, let $t_{N+1} = t$ and let $T'_{switch} = T_{switch} \cup t_{N+1}$. Then, we can define:

$$p_t(T_{switch} | y) = \max_{\widehat{u}: T_{switch}(\widehat{u}) = T_{switch}} \prod_{n=0}^{N} P_u(\widehat{u}[t_n]; T'_{switch}, n) P_{y|u}(y_{t_n:t_{n+1}-1} | \widehat{u}[t_n]; T'_{switch}, n) \tag{2.1.20}$$

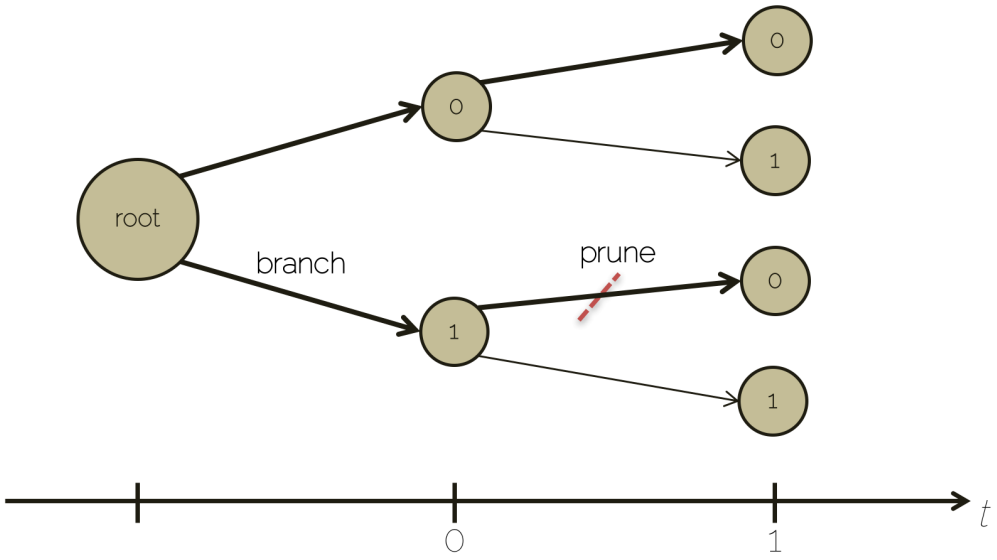

Figure 2.1: A segmentation T_{switch} can be thought of as a leaf node on a binary tree of depth T. That is, T_{switch} corresponds exactly to one leaf node of this binary tree. Additionally, we can associate a node at depth t with a switching time by assuming that no switches happen after t.

Finally, we introduce our algorithm in Algorithm 2.1.2. Note that, given y, we can calculate T since $y \in \mathbb{R}^{T+1}$. Similarly, the calculations of $P_u(v; T_{switch}, n)$, $P_{y|u}(z|v; T_{switch}, n)$, and $p_t(T_{switch}|y)$ are omitted from the pseudocode as they can be directly calculated from P_u and $P_{y|u}$ when Assumption 1 holds.

As previously hinted, our algorithm selectively explores branches of a binary tree. Limiting the growth of the filter bank $\mathcal{F}$ can be done by deciding which branches to expand and which branches to prune. This sort of formulation lends itself very easily to an online formulation of the filter banks algorithm. In fact, it is more intuitive to think of the algorithm in an online fashion.

At time t, by default, we only follow the 0 branch. For example, in Figure 2.1.2, this corresponds to following only the top path, as done at the [root→0] node. We choose to branch a filter, i.e. explore both the 0 and 1 branches on the binary tree, only if it corresponds to one of the most likely segmentations. As an example, this is done at the [root] node in Figure 2.1.2. The branching corresponds to adding the bottom path to the [root] node (creating the [root→1] node) as well as the top path (creating the [root→0] node). At the beginning of $t = 1$, the filter bank will contain $T_{switch} = \{0\}$ and $T_{switch} = \{\}$.

Additionally, at time t, we prune any paths that have sufficiently low likelihood. That is, we remove the segmentation T_{switch} from $\mathcal{F}$ if $p_t(T_{switch}|y) < p_{thres}$, where p_{thres} is an algorithm parameter. This is depicted by the dotted line in Figure 2.1.2; this involves removing $T_{switch} = \{0\}$ from our filter bank when $t = 1$. After pruning, none of this node's

Algorithm 1 Energy disaggregation via filter banks

procedure EDFB($y, P_u, P_{y|u}, p_{thres}$)
  $\mathcal{F} \leftarrow \{\emptyset\}$ ▷ Initialize the filter bank with one filter corresponding to $T_{switch} = \emptyset$.
  for $t \in \{0, 1, \ldots, T\}$ **do**
    for $T_{switch} \in \mathcal{F}$ **do** ▷ Prune the segmentations T_{switch} whose
      if $p_t(T_{switch}|y) < p_{thres}$ **then** ▷ likelihood falls below a certain threshold
        remove T_{switch} **from** $\mathcal{F}$ ▷ when considering data up until time t.
      end if
    end for
    $T_{list} \leftarrow \arg\max_{T_{switch} \in \mathcal{F}} p_t(T_{switch}|y)$ ▷ Branch the segmentations T_{switch}
    for $T_{switch} = \{t_1, t_2, \ldots, t_N\} \in T_{list}$ **do** ▷ with the highest likelihood up to time t.
      $t_{N+1} \leftarrow t$
      append $T_{switch} \cup t_{N+1}$ **to** $\mathcal{F}$
    end for
  end for
  pick any $T_{switch} \in \arg\max_{T_{switch} \in \mathcal{F}} p(T_{switch}|y)$ ▷ Find a set of
  pick any $u^* \in \arg\max_{\widehat{u}: T_{switch}(\widehat{u}) = T_{switch}} P_u(\widehat{u}) P_{y|u}(y|\widehat{u})$ ▷ optimal switching times,
  return u^* ▷ and calculate the u^* value conditioned on
    ▷ any one of the optimal switching times.
end procedure

children will be explored, so no node beginning with [root→1] will be present in the filter bank after time $t = 2$.

Thus, Algorithm 2.1.2 will only consider a subset of the leaf nodes of the binary tree. However, by intelligently deciding which subset to consider, we can maintain good performance of energy disaggregation algorithm. In fact, Algorithm 2.1.2 will recover an MAP estimate of the input under the conditions presented in the previous section. We will prove this formally below.

Theorem 1. (Optimality of energy disaggregation via filter banks without pruning)
Under Assumption 1, for $p_{thres} = 0$, the u^* returned by Algorithm 2.1.2 is a MAP estimate for the energy disaggregation problem $(P_u, P_{y|u}, y)$.

Intuitively, this proof works as follows. Suppose a MAP estimate were removed from the binary tree due to selective branching. This means that, at a particular point in time, there is a switching time t where the algorithm did not branch the correct segmentation. However, at time t, the algorithm did branch some segmentation T_{switch}, which was optimal based on the observations up until time t. By the decoupling in Proposition 2, we can improve our performance by replacing the MAP estimate's switching times up until time t with T_{switch} instead. Thus, a MAP estimate was not removed. The formal proof follows.

Proof. Suppose not, i.e. there exists an MAP estimate $\tilde{u}$ such that $P_u(\tilde{u})P_{y|u}(y|\tilde{u}) > P_u(u^*)P_{y|u}(y|u^*)$ and $P_u(\tilde{u})P_{y|u}(y|\tilde{u}) \geq P_u(v)P_{y|u}(y|v)$ for any v. By the construction of the algorithm, there exists a time $t \in T_{switch}(\tilde{u})$ such that Algorithm 2.1.2 did not branch the node corresponding to $T_{switch}(\tilde{u}) \cap \{0, 1, \ldots, t-1\}$.

Now pick a T_{switch} such that $T_{switch} \cap \{0, 1, \ldots, t-1\}$ was branched at time t. It follows that $p_t(T_{switch}|y) > p_t(T_{switch}(\tilde{u})|y)$. Let $T'_{switch} = (T_{switch} \cap \{0, 1, \ldots, t-1\}) \cup (T_{switch}(\tilde{u}) \cap \{t, t+1, \ldots, T\})$, i.e. T'_{switch} takes the switching times of T_{switch} prior to t and the switching times of $\tilde{u}$ after t. By the decoupling noted in Proposition 2, we have for any $\tau > t$:

$$p_\tau(T'_{switch} \cap \{0, 1, \ldots, \tau-1\}|y) > p_\tau(T_{switch}(\tilde{u}) \cap \{0, 1, \ldots, \tau-1\}|y) \tag{2.1.21}$$

This also includes the case where $\tau = T + 1$. That is:

$$p(T'_{switch}|y) > p(T_{switch}(\tilde{u})|y) \tag{2.1.22}$$

However, by Proposition 1, we can conclude that $\tilde{u}$ is in fact not an MAP estimate. This contradiction allows us to happily conclude that u^* is, in fact, an MAP estimate. □

Corollary 1. (Optimality of energy disaggregation via filter banks)

In addition to Assumption 1, suppose there exists an MAP estimate $\tilde{u}$ such that, for all $t \in \{0, 1, \ldots, T\}$:

$$p_t(T_{switch}(\tilde{u}) \cap \{0, 1, \ldots, t-1\}|y) \geq p_{thres} \tag{2.1.23}$$

Then, Algorithm 2.1.2 will recover an MAP estimate u^*.

Proof. Note that the hypothesis of the corollary implies that an MAP will never get pruned. Joint with Theorem 1, we recover the desired result. □

Experimental setup

To test our disaggregation method, we deployed a small-scale experiment. To collect data, we use the emonTx wireless open-source energy monitoring node from OpenEnergyMonitor[4]. We measure the current and voltage of devices with current transformer sensors and an alternating current (AC) to AC power adapter. For each device i, we record the root-mean-squared (RMS) current I^i_{RMS}, RMS voltage V^i_{RMS}, apparent power P^i_{VA}, real power P^i_{W}, power factor ϕ^i_{pf}, and a coordinated universal time (UTC) stamp. The data was collected at a frequency of 0.13Hz.

Our experiment focused on small devices commonly found in a residential or commercial office building. First, we recorded plug-level data z_i for a kettle, a toaster, a projector, a monitor, and a microwave. These devices consume anywhere from 70W to 1800W.

Then, we ran an experiment using a microwave, a toaster, and a kettle operating at different time intervals. These measurements form our ground truth y_i, and we also sum

[4] `http://openenergymonitor.org/emon/emontx`

the signals to get our aggregated power signal $y = \sum y_i$. The individual plug measurements are shown in Figure 2.2. It is worth commenting that the power consumption signals for individual devices are not entirely independent; one device turning on can influence the power consumption of another device. This coupling is likely due to the non-zero impedance of the power supply system. However, we found this effect to be negligible in our disaggregation algorithms.

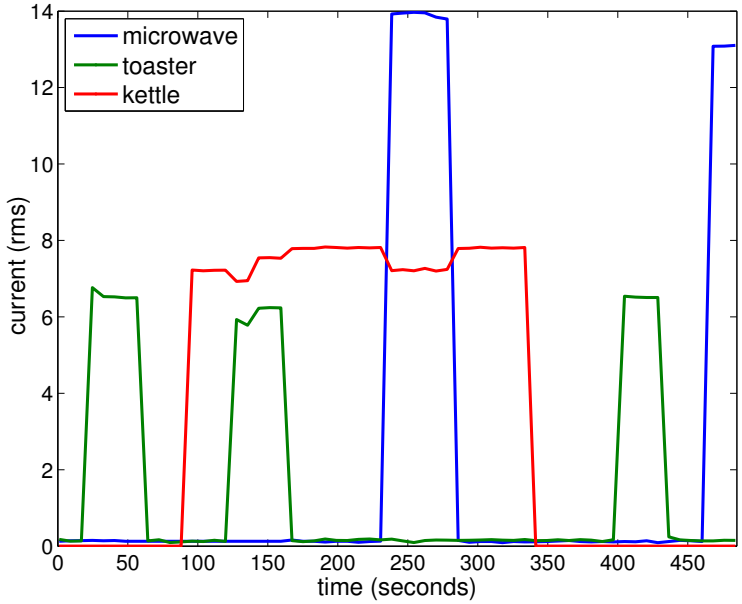

Figure 2.2: The measurements of individual plug RMS currents.

Implementation details

Several heuristics are used for pruning the binary tree depicted in Figure 2.1.2 that are specific to the task of disaggregation. First, we do not bother considering branches if the most likely segmentation explains the data sufficiently well. This greatly reduces the growth of the filter bank across time. Furthermore, we assume that at most one device switches on or off in any given time step. This unfortunately violates the assumptions of Theorem 1, but we find that it gives good results in practice.

Results

The disaggregation results are presented in Figure 2.3. We can see that the segmentation is correctly identified. Visually, the results also line up well.

We also note that it is not fair to compare results from our small-scale experiment with many of the methods mentioned in Section 2.1.1. Most of the methods listed are unsupervised

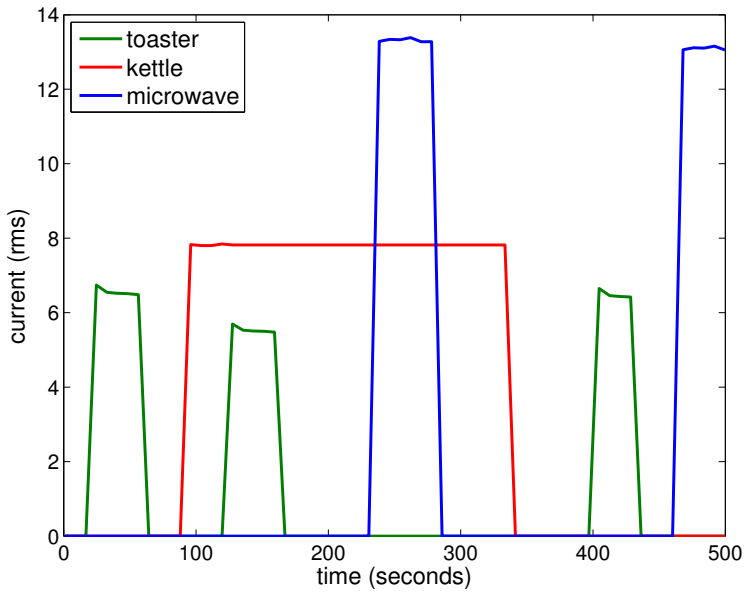

Figure 2.3: The estimated power consumption signals of each device.

methods which do not have a training set of data [KJ11; Sha+12; KJ12; JW12]. Since these unsupervised methods do not learn from training data, they have many priors which must be tuned towards the devices in the library. Also, the sparse coding method in [KN10] requires a large amount of disaggregated data to build a dictionary.

Closing remarks on energy disaggregation

In the work presented, we formalized the disaggregation problem within the filter banks framework. We provide an algorithm with guarantees on the recovery of the true solution given some assumptions on the data.

From the point of view of the utility company, the question of how to use this data to inform the consumer about their usage patterns and how to develop incentives for behavior modification is still largely an open one, which we are currently studying.

Another largely open question is the one concerning privacy. Given that energy data can be disaggregated with some degree of precision, how does this affect the consumer's privacy? The next natural step is to study how this data can be used in a privacy preserving way to improve energy efficiency. These privacy preserving policies may come in the form of selectively transmitting the most relevant data for a control objective, or incentive mechanisms for users to change their consumption behavior without direct transmission of their private information to the utility company. This will be discussed in Chapter 3.

2.2 Blind system identification

In Section 2.1, we assume we are given P_u and $P_{y|u}$. In the energy disaggregation application, P_u is the model for the usage patterns of consumers and $P_{y|u}$ is the dynamics of devices. In practice, we are not always given these models, and will have to learn them from data.

If we are given inputs u and outputs y, the task of estimating $P_{y|u}$ is known as the *system identification* problem. In our energy disaggregation application, we often do not record u_i, the usage patterns inside the home, but only record y_i, the energy traces of individual devices. The lack of u measurements makes the problem more difficult: we must simultaneously estimate u and $P_{y|u}$. This work appeared previously in [Ohl+14].

2.2.1 Problem introduction

Consider an auto-regressive exogenous input (ARX) model:

$$y(t) = \sum_{k=1}^{n_a} a_k y(t-k) + \sum_{k=1}^{n_b} b_k u(t-n_k-k) \tag{2.2.1}$$

Here, the inputs and outputs are both scalar, i.e. $u(t) \in \mathbb{R}$ and $y(t) \in \mathbb{R}$ for all t. Estimation of this type of model is one of the most common tasks in system identification and a very well studied problem (see, for instance: [Lju99]).

The common setting is that $\{(y(t), u(t))\}_{t=1}^N$ is given and the summed residuals

$$\sum_{t=n}^{N} \left(y(t) - \left[\sum_{k=1}^{n_a} a_k y(t-k) + \sum_{k=1}^{n_b} b_k u(t-n_k-k) \right] \right)^2 \tag{2.2.2}$$

is minimized to obtain an estimate for $a_1, \ldots, a_{n_a}, b_1, \ldots, b_{n_b}$. (Here, $n = \max(n_a, n_k + n_b) + 1$.) This estimate is often referred to as the *least squares* (LS) estimate.

In the following text, we study the more complicated problem of estimating an ARX model from solely outputs $\{y(t)\}_{t=1}^N$. This is an ill-posed problem and it is easy to see that under no further assumptions, it would be impossible to uniquely determine $a_1, \ldots, a_{n_a}, b_1, \ldots, b_{n_b}$. We will here therefore study this problem under the assumption that the stacked inputs belong to some known subspace. For example, the input could be known to change only at a set of discrete times due to a discrete controller, or known to be band-limited and therefore well represented by the projection on the first discrete Fourier transform basis vectors.

It should be noticed that this assumption is not enough to uniquely determine the input or the ARX model. Specifically, we will not be able to decide the input or the ARX coefficients $b_1, \ldots, b_{n_b}$ more than up to a multiplicative scalar. It should be stressed that this is not a limitation of the method that we propose but an inherent limitation of the system identification problem since the sought quantities always appear as products. To uniquely determine the input and the ARX coefficients $b_1, \ldots, b_{n_b}$, further knowledge is needed.

Our main contribution is a novel method for ARX model identification from only output measurements. The method takes the form of a convex optimization problem and gives

a computationally flexible framework for handling different types of measurement noises, constraints,&c.

2.2.2 Related work

Blind system identification (BSI) has a broad application area and has been applied in fields such as data communications, speech recognition and seismic signal processing [AM+97b]. Common for the type of modeling problems that BSI has been applied to is that the input is difficult, costly or impossible to measure. For example, in exploration seismology, the physical properties of the earth are explored by studying the response of an excitation (often a charge of dynamite) [Zer+99]. The excitation is often difficult to measure and the modeling problem therefore a BSI problem.

Many methods have been proposed to solve the BSI problem throughout the years. We give a short overview here but refer the interested reader to [AM+97b; Hua02], for a more extensive and complete review.

The maximum likelihood (ML) approach to BSI aims at finding the ML estimate of the model and input. The resulting non-convex optimization problem is often treated by alternating between optimizing with respect to the input and the system model. The channel subspace (CS) methods to BSI indirectly determine the sought finite impulse response (FIR) model by estimating the nullspace of the Sylvester matrix associated with the FIR model to be identified. This is done by an eigendecomposition of a matrix derived from the outputs (see, for instance: [AM+97a]). The method proposed in [Van+13] works under the assumption that two or more output series are available and that these were generated by the same input. The methods proposed in [Sat75; Ton+91] assume that the input consists of independent and identically (iid) distributed random variables and considers the autocorrelation of the output to decide a FIR model and the unknown input.

A number of approaches consider the blind identification problem of Hammerstein systems under the assumption that the input is piecewise constant. [Sun+99; Bai+02; BF02; Wan+07; Wan+10]. Our approach assumes that the input belongs to some known subspace. Piecewise constant signal can be represented using the subspace assumption used here. However, we note that we are not restricted to piecewise constant signals, and our approach is significantly different. Also, we consider the blind identification of ARX models while the blind identification problem of Hammerstein systems is considered in [Sun+99; Bai+02; BF02; Wan+07; Wan+10].

The related problem of blind deconvolution have been studied in a number of contributions. In particular, see the very interesting paper by [Ahm+12] for a solution where the signals to be recovered are assumed to be in some known subspaces. The development presented in [Ahm+12] has similarities to the approach presented in this paper and was done in parellel to our work. Note that only FIR models are discussed in [Ahm+12] and that the analysis does not apply.

2.2.3 Blind system identification via lifting

First, we will introduce some notation.

We will use y to denote the output and u the input. For simplicity, we will only consider *single input single output* (SISO) systems. However, with some extra bookkeeping, MIMO systems could also be treated. We will assume that N measurements of y are available and stack them in the vector $\mathbf{y}$:

$$\mathbf{y} = \begin{bmatrix} y(1) & \dots & y(N) \end{bmatrix}^T \tag{2.2.3}$$

We also introduce $\mathbf{u}$, $\mathbf{a}$ and $\mathbf{b}$ as:

$$\mathbf{u} = \begin{bmatrix} u(1) & \dots & u(N) \end{bmatrix}^T \tag{2.2.4}$$

$$\mathbf{a} = \begin{bmatrix} a_1 & \dots & a_{n_a} \end{bmatrix}^T \tag{2.2.5}$$

$$\mathbf{b} = \begin{bmatrix} b_1 & \dots & b_{n_b} \end{bmatrix}^T \tag{2.2.6}$$

We will use $\mathbf{y}(i)$ to denote the ith element of $\mathbf{y}$. To pick out a subvector of $\mathbf{y}$ consisting of the ith to the jth element we will use the notation $\mathbf{y}(i:j)$ and similarly for picking out a subvector of $\mathbf{u}$, $\mathbf{a}$ and $\mathbf{b}$. To pick out a submatrix consisting of the ith to the jth rows of $\mathbf{X}$ we use the notation $\mathbf{X}(i:j,:)$. We will use normal font to represent scalars and bold for vectors and matrices.

$\|\cdot\|_0$ is the cardinality operator which returns the number of nonzero elements of its argument and $\|\cdot\|_*$ the nuclear norm returning the sum of the singular values.

Now, we can formally define the problem of BSI in the noiseless case.

Assumption 2. *The sought input,* $\mathbf{u}$*, lies in some known subspace, i.e.*

$$\mathbf{u} = \mathbf{D}\mathbf{x} \tag{2.2.7}$$

for some known $N \times m$*-matrix* $\mathbf{D}$ *and an unknown vector* $\mathbf{x} \in \Re^m$*. It is assumed that* $m \leq N$*.*

Problem Statement 2. (Blind system identification without noise)

Given the sequence of outputs $\{y(t)\}_{t=1}^N \in \mathbb{R}$*, find an estimate for* $a_1, \dots, a_{n_a}, b_1, \dots, b_{n_b} \in \mathbb{R}$ *and* $u(t) \in \mathbb{R}, t = 1, \dots, N$ *such that*

$$y(t) = \sum_{k=1}^{n_a} a_k y(t-k) + \sum_{k=1}^{n_b} b_k u(t-n_k-k) \tag{2.2.8}$$

for $t = n, \dots, N$*, where* $n = \max(n_a, n_k + n_b) + 1$*. We will for simplicity assume that* n_a, n_b, n_k*, are known. To make the problem well-posed, we suppose Assumption 2 holds.*

We can formulate the problem of finding an input and the ARX coefficients as the feasibility problem:

$$\text{find } \mathbf{u}, \mathbf{b}, \mathbf{a} \tag{2.2.9}$$

$$\text{such that } y(t) - \sum_{k=1}^{n_a} a_k y(t-k) = \sum_{k=1}^{n_b} b_k u(t-n_k-k) \quad t = n, \dots, N \tag{2.2.10}$$

This problem is non-convex.

Introduce $\mathbf{X} = \mathbf{x}\mathbf{b}^T \in \mathbb{R}^{m \times n_b}$ and note that Equation 2.2.7 gives that $\mathbf{DX} = \mathbf{Dx}\mathbf{b}^T = \mathbf{u}\mathbf{b}^T$. Since $\mathbf{u}\mathbf{b}^T$ contains all products $u(i)b_j$, $i = 1, \ldots, N$, $j = 1, \ldots, n_b$, the sum

$$\sum_{k_1=1}^{n_b} b_{k_1} u(t - n_k - k_1) \tag{2.2.11}$$

can be realized by summing appropriate entries of $\mathbf{DX}$.

Thus, Problem 2.2.9 can be reformulated as:

$$\text{find } \mathbf{X}, \mathbf{a} \tag{2.2.12}$$

$$\text{such that } y(t) - \sum_{k=1}^{n_a} a_k y(t-k) = \sum_{k=1}^{n_b} (DX)(t - n_k - k, k) \quad t = n, \ldots, N \tag{2.2.13}$$

$$\text{rank}(\mathbf{X}) = 1 \tag{2.2.14}$$

Note that we need to require that $\text{rank}(\mathbf{X}) = 1$ to be able to decompose $\mathbf{X}$ as $\mathbf{X} = \mathbf{x}\mathbf{b}^T$. Problem 2.2.12 is equivalent to Problem 2.2.9 in the following sense. Assume that Problem 2.2.12, has a unique solution $\mathbf{X}^*$, then $\mathbf{X}^*$ must satisfy $\mathbf{X}^* = \mathbf{x}^*(\mathbf{b}^*)^T$, with $\mathbf{x}^*$ and $\mathbf{b}^*$ solving Problem 2.2.9. Extracting the rank-1 component of $\mathbf{X}^*$, using e.g. singular value decomposition, we can hence decide both $\mathbf{x}^*$ (and $\mathbf{u}^* = \mathbf{D}\mathbf{x}^*$) and $\mathbf{b}^*$ up to a multiplicative scalar[5]. The estimates of $\mathbf{a}$ will be identical for both problems (if the estimates are unique).

The technique of introducing the matrix $\mathbf{X}$ to avoid products between $\mathbf{x}$ and $\mathbf{b}$ is well known in optimization and referred to as *lifting* [Sho87; LS91; GW95; Nes98].

Problem 2.2.12 is a non-convex optimization problem and not easier to solve than 2.2.9. To get an optimization problem we can solve, we remove the rank constraint and instead minimize the rank. Since the rank of a matrix is not a convex function, we replace the rank with a convex heuristic. Here we choose the nuclear norm, but other heuristics are also available (see, for instance: [Faz+01]). We then obtain the convex program:

$$\min_{\mathbf{X},\mathbf{a}} \ \|\mathbf{X}\|_* \tag{2.2.15}$$

$$\text{subject to } y(t) - \sum_{k=1}^{n_a} a_k y(t-k) = \sum_{k=1}^{n_b} (DX)(t - n_k - k, k) \quad t = n, \ldots, N \tag{2.2.16}$$

We will refer to this problem as *blind identification via lifting* (BIL).

Lastly, we will not see Equation 2.2.8 hold exactly in practice. In the noisy setting, we have to tolerate some nonzero modeling error. We consider two noisy cases.

Problem Statement 3. (Blind system identification with bounded noise)

[5]Note that this is a fundamental ambiguity existent in Problem 2.2.9 as well.

Given the sequence of outputs $\{y(t)\}_{t=1}^N \in \mathbb{R}$ *and a noise bound* ϵ*, find an estimate for* $a_1, \ldots, a_{n_a}, b_1, \ldots, b_{n_b} \in \mathbb{R}$ *and* $u(t) \in \mathbb{R}, t = 1, \ldots, N$ *and bounded noise* $|w(t)| \leq \epsilon$ *such that*

$$y(t) = \sum_{k=1}^{n_a} a_k y(t-k) + \sum_{k=1}^{n_b} b_k u(t-n_k-k) + w(t) \tag{2.2.17}$$

for $t = n, \ldots, N$*, where* $n = \max(n_a, n_k + n_b) + 1$*. Again, we suppose Assumption 2 holds and* n_a, n_b, n_k*, are known.*

Our development suggests the following optimization to solve BSI in the case of bounded noise:

$$\min_{\mathbf{X},\mathbf{a},\eta} \ \|\mathbf{X}\|_* \tag{2.2.18}$$

$$\text{subject to } y(t) - \sum_{k=1}^{n_a} a_k y(t-k) = \sum_{k=1}^{n_b} (DX)(t-n_k-k,k) + \eta(t) \quad t = n, \ldots, N \tag{2.2.19}$$

$$|\eta(t)| \leq \epsilon \tag{2.2.20}$$

Similarly, we can derive an optimization problem when the noise is Gaussian.

Problem Statement 4. (Blind system identification with Gaussian noise)
Given the sequence of outputs $\{y(t)\}_{t=1}^N \in \mathbb{R}$*, find the maximum likelihood (ML) estimate for* $a_1, \ldots, a_{n_a}, b_1, \ldots, b_{n_b} \in \mathbb{R}$ *and* $u(t) \in \mathbb{R}, t = 1, \ldots, N$ *such that*

$$y(t) = \sum_{k=1}^{n_a} a_k y(t-k) + \sum_{k=1}^{n_b} b_k u(t-n_k-k) + w(t) \tag{2.2.21}$$

for $t = n, \ldots, N$*, where* $n = \max(n_a, n_k + n_b) + 1$*. Here,* $w(t) \sim N(0, \sigma^2)$ *is i.i.d. across time. Again, we suppose Assumption 2 holds and* n_a, n_b, n_k*, are known.*

A good heuristic for finding the ML estimate can be found by solving the optimization problem below for different values of parameter λ and using the largest λ such that $\mathbf{X}$ is rank 1.

$$\min_{\mathbf{X},\mathbf{a},\eta} \ \|\mathbf{X}\|_* + \lambda \|\eta\|_2^2 \tag{2.2.22}$$

$$\text{subject to } y(t) - \sum_{k=1}^{n_a} a_k y(t-k) = \sum_{k=1}^{n_b} (DX)(t-n_k-k,k) + \eta(t) \quad t = n, \ldots, N \tag{2.2.23}$$

Analysis

The number of optimization variables in Problem 2.2.9 is essentially $n_a + n_b + m$, under Assumption 2. Therefore, we cannot expect a reliable identification result from fewer than $n_a + n_b + m$ measurements. One may wonder how many measurements that are needed. Using that the constraint (2.2.16) of BIL is linear in $\mathbf{X}$, we have the following result:

Theorem 2. (Guaranteed Recovery using BIL)

Consider the noise-free blind ARX identification problem presented in Problem 2.2.9, and assume that it has a unique solution (up to a multiplicative scalar).

Let the row vector $d_i \in \mathbb{R}^m$ be the i:th row of $\mathbf{D}$ and let:

$$A = \begin{bmatrix} d_{n-n_k-1} & d_{n-n_k-2} & \dots & d_{n-n_k-n_b} & y(n-1) & \dots & y(n-n_a) \\ d_{n-n_k} & d_{n-n_k-1} & \dots & d_{n-n_k-n_b+1} & y(n) & \dots & y(n-n_a+1) \\ \vdots & & \ddots & \vdots & \vdots & & \vdots \\ d_{n-n_k-2+n_b} & \dots & & d_{n-n_k-1} & y(n+n_b-2) & \dots & y(n-n_a+n_b-1) \\ \vdots & & & \vdots & \vdots & & \vdots \\ d_{N-n_k-1} & d_{N-n_k-2} & \dots & d_{N-n_k-n_b} & y(N-1) & \dots & y(N-n_a) \end{bmatrix}$$

If A has full column rank, then the ARX model and input solving Problem 2.2.9 are recovered, up to a multiplicative scalar, by BIL, the solution to Problem 2.2.15.

Proof. By assumption, Problem 2.2.9 has a unique solution $(\mathbf{a}^*, \mathbf{b}^*, \mathbf{x}^*)$. Form $\mathbf{X}^* = \mathbf{x}^*(\mathbf{b}^*)^T$. Define θ^* as

$$\theta^* = \begin{bmatrix} \mathbf{X}^*(:,1)^T & \mathbf{X}^*(:,2)^T & \dots & \mathbf{X}^*(:,n_b)^T & -a_1^* & \dots & -a_{n_a}^* \end{bmatrix}^T.$$

We must have that

$$\begin{bmatrix} y(n) & y(n+1) & \dots & y(N) \end{bmatrix}^T = \mathbf{A}\theta^*. \tag{2.2.24}$$

Note that the pair $\mathbf{X}^*$ and $\mathbf{a}^*$ is a feasible solution to Problem 2.2.15. Since $\mathbf{A}$ has full column rank, the solution to Problem 2.2.24 must be unique. Thus, BIL gives θ^*. □

Note that if the linear constraints (2.2.16) are sufficient to uniquely give the solution of BIL, no optimization is necessary. Seeking the matrix $\mathbf{X}$ that gives the minimum nuclear norm is only of interest if we have too few measurements for the constraints to uniquely define the solution but more measurements than $n_a + n_b + m$.

Computing $\lambda^{\min}$

In the noisy version of BIL, Problem 2.2.22, the design parameter λ has to be chosen. Since λ regulates the tradeoff between the nuclear norm and the squared norm of the estimated noise η, it is natural to seek the largest λ such that the estimate $\mathbf{X}$ is rank 1. In seeking this λ, the value for $\lambda^{\min}$ may come handy, where $\lambda^{\min}$ is defined as the largest λ such that $\mathbf{X} = 0$ in BIL. Since the estimate for $\mathbf{X}$ will stay the same for all $\lambda \leq \lambda^{\min}$, we should limit our search of λ to be within $[\lambda^{\min}, \infty]$. One may for example start with $\lambda = \lambda^{\min}$ and then successively increase λ as long as $\text{rank}(\mathbf{X}) = 1$.

Theorem 3. (Computing $\lambda^{\min}$)

Consider the optimization problem given in Problem 2.2.22. There exists a λ^{min} such that whenever $\lambda \leq \lambda^{min}$, solving Problem 2.2.22 results in $\mathbf{X} = 0$. λ^{min} is given by:

$$(\lambda^{min})^{-1} = \min_{\mathbf{V} \in \mathbb{R}^{m \times n_b}} \quad \|\mathbf{V}\| \tag{2.2.25}$$

$$\text{subject to} \quad \mathbf{V}(i,j) = 2 \sum_{t=n}^{N} \left(y(t) - \sum_{k=1}^{n_a} \hat{a}_k y(t-k) \right) D(t - n_k - j, i) \tag{2.2.26}$$

with

$$(\hat{a}_1, \ldots, \hat{a}_{n_a}) = \arg\min_{\mathbf{a}} \sum_{t=n}^{N} \left(y(t) - \sum_{k=1}^{n_a} a_k y(t-k) \right)^2. \tag{2.2.27}$$

Here $\|\cdot\|$ denotes the operator norm.

Proof. Problem 2.2.22 can be rewritten as:

$$\min_{\mathbf{X},\mathbf{a}} \ \|\mathbf{X}\|_* + \lambda \sum_{t=n}^{N} \left(y(t) - \sum_{k=1}^{n_a} a_k y(t-k) - \sum_{k=1}^{n_b} (\mathbf{DX})(t - n_k - k, k) \right)^2 \tag{2.2.28}$$

Let f_λ denote the objective function in Problem 2.2.28:

$$f_\lambda(\mathbf{X}, \mathbf{a}) = \|\mathbf{X}\|_* + \lambda \sum_{t=n}^{N} \left(y(t) - \sum_{k=1}^{n_a} a_k y(t-k) - \sum_{k=1}^{n_b} (\mathbf{DX})(t - n_k - k, k) \right)^2 \tag{2.2.29}$$

Let $\partial f(x)$ denote the subgradient of f at x [Roc70, Section 23]. Then, $(\mathbf{X}, \mathbf{a}) = (0, \hat{\mathbf{a}})$ is a solution for Problem 2.2.28 if and only if $0 \in \partial f_\lambda(0, \hat{\mathbf{a}})$ [Roc70, Section 27]. We can calculate the subgradient $\partial f_\lambda(0, \hat{\mathbf{a}})$ [Wat92; Rec+10]: $(\mathbf{Y}, \alpha) \in \partial f_\lambda(0, \hat{\mathbf{a}})$ if there exists a $\|V\| \leq 1$ such that:

$$Y(i,j) = V(i,j) - 2\lambda \sum_{t=n}^{N} \left(y(t) - \sum_{k=1}^{n_a} \hat{a}_k y(t-k) \right) \mathbf{D}(t - n_k - j, i) \tag{2.2.30}$$

$$\alpha_k = -2\lambda \sum_{t=n}^{N} \left(y(t) - \sum_{l=1}^{n_a} \hat{a}_l y(t-l) \right) y(t-k) \tag{2.2.31}$$

Note that Equation 2.2.31 will hold for $\alpha = 0$ by the definition of $\hat{\mathbf{a}}$. Thus, $0 \in \partial f_\lambda(0, \hat{\mathbf{a}})$ if and only if $\|V\| \leq 1$, where V is defined by:

$$\mathbf{V}(i,j) = 2\lambda \sum_{t=n}^{N} \left(y(t) - \sum_{k=1}^{n_a} \hat{a}_k y(t-k) \right) \mathbf{D}(t - n_k - j, i) \tag{2.2.32}$$

Finally, note that the definition of $\mathbf{V}$ in Equation 2.2.32 is positively homogenous in λ, so if $0 \in \partial f_\lambda(0, \hat{\mathbf{a}})$ for some $\lambda > 0$, then $0 \in \partial f_{\lambda'}(0, \hat{\mathbf{a}})$ for all $0 < \lambda' < \lambda$.

Thus, $\lambda^{\min}$ is given by:

$$\sup_{\lambda, \|V\| \leq 1} \quad \lambda \tag{2.2.33}$$

$$\text{such that } \mathbf{V}(i,j) = 2\lambda \sum_{t=n}^{N} \left(y(t) - \sum_{k=1}^{n_a} \hat{a}_k y(t-k) \right) \mathbf{D}(t - n_k - j, i) \tag{2.2.34}$$

$$\text{for } i = 1, \ldots, m \text{ and } j = 1, \ldots, n_b \tag{2.2.35}$$

That is, λ^* is the largest λ such that the V defined by Equation 2.2.32 is such that $\|V\| \leq 1$. This is equivalent to Problem 2.2.25. □

$\lambda^{\min}$ was also numerically verified.

Solution algorithms and software

Many standard methods of convex optimization can be used to solve Problems 2.2.15, 2.2.18, and 2.2.22. Systems such as CVX [GB08; GB14] or YALMIP [Lof04] can readily handle the nuclear norm. For large scale problems, the *alternating direction method of multipliers* (ADMM) is an attractive choice [BT89; Boy+11] and we have previously shown that ADMM can be very efficient on similar problems [Ohl+13]. Code for solving 2.2.15, 2.2.18 and 2.2.22 can be found online at `http://www.rt.isy.liu.se/~ohlsson/code.html`.

Numerical illustration

Consider the system given in the diagram below.

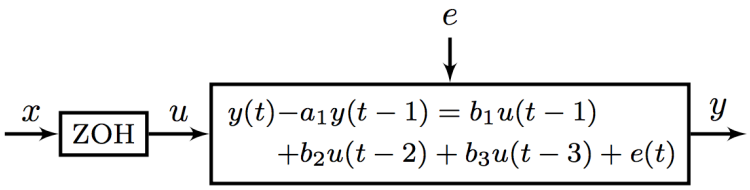

Figure 2.4: System considered in the numerical illustration.

Here the values x were generated by independently sampling from a unit Gaussian and the noise e by independently sampling from a uniform distribution between $-\epsilon/2$ and $\epsilon/2$. The ZOH (zero-order hold) block holds the input to the ARX system constant for 6 consecutive

samples. We can therefore express $\mathbf{u}$ in terms of $\mathbf{x}$ as:

$$\mathbf{u} = \begin{bmatrix} \mathbf{1}_{6\times1} & \mathbf{0}_{6\times1} & \mathbf{0}_{6\times1} & \dots & \mathbf{0}_{6\times1} \\ \mathbf{0}_{6\times1} & \mathbf{1}_{6\times1} & \mathbf{0}_{6\times1} & \dots & \mathbf{0}_{6\times1} \\ \vdots & & \ddots & & \vdots \\ \mathbf{0}_{6\times1} & \dots & \mathbf{0}_{6\times1} & \mathbf{1}_{6\times1} & \mathbf{0}_{6\times1} \\ \mathbf{0}_{6\times1} & \dots & \mathbf{0}_{6\times1} & \mathbf{0}_{6\times1} & \mathbf{1}_{6\times1} \end{bmatrix} \tag{2.2.36}$$

We identify the matrix in the relation between $\mathbf{u}$ and $\mathbf{x}$ as $\mathbf{D}$.

The ARX coefficients used were

$$a_1 = -0.3, b_3 = 1, b_2 = 2, b_1 = 3 \tag{2.2.37}$$

Figure 2.2.3 shows the output y for $\epsilon = 5$.

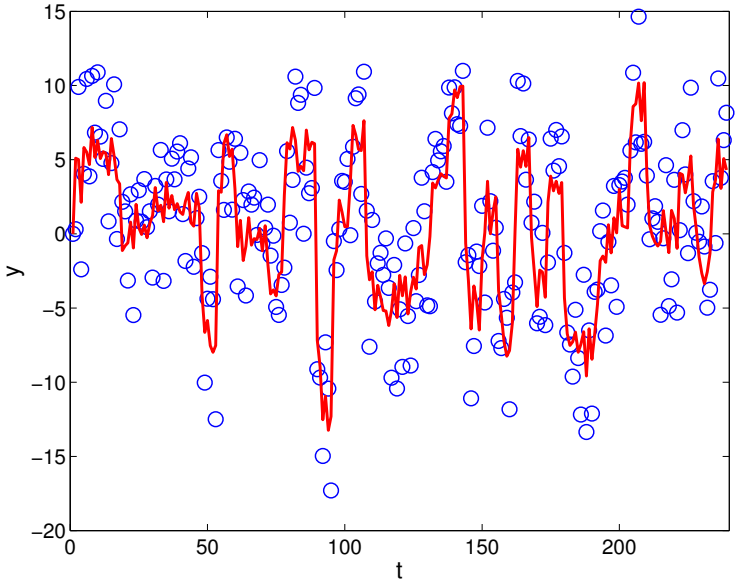

Figure 2.5: The noise-free (solid line) and noisy outputs (circles).

If Problem 2.2.22, the noisy version of BIL, is used to estimate $\mathbf{u}$ and an ARX model, we get the input-estimate given in Figure 2.2.3 and the ARX coefficients:

$$a_1 = -0.21, b_3 = 0.91, b_2 = 1.80, b_1 = 2.7. \tag{2.2.38}$$

It is interesting to notice that if we instead would be given the true input $\mathbf{u}$ and only estimated the ARX coefficients by minimizing the squared residuals between the output $\mathbf{y}$ and the predicted output, we would have got the estimates:

$$a_1 = -0.30, b_3 = 0.88, b_2 = 2.39, b_1 = 2.85. \tag{2.2.39}$$

As seen, these estimates are not that much better than what BIL is providing (see (2.2.38)). Remember that BIL is only given the $\mathbf{y}$ measurements and not the inputs $\mathbf{u}$. It is therefore quite remarkable that the estimates of BIL is comparable to those given in (2.2.39).

To further study the robustness of BIL we carried out a Monte Carlo simulation. In the simulation, the noise level ϵ was varied between 0 and 5. For each noise level, 100 trials were carried out with different noise and input realizations. The true ARX model was kept fixed (the same as above). The results are summarized in Figure 2.2.3.

The setup of above example does not imply that $\mathbf{A}$ has full column rank. Nevertheless, a perfect result was obtained in the noise free case.

If $\mathbf{D}$ is instead generated by independently sampling each element from a unit Gaussian distribution (everything else unchanged), the resulting $\mathbf{A}$ has full column rank with high probability.

To recap, this subsection presents a novel framework for the blind system identification of ARX models, under Assumption 2. The framework uses the fact that the problem can be rewritten as a rank minimization problem. A convex relaxation is presented to approximate the sought ARX parameters and the unknown inputs. This method was also applied to problems of blind subspace identification for multiple-input, multiple-output systems in [Sco+15].

2.3 Conclusion

In this chapter, we discussed disaggregation and inference algorithms with applications to IoT. Disaggregation is a general problem that will arise in many IoT contexts, as large amounts of users and data will be aggregated and disaggregated at various stages in data transmission, storage, and retrieval. We discussed a detailed example in the context of energy, and provided a tractable algorithm that can solve a combinatorial problem with guaranteed performance. We also discussed blind system identification techniques for estimating the dynamics of systems from sensor readings when very little information is available about their operational usage and internal physics. These form useful tools in taking IoT sensor measurements and creating models from which new actuation and mechanisms can be designed.

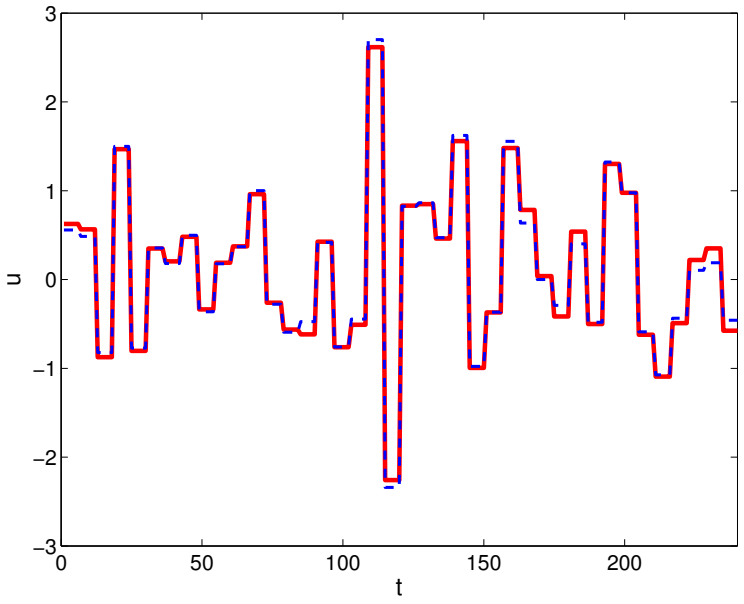

Figure 2.6: The estimated (dashed line) and the true input **u** (solid line).

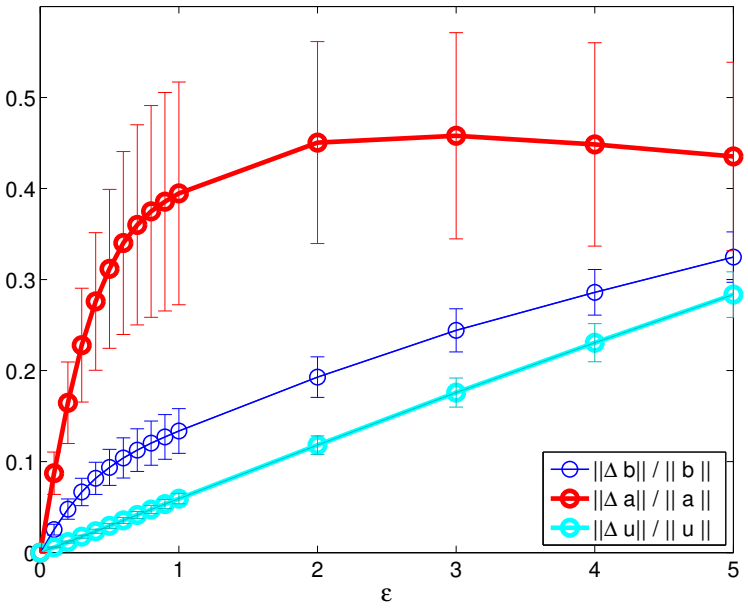

Figure 2.7: The relative errors along with their 0.5 standard deviation error bounds for varying noise levels.

Chapter 3

Design Paradigms for the Privacy of IoT Consumers and their Data

The Internet of Things (IoT) promises many advantages in the control and monitoring of physical systems, from both efficacy and efficiency perspectives. However, in the wrong hands, the data might pose a privacy threat, as was alluded to in Chapter 2. In this chapter, we discuss different methods by which one can quantify privacy, different design paradigms for incorporating privacy into the design of an IoT system, and discuss different applications of each of these paradigms.

The Internet of Things collects and transmits a large amount of data. This data enables a multitude of advantages to all parties, but also presents a privacy risk to consumers who are now sharing data that may contain private information, or may be statistically correlated with information considered private. For example, in the smart grid, IoT devices promise better efficiency in energy distribution, more reliability, and transparency to electric utility customers in their energy consumption. On the other hand, monitoring energy consumption at high granularity can allow the inference of detailed information about consumers' lives such as the times they eat, when they watch TV, and when they take a shower [Lis+10]. Such information is highly valuable and will be sought by many parties, including advertising companies [AF10], law enforcement [Smi12], and criminals [Gov11].

As IoT modernizes our lives and infrastructures, privacy emerges as a major concern. In this chapter, we hope to address privacy issues in IoT by presenting a formal framework by which privacy-aware IoT service models can be designed. Currently, legislators are investigating technology-aware policies to ensure customer privacy; the final form of these policies is yet to be determined but will likely have a large impact on the shape of IoT in the future. Additionally, researchers are looking at ways to design these systems in a fashion which acknowledges privacy.

Again referring to the smart grid example for concreteness: governments, researchers, and organizations are working on privacy standards and policies to guide advanced metering infrastructure (AMI) deployments. Researchers have considered the issue of data privacy in smart grid infrastructures, and have proposed novel mechanisms for protecting the col-

lected data (encryption, access control, and cryptographic commitments) [Kur+11; RD11], by anonymization and aggregation [TG09; Li+10], and by preventing inferences and re-identification from databases that allow queries from untrusted third parties (via differential privacy) [AC11].

This chapter is organized as follows. First, we discuss some of the existing literature in privacy in Section 3.1. In Section 3.2, we provide a high-level overview of the different methods to quantify privacy. Additionally, some of these methods are our contribution, and we present theoretical results for these privacy metrics in this section as well. In Section 3.3, we discuss three levels of resolution in which privacy can be incorporated into the design and analysis of IoT systems: passive privacy analysis, active privacy mechanisms, and optimal privacy-by-design. Finally, in Section 3.4, we provide several examples of privacy design being instantiated in the context of different cyber-physical infrastructures. These include examples in ground transportation networks, direct load control programs in the smart grid, and building control using occupancy sensors.

3.1 Related work

Some of the earliest literature in applied privacy was ensuring that surveys could be conducted in a privacy-preserving fashion. These methods were called *randomized response* methods [War65; Gre+69]. These researchers noticed there was structural bias when surveys requested sensitive information, such as whether or not a subject was HIV-positive. The key component for guaranteeing privacy was to given individual subjects *deniability*: a positive answer could either be a true response, or due to the randomness in the survey procedure.

The next advances in the applied privacy literature was in statistical databases. In [Dal77], the author argues that any definition of privacy should satisfy the following desideratum: nothing can be learned about a user with the database that could not be learned without the database. One attempt to satisfy this desideratum was k-anonymity [Swe02], which provides methods to ensure that for any one user, there are at least $k-1$ users who appear indistinguishable from said user.

More recently, the advent of Big Data has introduced many databases with potentially sensitive data that could be utilized by an adversary as side information to infer private facts. For example, in [NS06], the authors are able to take anonymized Netflix data and, using publicly available information from IMDB, recover the identities of individual users. These results pushed researchers to no longer consider privacy of a database in isolation, but in the larger context of widely available side information.

Recent research has been focused on designing privacy metrics to quantify privacy risks. This will be discussed in more detail in Section 3.2.

Arguably the most popular privacy metric, *differential privacy* was introduced in [Dwo06]. Differential privacy requires an exogenous adjacency relationship, which specifies pairs of potential values for private parameters that we hope to keep indistinguishable. With this

adjacency relationship, differential privacy is a bound on the change in the distribution of the observables between any two adjacent private parameters.

Differential privacy is attack-agnostic, in the sense that as a metric it does not suppose the adversary launches a particular type of inference attack. Furthermore, differential privacy is also agnostic to the amount of side information an adversary has, since it simply captures how much the distribution of the observable changes for small perturbations to private parameters.

An alternative definition uses an information-theoretic metric to quantify privacy loss. In particular, *mutual information* between a private parameter and a public observable has recently become a popular metric [PCF12a; San+13]. One interpretation of the mutual information is the difference between the entropy of the prior distribution and the entropy of the posterior distribution [CT91]; from that perspective, this metric has an intuitive interpretation as quantifying the reduction in the uncertainty of an adversary due to a public observable. This metric is attack-agnostic, since it simply quantifies a statistical relationship between private and public variables. However, this requires a specification of the available side information to an adversary, as reflected in the prior distribution.

It is our belief that, similar to previous technological changes, the Internet of Things will motivate a sea change in how privacy is perceived, defined, quantified, and treated. All the previously state references consider privacy in the context of databases, but a nascent area of research is the investigation of how privacy can be understood in the context of systems with dynamics.

Recent work in this regard include the extension of differential privacy to Kalman filtering [LNP14], constrained optimization [Han+14] and convex optimization [Hsu+14], distributed control [Hua+14b], and online learning [Don+15]. Similarly, there have been efforts to consider information-theoretic metrics in the context of dynamic systems such as the smart grid [Raj+11; Jia+16].

3.2 Quantifying privacy

The first step to any technical analysis of privacy is to formulate a model of privacy. Somehow, as researchers, we must make regularizing assumptions that turns a complicated, contextually dependent, essentially contested concept [Sol02] with a plurality of definitions [Nis04] into a mathematical object that can be incorporated into the engineering design of IoT technologies.

Broadly speaking, these methods can be put into three general categories: non-statistical methods, differential privacy, and inferential privacy.

3.2.1 Non-statistical methods

We use the term *non-statistical methods* to refer to any privacy methods that do not rely on probabilistic models. In the human-computer interaction (HCI), human-robot interaction

(HRI), and behavioral economics communities, these methods are particularly popular, but these methods are also quite popular broadly. We briefly outline sample works from this class of privacy methods.

There are methods in what is often referred to as *usable privacy*. These methods outline a particular situation in which there is a privacy problematic and design solutions for these situations. For example, [Sha+16] works on designing password policies that are reliable, easy-to-remember, and still secure against most attacks. In [Roe+12], the authors re-design smart phone applications to limit their access to private materials in a convenient, easy-to-implement and user-intuitive fashion. For example, rather than granting the Facebook app constant access to one's camera, a user can instead only allow Facebook to access the camera when she clicks a camera button, which will be chosen from a pre-determined set of icons that serve as clear indicators that grant camera access permission.

Alternatively, researchers quantify privacy preferences using behavioral studies. For example, in [Acq+15], the authors experiment to see how much information users will reveal about themselves in different contexts. One interesting result from this area is known as the 'paradox of control': when users are given more control over their privacy settings, they tend to reveal more about themselves. It's not entirely known why people do this, but one hypothesis that is common in the community is that control over privacy settings creates a false sense of security: users naturally believe they have the ability to recall private information once it is 'out there'.

In [But+15], researchers consider the efficacy of robot teleoperation under different privacy-preserving filters. Robot operators are given tasks to perform by remotely controlling a robot, with the visual feedback obscured by different image filters. The privacy levels are evaluated by user survey questions.

A lot of non-statistical methods are very compelling because it is difficult to turn a person's emotions and decision-making around their privacy into a mathematical object. Researchers using these methods focus on ensuring that, when they evaluate privacy levels, it closely aligns with either a person's privacy assessment or a person's decision to reveal information. Grounding this with user studies ensures that this research maintains applicability to real-life applications.

There are other non-statistical methods that are more mathematical, as well. One commonly used in practice is k-anonymity. k-anonymity requires that every user be indistinguishable from at least $k-1$ other users [Swe02]. In Google Analytics for web traffic, they apply k-anonymity, as shown from their help page in Figure 3.1.

In summary, non-statistical methods are often used because of their real-world applicability. In general, the work in this area has far more user studies and experiments than those in differential privacy and inferential privacy. However, due to the nature of non-statistical methods, it is difficult to provide strong theoretical guarantees or generalize privacy results to many settings. Nevertheless, out of the three methods, these methods are the most impactful in today's technologies.

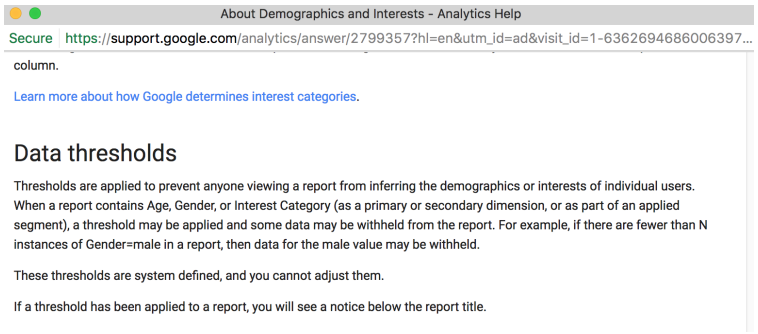

Figure 3.1: When too few site visitors are in a particular demographic, Google Analytics will threshold the displayed data for privacy purposes, following the concept of k-anonymity.

3.2.2 Differential privacy

Differential privacy is arguably the most popular privacy metric in the literature currently. It was introduced in [Dwo06] and has since developed into a very large body of literature. A good survey of the results is presented in [DR14], but even this book is not complete as the literature has grown so vast in scope.

Originally, differential privacy was imagined in the context of databases. Intuitively, it quantifies how much the observables of a database change as a result of a particular person participating or not participating in the database. If this change is small, then the system is private in the sense that not much information can be gleaned from the small change in database observables.

Differential privacy frames things in a cryptographic framework, and one of the main reasons differential privacy is incredibly popular is that it is both agnostic to attack models and the adversary's available side information. Since it only monitors the change in a database's observables, it does not need to make explicit assumptions on which inference algorithms an adversary will use or what side information can be leveraged.

Thus, differential privacy gets its name from two differentials: it monitors the difference in database observables when a user does or does not participate in a database, and it it monitors the difference in an adversary's knowledge as a result of the change in database observables.

The latter point is worth emphasizing as it does not garner enough focus in the literature. As originally presented in [Dwo06], if an adversary has a lot of prior knowledge about the user, then the adversary may be able to infer everything about the user. However, the differential privacy argument is that the privacy breach would essentially have occurred with or without the database.

In other words, differential privacy is a relative measure of privacy: it measures how much your privacy will change by participation in a system. It is *not* able to provide an absolute measure of privacy, at least in its original form.

We now formally define differential privacy. For concreteness and clarity of philosophical

intent, we will define it in its original context.

Definition 3. (Differential privacy) *Let Θ denote the set of possible database values. We are given a symmetric binary relation* Adj $\subset \Theta \times \Theta$ *known as the* adjacency *relationship. For $\theta, \theta' \in \Theta$, if* Adj$(\theta, \theta')$*, we say that θ and θ' are adjacent.*

A query *Y is a random mapping from Θ to some measurable space $(S, \mathcal{S})$, i.e. $Y(\theta)$ is a random element in S for every $\theta \in \Theta$. This query is ϵ* differentially private *if for any adjacent θ, θ' and measurable set $B \in \mathcal{S}$:*

$$\Pr(Y(\theta) \in B)) \leq \exp(\epsilon) \Pr(Y(\theta') \in B)$$

As an example, suppose Θ is a database of possible user heights. Let X denote the set of possible user IDs, and let $Y = \mathbb{R}$ denote possible height values. An entry in the database would then be of the form $(x, y) \in X \times Y$. If the database has n entries, then it is an element of $(X \times Y)^n$. If our database can have an arbitrary number of entries, then the set of possible databases would be $\Theta = \cup_{n \in \mathbb{N}} (X \times Y)^n$.

Now, we pick two databases $\theta, \theta' \in \Theta$. We enumerate $\theta = (x_1, y_1), \ldots, (x_N, y_N)$ and $\theta' = (x'_1, y'_1), \ldots, (x'_M, y'_M)$. We define θ and θ' to be adjacent if $N = M$ and there exists an i such that for all $j \neq i$, we have $(x_j, y_j) = (x'_j, y'_j)$. That is, θ and θ' have the same entries except for at row i. Intuitively, this captures the notion that we are considering how much the change in one person's database entry affects the overall database.

Finally, suppose we have a query Y which returns the average height, i.e. $Y(\theta) = \frac{1}{N} \sum_{i=1}^{N} y_i$. This cannot be ϵ differentially private for any $\epsilon > 0$ because there is no randomness. However, if we let $\tilde{Y}(\theta) = Y(\theta) + v$, for some Laplacian noise v, we can guarantee that this is ϵ differentially private for some $\epsilon > 0$ [Dwo06].

Note that the level of differential privacy is parameterized by ϵ.

Differential privacy provides strong guarantees and provides powerful results in composition as well. That is, if we take the output of an ϵ_1 differentially private mechanism and input that into an ϵ_2 differentially private mechanism, we have an $\epsilon_1 + \epsilon_2$ differentially private mechanism. This allows us to take simple differentially private mechanism and use them to construct more complicated mechanisms.

Generally, the results in differential privacy show that adding Laplacian or Gaussian noise in the right locations yields a differentially private or approximately differentially private mechanism. In practice, applications of differential privacy often involve finding a level of noise which guarantees a certain level of privacy, and hoping that the data is not too corrupted by the privacy-preserving noise. Loosely speaking, the level of noise required is inversely proportional to the number of users in the database which are being aggregated.

Optimally designing the noise for differential privacy is an open, difficult problem. Additionally, there are several contexts in which it is not clear how to implement differential privacy. For example, how does one consider the level of differential privacy in a dynamical system when the sampling rate is adjusted?

In summary, differential privacy is a compelling privacy metric that requires no explicit models of the adversary's available information or computing power. It has a literature with

useful results such as composition theorems and sample mechanisms. However, it generally requires the noise and the system to be of a particular form, and relies on a large number of users to provide meaningful bounds.

3.2.3 Inferential privacy

We refer to the third method for quantifying privacy as *inferential privacy*. Intuitively speaking, inferential privacy considers an adversary attempting to make an inference on a user's private data from observables. The metrics applied consider some statistical performance of these inferences. The work in this category can be unified by a general adversary model.

Adversary model

We start by defining the threat model, as presented in [PCF12a; Sal+13]. At a high level, an adversary observes Y, and wishes to infer private variables θ from Y. Adversaries are defined by their side information, as modeled by their beliefs on the joint distribution of θ and Y, and their cost function ℓ.

Let Θ denote the set of possible values for θ and similarly $\mathcal{Y}$ for Y. Also, for any set X, let $\mathcal{P}(X)$ denote the set of probability distributions across X. Finally, for a joint distribution $p(x, y)$, we will use $p(y|x)$ and $p(x)$ to denote the conditionals and marginals which arise out of $p(x, y)$[1].

Definition 4. (Adversary model) *An* adversary *is defined by a joint distribution* $p(\theta, Y)$ *and cost function* $\ell : \Theta \times \mathcal{P}(\Theta) \to \mathbb{R}$. *We will sometimes refer to the adversary as the tuple* (p, ℓ). *The* inference attack *that the adversary performs when observing* $Y = y$ *is the following optimization:*

$$L(y) = \min_{q \in \mathcal{P}(\Theta)} \sum_{\phi \in \Theta} p(\phi|y)\ell(\phi, q)$$

Note that this is a loss that is defined for every $y \in \mathcal{Y}$, and the optimization variable q is a distribution across Θ. This ℓ is a general notation for inference loss functions. As an example, if we take:

$$\ell(\phi, q) = \begin{cases} 1 & \text{if } \phi = \arg\max_z q(z) \\ 0 & \text{otherwise} \end{cases}$$

Then $L(y)$ is the probability that the maximum a posteriori estimate of θ is equal to the true θ when $Y = y$. As another example, we can take the logarithm loss $\ell(\phi, q) = -\log q(\phi)$, and this will recover equivocation metrics often used in the literature [San+11; Ven13].

[1] For ease of presentation, we will treat Θ and $\mathcal{Y}$ as finite sets in the sequel. However, all of this holds in general if we allow p to be absolutely continuous with respect to some underlying measure μ, and then replace each some with an integral with respect to the measure μ.

Thus, we can define the level of privacy against inferences. Without an observation of Y, the adversary's best belief about the private process is the solution to the minimization:

$$L_0^* = \min_{q \in \mathcal{P}(\Theta)} \sum_{\phi \in \Theta} p(\phi)\ell(\phi, q)$$

This will often be referred to as the *loss of the prior estimate*. After an observation of Y, the adversary will now have a loss of $L(Y)$, the *loss of the posterior estimate*. We can take the expectation of $L(Y)$ across Y to yield the *expected loss of the posterior estimate*:

$$\mathbb{E}_Y L(Y) = \sum_{y \in \mathcal{Y}} p(y)L(y)$$

Finally, we can talk about the average change in inference cost that the adversary experiences as a result of observing Y.

Definition 5. *The* expected total privacy loss *of θ from the observation Y is given by:*

$$\Delta L = L_0^* - \mathbb{E}_Y L(Y)$$

This is the difference in the inference cost of the prior estimate with the expected inference cost of the posterior.

We explore two specific examples of the loss function ℓ.

Log-loss ℓ

The log-loss is used as the loss function in recent work, e.g. [PCF12b], to compute the privacy loss, as it results in a natural measure of relevance: mutual information.

Proposition 3. *If an adversary (p, ℓ) uses the log-loss $\ell(\phi, q) = -\log q(\phi)$, then the total expected privacy loss of θ from Y is equivalent to the mutual information between θ and Y:*

$$\Delta L = I(\theta; Y) = \sum_{\phi \in \Theta} \sum_{y \in \mathcal{Y}} p(\phi, y) \log \left(\frac{p(\phi, y)}{p(\phi)p(y)} \right)$$

We will sometimes use mutual information as a measure of privacy loss. Mutual information is shown to be the *only* measure that satisfies the data processing axiom which is needed to properly define the benefit of side information in an inference problem [Jia+15].

We refer to the uses of the log-loss ℓ as information-theoretic metrics. These metrics lend themselves nicely to the design of noise in a fashion that is oftentimes optimal with respect to some criterion. In [PCF12a], the authors are able to design an optimal noising scheme subject to a performance constraint in database estimation, and in [Raj+11], the authors consider compression schemes in the context of the smart grid, and provide a theoretical bounds on the information leakage subject to a distortion constraint.

Point estimates

The formulation of the adversary in Definition 4 states that the adversary wishes to find a distribution across possible values of θ that minimizes some cost. As a special case, we can consider when the adversary only wishes to use a point estimate, i.e. he is only interested in a maximum a posteriori estimate. This can be recovered by taking: If we take:

$$\ell(\phi, q) = \begin{cases} 1 & \text{if } \phi = \arg\max_z q(z) \\ 0 & \text{otherwise} \end{cases} \tag{3.2.1}$$

Our work on the inferential privacy of point estimates builds on a hypothesis testing framework, which has been well studied in the information theory [CT91] and statistics [Kee10] communities. Variational calculus methods for statistics were first introduced by Neyman and Pearson [NP33], and have been a fruitful way to find optimal estimators. Additionally, a popular metric for critiquing the performance of an estimator is known as the minimax risk, which measures an estimator's expected loss against a worst-case distribution [LC73]; the minimax risk can act as a measure of the difficulty of a hypothesis testing problem. Alternatively, Fano was able to analyze the difficulty of the hypothesis testing problem by considering the entropy and mutual information between the parameter of interest and the observables [CT91; Yu97]; these results were extended to observations on the continuum in [HV94]. Each of these methods can provide a measure of the hypothesis testing problem's difficulty, which we use as a guarantee for privacy.

Recall that our adversary has access to the measured data signal, and also holds priors on the consumer's private information θ. He also knows how this private information affects the consumer's usage of IoT devices, $p_{y|\theta}$. Although this adversary has quite a bit of knowledge about the consumers, he does not hold arbitrary side information.

We note that it may not be realistic to suppose the adversary has access to p_θ and $p_{y|\theta}$. However, any adversary who tries to infer θ from y with less information will only do worse than our adversary model. Thus, this model provides a conservative estimate against all weaker adversary models.

Inferential privacy of point estimates

Our privacy metric is the probability of error if an adversary tries to infer the private variable θ.

Definition 6. *A system is '*α inferentially private*' if, for any estimator* $\widehat{\theta} : \mathcal{Y} \to \Theta$*, we have:*

$$\Pr(\widehat{\theta}(Y) \neq \theta) \geq \alpha \tag{3.2.2}$$

This estimator can be based on information in p_θ *and* $p_{y|\theta}$*.*

Here we note that this is in essence an ex-ante privacy metric, i.e. the privacy is spread across Θ according to p_θ. As often arises in many statistical estimation problems, an ex-

post privacy metric, i.e. a privacy metric that guarantees privacy for every type, is not a well-posed problem.

For example, suppose $\Theta = \{0, 1\}$, and consider the estimator $\widehat{\theta} \equiv 0$. For any consumer of type $\theta = 0$, the adversary will correctly infer their type with this estimator. In other words, an adversary can always violate the privacy of one type of consumer by making the blanket assumption that everyone is a fixed type. In a sense, we gain privacy by noting that the adversary has to be successful across the different types Θ (weighted according to p_θ).

Regardless of the algorithm the adversary uses, we can bound the probability it will successfully breach a consumer's privacy. Furthermore, this formula allows us to vary the quality level q, such as how often data is collected and transmitted. We will examine this on a concrete example in Section 3.4.2. This guarantee is also simple for consumers to interpret, and can be used in the design of privacy contracts between IoT service provider and consumers [Rat+16].

Proposition 4. *For an adversary (p, ℓ) where ℓ is defined as in (3.2.1), we have that $\Delta L = \inf\{\alpha : \text{ the system is } \alpha \text{ inferentially private}\}$.*

We will derive results that allow us to calculate values of α that satisfy the condition given in Definition 6. This section is a generalization of some of our previous work [Don+14], which considered this definition in the context of aggregate energy observations.

There are three methods by which we can derive lower bounds. Depending on the particular form of p_θ and $p_{y|\theta}$, some forms of the lower bound may be easier to calculate than others.

Likelihood-based theoretical bounds

Let $\Theta = \{1, \dots r\}$. We can define the maximum a posteriori (MAP) estimator $\widehat{\theta}_{MAP}$, which maximizes $\Pr(\widehat{\theta}(Y) = \theta)$.

Proposition 5. [CT91; Don+14] $\Pr(\widehat{\theta}(Y) = \theta)$ *is maximized by:*

$$\widehat{\theta}_{MAP}(Y) = \arg\max_{i \in \Theta}(p_\theta(i) \cdot p_{y|\theta}(Y|i)) \tag{3.2.3}$$

The proof of this proposition follows a variational calculus approach that was pioneered in [NP33].

The optimality of the MAP estimator with respect to the prior p_θ immediately leads to a guarantee of privacy.

Corollary 2. *The system is α inferentially private, where $\alpha = \Pr(\widehat{\theta}_{MAP}(Y) \neq \theta)$. Furthermore, the system is* not *α' inferentially private for any $\alpha' > \alpha$.*

Although this bound is optimal, it is often difficult to calculate. In these instances, some of the latter bounds may be used as a surrogate.

Le Cam's method

Le Cam's method is typically used in assessing minimax risk. More specifically, it is used to find a lower bound on the worst-case loss for an estimator. Here, worst-case means that the performance of the estimator is evaluated on the distribution for which the loss is maximized. For more details, we refer readers to [LC73; Yu97].

We present Le Cam's lemma in the context of our usage model. Again, let $\Theta = \{1, 2, \ldots, r\}$. First, we offer two definitions of distances between probability distributions.

Definition 7. *The* total variation distance *between two densities p and q on a measure space $(X, \mathcal{A}, \mu)$ is given by:*

$$\begin{aligned} \|p - q\|_{TV} &= \sup_{A \in \mathcal{A}} \left| \int_A p(x) - q(x)\mu(dx) \right| \\ &= \frac{1}{2} \int_X |p(x) - q(x)|\mu(dx) \end{aligned} \tag{3.2.4}$$

Definition 8. *The* Kullback-Leibler (KL) divergence *between two densities p and q on a measure space $(X, \mathcal{A}, \mu)$ is given by:*

$$D_{KL}(p\|q) = \int p(x) \log \frac{p(x)}{q(x)} \mu(dx) \tag{3.2.5}$$

Similarly, we will define the KL divergence between two random variables X and Y to be the KL divergence between their densities, and it will be denoted $D_{KL}(X\|Y)$.

Using Definitions 2 and 3, we now state Le Cam's Lemma.

Proposition 6. (Le Cam's lemma [LC73; Yu97]) *For any estimator $\widehat{\theta} : \mathcal{Y} \to \Theta$ and any distinct $i, j \in \Theta$, we have:*

$$\Pr(\widehat{\theta}(Y) \neq \theta \mid \theta = i) + \Pr(\widehat{\theta}(Y) \neq \theta \mid \theta = j) \geq \quad 1 - \|p_{y|\theta}(\cdot|i) - p_{y|\theta}(\cdot|j)\|_{TV} \tag{3.2.6}$$

A quick corollary is a lower bound on the probability of error:

Corollary 3. $\Pr(\widehat{\theta} \neq \theta)$ *is bounded below by:*

$$\min(p_\theta(i), p_\theta(j)) \cdot (1 - \|p_{y|\theta}(\cdot|i) - p_{y|\theta}(\cdot|j)\|_{TV}) \tag{3.2.7}$$

In practice, it will suffice to find an over-approximation of the total variation distance. For example, we have Pinsker's inequality:

Proposition 7. (Pinsker's inequality [Tsy09]) *For any densities p and q:*

$$\|p - q\|_{TV} \leq \sqrt{\frac{1}{2} D_{KL}(p\|q)} \tag{3.2.8}$$

Thus, we can provide a guarantee of inferential privacy.

Proposition 8. *The system is α inferentially private, where:*

$$\alpha = \max_{i \neq j} \left[\min(p_\theta(i), p_\theta(j)) \cdot (1 - \|p_{y|\theta}(\cdot|i) - p_{y|\theta}(\cdot|j)\|_{TV}) \right] \tag{3.2.9}$$

Fano's method

Fano's inequality relates the probability of error for a hypothesis test to the entropy and mutual information between a parameter and its estimator. Traditionally a concept defined on discrete random variables, it has also been extended to handle continuous distributions [CT91; IH91; HV94; Yu97]. Here we will state Fano's inequality in the context of our usage model, where $\Theta = \{1, \ldots, r\}$.

Proposition 9. (Fano's inequality [Yu97]) *For any estimator $\widehat{\theta} : \mathcal{Y} \to \Theta$, the probability of error $P(\widehat{\theta}(Y) \neq \theta)$ is bounded below by:*

$$\frac{1}{\log(r-1)} \left[\log r - \frac{1}{r^2} \sum_{i,j} D_{KL} \left[p_{y|\theta}(\cdot|i) \middle\| p_{y|\theta}(\cdot|j) \right] - \log 2 \right] \tag{3.2.10}$$

Thus, we have the quick corollary:

Corollary 4. *The system is α inferentially private, where α is given by Equation 3.2.10.*

Closing remarks on inferential privacy

Inferential privacy is most applicable when there is an inference algorithm that is the basis of a privacy breach. It requires the strongest probabilistic assumptions, and, by explicitly providing an adversary model, oftentimes can be evaluating by bounding the performance of optimal inference. It is formulated in a fashion that allows us to leverage a very large literature in statistical estimation, which provides powerful results when the context allows one to make strong structural assumptions on the processes operating on private data.

3.3 Design paradigms for privacy

Once we have a method to quantify privacy that sensibly captures our notions of privacy in a target domain, we still have to face the question of how to design systems in a way that accounts for the sensing and transmission of private data.

In our research on privacy mechanisms in the design of systems, we found that it is helpful to categorize these mechanisms under three general categories: passive privacy analysis, active privacy mechanisms, and optimal privacy-by-design. We briefly outline each of the three, and then go into detailed examples in Section 3.4.

3.3.1 Passive privacy analysis

In passive privacy analysis, a privacy analyst takes an existing system and then applies privacy metrics to quantify the privacy of the system. The key feature of passive privacy is that the system already exists and is fixed: in this sense, privacy is not a part of the design process at all.

For example, we can use fundamental limits of energy disaggregation to provide inferential privacy guarantees for users, based on the power signatures of devices and the usage patterns of consumers [Don+14]. When applying these results, we do not modify device signatures or user behaviors; rather, we just quantify the level of privacy in the existing system.

Another example in the context of ground traffic monitoring is given in Section 3.4.1: sensors measuring congestion information may reveal information about the origins and destinations of people, based on how users make routing decisions. We take dynamics for these systems from the literature and quantify the level of differential privacy for users.

3.3.2 Active privacy mechanisms

In active privacy mechanisms, there is a design parameter which can be varied to affect privacy levels. Naturally, in all applications, there is some force that works opposite privacy: we refer to this as the utility of data. In other words, if privacy was our only concern, we would simply shut down all sensors and not record or transmit anything. Some countervailing objective causes us to sample this private data. However, in active privacy mechanisms, we can vary the quality and quantity of data collected to balance this tradeoff between the utility of data and the privacy of users.

As an example, we an application in the smart grid where the sampling frequency is the design parameter. We consider how direct load control programs perform at different sampling rates. Additionally, we can quantify the level of inferential privacy for users when their data is collected at different time resolutions. This is presented in Section 3.4.2.

3.3.3 Optimal privacy-by-design

In optimal privacy-by-design methods, we fix a performance metric and a privacy metric. Then, we design the system to optimize the privacy metric subject to a performance constraint, or vice versa. In active privacy mechanisms, we simply have a parameter we vary, and then analyze its effects. In contrast, optimal privacy-by-design optimizes parts of the system to maximize privacy, while maintaining control/estimation performance. These approaches work to decouple the parts of data which contain private information from the parts of data which contain useful data for the original objective.

As an example, building air control can be improved by using occupancy estimates, as occupancy has a significant influence temperature and carbon dioxide levels. However, this reveals information about the location traces of users inside the building. We design occupancy sensors to randomize across reported occupancy with distributions that minimize the change in control law, while maximizing the mutual information between the true location traces and the sensor reports. This is in Section 3.4.3.

3.3.4 Discussion on design paradigms

Naturally, from a pure privacy perspective, optimal privacy-by-design is the most desirable option. However, in practice, it requires detailed models of the system dynamics and objective, and can be computationally intensive to calculate. We have been developing the literature on optimal privacy-by-design, and it is currently in a developing state of research.

Active privacy mechanisms can be cheaper to implement and easier to analyze. These are the most common in practice: when Facebook and Apple announce that they are using differential privacy, this means they are adding noise to their databases, and are free to choose the variance of the noise to affect the privacy level of its users, as well as the estimation quality of their targeted advertising.

Passive privacy analysis is perhaps the most cost-effective analysis of privacy, and, from an economic perspective, this should be the first attempt at privacy: analyze the level of privacy in the existing system to see if it is sufficient. If passive privacy analysis reveals that the privacy levels are insufficient, then redesign should happen, but if passive privacy analysis reveals that the threat to privacy is minimal, IoT company can save a lot of expense by not redesigning the system.

Thus, in practice, the design paradigm used for privacy will vary based on the context: the system under question, the consumer's sentiments towards privacy, the operational objective of the devices, and so on. As privacy becomes a more integral part of the discussion on engineering and privacy, these three design paradigms provide a good framework for understanding how companies are interacting with the privacy of the data they handle.

3.4 Examples

In this section, we discuss some examples of the privacy paradigms outlined in Section 3.3. In particular, we cover three examples. In Section 3.4.1, we cover a passive privacy analysis example in the context of ground transportation systems. We use differential privacy as a metric and consider how the origins and destinations of drivers in the routing game are revealed through public congestion in formation through general classes of driver routing dynamics. In Section 3.4.2, we discuss an active privacy mechanism in the context of the smart grid. We look at direct load control programs that use thermostatically controlled loads to shift peak demands, and consider how different sampling frequencies affect both the control efficacy, as well as the ability to infer private parameters such as income from smart meter readings. In Section 3.4.3, we consider heating, ventilation, and air condition control of buildings. In particular, we consider how occupancy readings can improve control, as well as how occupancy readings allow for the inference of individual location traces. We design a sensor that obfuscates its readings to minimally affect control performance while maximizing the privacy of individuals.

3.4.1 Passive privacy analysis example: differential privacy of populations in routing games

As our ground transportation infrastructure modernizes, the large amount of data being measured, transmitted, and stored motivates an analysis of the privacy aspect of these emerging cyber-physical technologies. In this paper, we consider privacy in the routing game, where the origins and destinations of drivers are considered private. This is motivated by the fact that this spatiotemporal information can easily be used as the basis for inferences for a person's activities. More specifically, we consider the differential privacy of the mapping from the amount of flow for each origin-destination pair to the traffic flow measurements on each link of a traffic network. We use a stochastic online learning framework for the population dynamics, which is known to converge to the Nash equilibrium of the routing game. We analyze the sensitivity of this process and provide theoretical guarantees on the convergence rates as well as differential privacy values for these models. We confirm these with simulations on a small example.

This work appeared previously in [Don+15].

Introduction

With the decreasing cost and size of technologies, our ground transportation infrastructure is increasingly modernizing with new sensor systems, control algorithms, and actuation modalities. Although these technologies promise great gains in traffic performance, such as level of service or equity [Tra11], an unprecedented amount of data is being measured, transmitted, and stored, and an analysis of the privacy aspect of this emerging cyber-physical technology is needed.

There have been a multitude of privacy conceptions in the philosophical and legal literatures. From an engineering perspective, the most commonly used paradigms are *control over information* and *secrecy* [Sol02].

In the abstract, control over information generally requires transparency to the person about what data is being collected and stored, consent to the transmission of this data to any parties, and an ability to correct mistakes in the data. As an example of how this conception works in practice, control over information forms the foundation of the Federal Trade Commission's Fair Information Practices.

On the other hand, secrecy focuses on which new inferences can be made about a person due to the information contained in the data; in this paradigm, a privacy breach occurs when there is a revelation of information that was previously not known, and the person felt that the information was private.

Throughout this paper, our conception of privacy will focus on the secrecy paradigm. In other words, we will focus on what new inferences can be made from the data collected by sensors in ground traffic infrastructures.

In the context of traffic systems, we consider the case where the origin and destination are considered private. This is motivated by the fact that this spatiotemporal information

can easily used as the basis for inferences for a person's activities. For example, an executive at the carsharing company Uber claimed he could tell when its users were having an affair in a blog post [TK14].

More specifically, we consider the differential privacy of the mapping from population sizes, i.e. the amount of flow for each origin-destination pair, to the traffic flow measurements on each link of a traffic network.

A popular modeling assumption is that the traffic flow is *atomless*, i.e. a single vehicle cannot unilaterally affect the flows on links [San01; Rou07; Kri+15c]. This is designed to match our intuition that, under normal conditions, one vehicle does not contribute significantly to traffic.

However, this implies that, through our models, one vehicle has no effect on the traffic flow measurements. Thus, in this paper, we consider differential privacy with respect to population sizes: how much does traffic flow change when a non-negligible mass of vehicles switch origin-destination pairs?

This framework is applicable for when some aggregator wants to protect the privacy of several drivers. For example, Google can analyze how much it reveals about its users when it provides routes through Google Maps. Alternatively, companies can consider how much is revealed through their shipping patterns, since this detailed data can allow inferences about important business information, such as which consumer markets are being targeted, which companies are in the supply chain, and which locations have potential for future expansion.

To model the dynamics of the driver populations, we use an online learning model in which, at iteration t, each population chooses a distribution over its paths. The joint decision of all populations determines the flows over the edges of the network, which, in turn, determines the costs over paths. These costs are then revealed to the populations, and given this information, they can update their distributions. This online learning model has been applied to routing games in [Blu+06], where the authors show that any no-regret strategy is guaranteed to converge to an equilibrium. The same model is also used in [Kri+15b], where the authors show that if each population applies a mirror descent algorithm, the joint distribution converges to a Nash equilibrium.

Our contribution is an analysis of the differential privacy of the dynamics of the driver populations. In this article, we consider a stochastic version of the model in [Kri+15b], in which the populations only have access to a noisy measurement of the path costs. The presence of noise is essential in providing differential privacy, while still guaranteeing convergence to the equilibrium, using results from stochastic optimization [Jud+11; Kri+15a].

The rest of this section is organized as follows. First, we review the engineering literature on privacy and develop some of the theory of differential privacy. Next, we introduce the routing game in the context of privacy. Then, we provide a learning model based on stochastic mirror descent. Here, we present theory on convergence rates and analyze the differential privacy of the routing game. Finally, we present a numerical example and closing remarks.

Previous work

Motivated by changing technologies, there has been a lot of recent research considering the issue of privacy. In this section, we will try to summarize the mathematical results in this line of research most relevant to this paper, noting both that the field is too rich for a comprehensive literature review and that privacy is a complicated social phenomenon of which a mathematical model is only one facet.

From a mathematical perspective, there have been several definitions of privacy. We seek to quickly survey a few definitions.

There has been work in inferential privacy, which seeks to bound the probability an adversary with a fixed set of information can correctly infer a hidden parameter, and uses a hypothesis testing model [Don+14].

Additionally, there has been work in information-theoretic based definitions of privacy, which uses the mutual information between a private parameter and the publicly observable data [San+11; Sal+13] or the conditional entropy of a private parameter given the observables [Ven13].

Throughout this paper, we will focus on a definition of privacy first introduced in [Dwo06], called *differential privacy*. This definition was originally designed for databases taking values in a finite alphabet, but has since been extended to consider the output of optimization algorithms [Duc+12; Hsu+14; Hua+14a; Han+14] and dynamical systems [LNP14]. For a more detailed analysis of the interpretation of differential privacy, we refer the reader to [DR14].

Our work is closest to that in [Han+14], where the authors considered the differential privacy of constraint sets in the context of gradient descent. Additionally, the work in [Duc+12] is of relevance, as it provides several minimax bounds for stochastic mirror descent, considered in this paper.

Theory

In this section, we will formally define differential privacy, as well as present results needed in future sections.

First, let $(\Omega, \mathcal{A}, P)$ denote our underlying probability space. Also, let Θ be a set equipped with a symmetric binary relation Adj, called the *adjacency* relation. The set Θ contains the possible values for a private parameter. Intuitively, the adjacency relation indicates which values should be roughly indistinguishable from the observable data. Although we never consider distributions or measures on Θ, for brevity we will often treat Θ as a measurable space, where any subset of Θ is measurable.

Furthermore, let $(S, \mathcal{S})$ denote a measurable space and let $Y : \Theta \times \Omega \to S$ be a mapping such that $Y(\theta, \cdot)$ is measurable for every $\theta \in \Theta$. In other words, given θ, $Y(\theta, \cdot)$ is a random element in S. For shorthand, we will write Y_θ to represent $Y(\theta, \cdot)$.

We can now present the definition of differential privacy.

Definition 9. Differential privacy*: We say a measurable mapping $Y : \Theta \times \Omega \to S$ is (ϵ, δ)-differentially-private if for all measurable sets $B \in \mathcal{S}$ and any $\theta, \theta' \in \Theta$ such that* $\mathrm{Adj}(\theta, \theta')$*:*

$$P(Y_\theta \in B) \leq \exp(\epsilon) P(Y_{\theta'} \in B) + \delta \tag{3.4.1}$$

If $\delta = 0$, we will say this mapping is ϵ-differentially-private.

We note two consequences of this definition. The first lemma appears in [LNP14].

Lemma 1. [LNP14]*: If a mapping Y is (ϵ, δ)-differentially private then*

$$Eg(Y_\theta) \leq \exp(\epsilon) Eg(Y_{\theta'}) + \delta \tag{3.4.2}$$

holds for all bounded measurable real-valued functions g and all $\theta, \theta' \in \Theta$ such that $\mathrm{Adj}(\theta, \theta')$*.*

The second lemma allows us to use tail bounds when analyzing differential privacy in certain contexts as we will see a later section.

Lemma 2. *Fix some event E. Suppose $P(E) \geq 1 - \delta'$ and that, for all measurable sets $B \in \mathcal{S}$ and all $\theta, \theta' \in \Theta$ such that* $\mathrm{Adj}(\theta, \theta')$*:*

$$P(\{Y_\theta \in B\} \cap E) \leq \exp(\epsilon) P(\{Y_{\theta'} \in B\} \cap E) + \delta$$

Then, Y is $(\epsilon, \delta + \delta')$-differentially-private.

Proof. Fix any measurable set $B \in \mathcal{S}$ and any adjacent $\theta, \theta' \in \Theta$. Then:

$$\begin{aligned} &P(Y_\theta \in B) \\ &\quad = P(\{Y_\theta \in B\} \cap E) + P(\{Y_\theta \in B\} \cap E^c) \\ &\quad \leq \exp(\epsilon) P(\{Y_{\theta'} \in B\} \cap E) + \delta + \delta' \end{aligned}$$

As desired. □

A result we will use in future sections is how differentially private mappings can be composed.

Proposition 10. Adaptive composition*: Suppose $Y_1 : \Theta \times \Omega \to S_1$ is (ϵ_1, δ_1)-differentially-private and $Y_2 : \Theta \times S_1 \times \Omega \to S_2$ is a measurable mapping such that $Y_2(\cdot, s, \cdot)$ is (ϵ_2, δ_2)-differentially-private for each fixed $s \in S_1$. Then the mapping $(\theta, \omega) \mapsto (Y_1(\theta, \omega), Y_2(\theta, Y_1(\theta, \omega), \omega))$ is $(\epsilon_1 + \epsilon_2, \exp(\epsilon_2)\delta_1 + \delta_2)$-differentially-private.*

Proof. Pick any set $A \in \mathcal{S}_1 \times \mathcal{S}_2$. Let $\mu_1(\theta, \cdot)$ denote the distribution of $Y_1(\theta)$ and $\mu_2(\theta, s, \cdot)$ denote the distribution of $Y_2(\theta, s)$. Furthermore, for $y \in S_1$, let $A_y = \{y' \in S_2 : (y, y') \in A\}$

denote the slice of A with respect to the first coordinate. Then, for any θ, θ' such that $\text{Adj}(\theta, \theta')$:

$$\begin{aligned}
&P[(Y_1(\theta), Y_2(\theta, Y_1(\theta))) \in A] \\
&\quad = \int \mu_1(\theta, dy_1) P(Y_2(\theta, y_1) \in A_{y_1}) \\
&\quad \leq \int \mu_1(\theta, dy_1)[\exp(\epsilon_2) P(Y_2(\theta', y_1) \in A_{y_1}) + \delta_2] \\
&\quad = \exp(\epsilon_2) \int \mu_1(\theta, dy_1)[P(Y_2(\theta', y_1) \in A_{y_1})] + \delta_2
\end{aligned}$$

Let $g(y) = P(Y_2(\theta', y) \in A_y)$, and note both that g is a bounded, measurable function and $Eg(Y_1(\theta')) = P((Y_1(\theta'), Y_2(\theta', Y_1(\theta'))) \in A)$. So, invoking Lemma 1:

$$\begin{aligned}
&P[(Y_1(\theta), Y_2(\theta, Y_1(\theta))) \in A] \\
&\quad \leq \exp(\epsilon_2) \int \mu_1(\theta, dy_1)[P(Y_2(\theta', y_1) \in A_{y_1})] + \delta_2 \\
&\quad = \exp(\epsilon_2) Eg(Y_1(\theta)) + \delta_2 \\
&\quad \leq \exp(\epsilon_2)[\exp(\epsilon_1) Eg(Y_1(\theta')) + \delta_1] + \delta_2 \\
&\quad = \exp(\epsilon_1 + \epsilon_2) P((Y_1(\theta'), Y_2(\theta', Y_1(\theta'))) \in A) + \ldots \\
&\qquad \exp(\epsilon_2)\delta_1 + \delta_2
\end{aligned}$$

As desired. □

Additionally, we can induct on Proposition 10. For brevity, we will sometimes write $Y_t(\theta, Y_1(\theta), \ldots, Y_{t-1}(\theta), \cdot)$ simply as $Y_t(\theta)$.

Corollary 5. Repeated adaptive composition*: Suppose $Y_1 : \Theta \times \Omega \to S_1$ is (ϵ_1, δ_1)-differentially-private and $Y_t : \Theta \times S_1 \times \ldots S_{t-1} \times \Omega \to S_t$ is a measurable mapping such that $Y_t(\cdot, s_1, \ldots, s_{t-1}, \cdot)$ is (ϵ_t, δ_t)-differentially-private for each fixed $(s_1, \ldots, s_{t-1}) \in S_1 \times \cdots \times S_{t-1}$ and $1 < t \leq T$.*

Then, the mapping $(\theta, \omega) \mapsto (Y_1(\theta), Y_2(\theta), \ldots, Y_T(\theta))$ is $\left(\sum_{t=1}^T \epsilon_t, \sum_{t=1}^T \exp\left[\sum_{t'=t+1}^T \epsilon_{t'}\right] \delta_t\right)$-differentially-private.

Finally, we note that the Gaussian distribution guarantees differential privacy.

Definition 10. Sensitivity*: The ℓ_2 sensitivity of a function $f : \Theta \to \mathbb{R}$ is given by:*

$$\Delta_2 f = \sup_{\theta, \theta' \in \Theta : \text{Adj}(\theta, \theta')} \|f(\theta) - f(\theta')\|_2 \tag{3.4.3}$$

Definition 11. *The zero-mean Gaussian distribution on $\mathbb{R}$ with variance parameter σ^2, denoted $\text{Gauss}(\sigma^2)$, has the density*

$$y \mapsto \frac{1}{\sqrt{(2\pi\sigma^2)}} \exp\left(\frac{-|y|^2}{2\sigma^2}\right) \tag{3.4.4}$$

with respect to the Lebesgue measure.

Proposition 11. Gaussian mechanism [DR14]*: For $\epsilon \in (0,1)$, and $b^2 > 2\ln(1.25/\delta)$, the mapping $Y_\theta = f(\theta) + Z$, where $Z_i \overset{iid}{\sim} \text{Gauss}(\sigma^2)$ for some $\sigma \geq b\Delta_2 f/\epsilon$, is (ϵ, δ)-differentially-private.*

The routing game

The routing game is given by:

- a directed graph $G = (V, E)$,
- a set of non-decreasing, Lipschitz continuous edge cost functions $c_e : \mathbb{R}_+ \to \mathbb{R}_+$, $e \in E$,
- a finite set of origin-destination pairs $(o_i, d_i) \in V \times V$, indexed by $i \in \{1, \ldots, I\}$,
- and a finite set of populations P_k, indexed by $k \in \{1, \ldots, K\}$.

In a ground transportation setting, the nodes, i.e. elements in V, represent physical locations, and edges, i.e. elements in E, represent the roadways that connect two locations. The edge cost functions c_e correspond to the amount of time taken when traveling along an edge e, and the non-decreasing assumption corresponds to the physical intuition that congestion worsens travel time. Finally, each population represents some aggregator that manages flows for all origin-destination pairs, such as Google or Waze.

For a given origin-destination pair (o_i, d_i), let $\mathcal{P}_i$ be the set of simple paths connecting o_i to d_i, and let $M_i \in \mathbb{R}^{|E| \times |\mathcal{P}_i|}$ be the edge-path incidence matrix, defined as follows:

$$\forall (e,p) \in E \times \mathcal{P}_i,\ (M_i)_{e,p} = \begin{cases} 1 & \text{if } e \in p \\ 0 & \text{otherwise.} \end{cases} \tag{3.4.5}$$

A population P_k is given by a private vector $\theta_k \in \mathbb{R}_+^I$, which specifies, for each origin-destination pair (o_i, d_i), the total mass of traffic $(\theta_k)_i$ that belongs to this population, and that travels from o_i to d_i. We assume there is some upper bound on the total size of the populations. Furthermore, we will define an adjacency relationship between private vectors.

Assumption 3. *It is common knowledge that θ is bounded. That is, there exists an $A_\theta < \infty$ such that, for every population k, $\|\theta_k\|_\infty \leq A_\theta$, and each population and outside observers know this bound.*

Definition 12. *Two private parameters of populations $(\theta_k)_{k\in[K]}$ and $(\theta'_k)_{k\in[K]}$ are* adjacent *if there exists a k^* such that $\theta_k = \theta'_k$ for $k \neq k^*$ and:*

$$\|\theta_{k^*} - \theta'_{k^*}\|_\infty \leq c$$

Recall that the adjacency relationship provides defines which pairs of private parameters should be roughly indistinguishable. Here, c is a constant that will be determined by the populations, modeling the maximum amount that a single population can increase or decrease the flow in one origin-destination pair without having a significant effect on observable data.

The action set of population P_k is a distribution vector $x_k \in \Delta^{\mathcal{P}_1} \times \cdots \times \Delta^{\mathcal{P}_I}$, where

$$\Delta^{\mathcal{P}_i} = \left\{ m \in \mathbb{R}_+^{|\mathcal{P}_i|} : \sum_{p \in \mathcal{P}_i} m_p = 1 \right\}$$

is the set of probability distributions over $\mathcal{P}_i$. In other words, every population chooses, for each origin-destination pair (o_i, d_i), how to distribute its mass across the available paths $\mathcal{P}_i$. For notational convenience, we will write $(x_k)_{\mathcal{P}_i}$ to denote the sub-vector $((x_k)_p)_{p \in \mathcal{P}_i} \in \Delta^{\mathcal{P}_i}$, so that $x_k = ((x_k)_{\mathcal{P}_1}, \ldots, (x_k)_{\mathcal{P}_I})$.

The flow allocations of all populations $(x_k)_{k \in [K]}$ determine the edge flows, defined as follows: the flow on edge e is $\phi_e(x_1, \ldots, x_K) = \sum_{k=1}^{K} \sum_{i=1}^{I} (\theta_k)_i \sum_{p \in \mathcal{P}_i} (x_k)_p 1_{(e \in p)}$. The vector of edge flows can be written simply in terms of the incidence matrices:

$$\phi(x_1, \ldots, x_K) = \sum_{k=1}^{K} \sum_{i=1}^{I} (\theta_k)_i M_i (x_k)_{\mathcal{P}_i}$$

The edge flows and edge costs determine the path costs. That is, the cost on path $p \in \mathcal{P}_i$ is given by:

$$\ell_p(x_1, \ldots, x_K) = \sum_{e \in p} c_e(\phi_e(x_1, \ldots, x_K))$$

We will denote by $\ell_{\mathcal{P}_i}(x_1, \ldots, x_K)$ the vector $(\ell_p(x_1, \ldots, x_K))_{p \in \mathcal{P}_i}$, and $\ell = (\ell_{\mathcal{P}_1}, \ldots, \ell_{\mathcal{P}_I}) \in \mathbb{R}_+^{\mathcal{P}_1} \times \cdots \times \mathbb{R}_+^{\mathcal{P}_I}$.

Finally, the cost for population P_k under distributions $x_1, \ldots, x_K$ is

$$\sum_{i=1}^{I} (\theta_k)_i \sum_{p \in \mathcal{P}_i} ((x_k)_{\mathcal{P}_i})_p \ell_p(x_1, \ldots, x_K)$$

which we will denote, more concisely, as

$$\langle x_k, \ell(x_1, \ldots, x_K) \rangle_{\theta_k}$$

where we define the inner product as follows: for all $x, y \in \mathbb{R}^{\mathcal{P}_1} \times \cdots \times \mathbb{R}^{\mathcal{P}_I}$

$$\langle x, y \rangle_\theta = \sum_{i=1}^{I} \theta_i \sum_{p \in \mathcal{P}_i} x_p y_p \tag{3.4.6}$$

Nash equilibria and the Rosenthal potential function

Definition 13. *A collection of population distributions* $(x_k)_{k\in[K]}$ *is a* Nash equilibrium *(also called* Wardrop equilibrium *in the traffic literature), if for every* $k \in [K]$ *and every* $y \in \Delta^{\mathcal{P}_1} \times \cdots \times \Delta^{\mathcal{P}_I}$*:*

$$\langle x_k, \ell(x_1, \dots x_K)\rangle_{\theta_k} \leq \langle y, \ell(x_1, \dots x_K)\rangle_{\theta_k}$$

That is, no driver can improve their cost by unilaterally changing their path.

Next, we show that the set of Nash equilibria of the game are exactly the set of minimizers of the Rosenthal potential, defined as follows:

$$f(x_1, \dots, x_K) = \sum_{e\in E} \int_0^{\phi_e(x_1,\dots,x_K)} c_e(u)du$$

Proposition 12. *The Rosenthal potential is convex, and its gradient with respect to* x_k *is:*

$$\nabla_{x_k} f(x_1, \dots, x_K) = \sum_{i=1}^{I} (\theta_k)_i \ell_{\mathcal{P}_i}(x_1, \dots, x_K)$$

Corollary 6. *The set of Nash equilibria of the game is exactly the set of solutions of the following convex problem:*

$$\begin{array}{ll} \textit{minimize} & f(x_1, \dots, x_K) \\ \textit{subject to} & x_k \in \Delta^{\mathcal{P}_1} \times \cdots \times \Delta^{\mathcal{P}_I} \textit{ for all } k \in [K] \end{array} \tag{3.4.7}$$

Online learning model

We consider the following online learning model of the game: at each iteration $t \in \{1, 2, \dots, T\}$, every population P_k chooses a distribution vector $x_k^{(t)} \in \Delta^{\mathcal{P}_1} \times \cdots \times \Delta^{\mathcal{P}_I}$. The combined choice of all populations determines the path loss vector $\ell(x_1^{(t)}, \dots, x_K^{(t)})$, which we will denote simply by $\ell^{(t)}$. The loss of population k is then given by the inner product $\left\langle \ell^{(t)}, x_k^{(t)} \right\rangle_{\theta_k}$.

At the end of iteration t, a stochastic loss vector $\hat{\ell}^{(t)}$, is revealed to all populations. Intuitively, one can think of $\hat{\ell}^{(t)}$ as a noisy version of $\ell^{(t)}$. The precise assumptions on the process $(\hat{\ell}^{(t)})$ will be given in Assumption 7.

Population dynamics

Our population dynamics take the following form.

Assumption 4. *We assume that for each population* P_k*, the stochastic process* $(x_k^{(t)})$ *follows stochastic mirror descent dynamics, given in Algorithm 2.*

Algorithm 2 Stochastic mirror descent dynamics for population k, with initial distribution $x_k^{(0)}$, learning rates $(\eta_k^{(t)})$, and distance generating function ψ_k.

for $t \in \{0, \dots, T-1\}$ **do**
Observe $\hat{\ell}^{(t)}$
Update

$$x_k^{(t+1)} = \arg \min_{x_k \in \Delta^{\mathcal{P}_1} \times \cdots \times \Delta^{\mathcal{P}_I}} \left\langle \hat{\ell}^{(t)}, x_k \right\rangle_{\theta_k} + \frac{1}{\eta_k^{(t)}} D_{\psi_k}(x_k, x_k^{(t)})$$

end for

These dynamics correspond to a stochastic version of the dynamics used in [Kri+15b].

Here, ψ_k is a distance generating function defined and C^1 on $\Delta^{\mathcal{P}_1} \times \cdots \times \Delta^{\mathcal{P}_I}$, and D_{ψ_k} is the Bregman divergence induced by ψ_k, defined as follows:

$$D_{\psi_k}(x_k, y_k) = \psi(x_k) - \psi(y_k) - \langle \nabla \psi(y_k), x_k - y_k \rangle \tag{3.4.8}$$

Assumption 5. *For all k, ψ_k is strongly convex with respect to a reference norm $\|\cdot\|$. That is, there exists $\ell_{\psi_k} > 0$ such that for all $x_k, y_k \in \Delta^{\mathcal{P}_1} \times \cdots \times \Delta^{\mathcal{P}_I}$:*

$$D_{\psi_k}(x_k, y_k) \geq \frac{\ell_{\psi_k}}{2} \|x_k - y_k\|^2$$

See Chapter 11 in [CBL06] for an account of the properties of Bregman divergences. We will further assume that the norm $\|\cdot\|$ decomposes into a sum of norms defined on each of the simplexes.

Assumption 6. *The norm $\|\cdot\|$ on $\mathbb{R}^{\mathcal{P}_1} \times \cdots \times \mathbb{R}^{\mathcal{P}_I}$ can be decomposed as follows:*

$$\|x_k\| = \sum_{i \in I} \|(x_k)_{\mathcal{P}_i}\|$$

Mirror descent is a general class of first-order optimization methods, used extensively both in convex optimization [NY83] and online learning [CBL06; BCB12]. In particular, projected gradient descent and entropic descent (a.k.a. the Hedge algorithm) are instances of the mirror descent method, for the appropriate choices of the distance generating function (see, for example, [BT03]).

In our model, since each population is updating its distribution vector using mirror descent dynamics, we can write the joint update as follows

$$\begin{aligned}
&(x^{(t+1)}, \dots, x_K^{(t+1)}) \\
&= \arg\min_x \sum_k \left\langle \ell^{(t)}, x_k \right\rangle_{\theta_k} + \sum_k \frac{1}{\eta_k^{(t)}} D_{\psi_k}(x_k, x_k^{(t)}) \\
&= \arg\min_x \left\langle \nabla f(x^{(t)}), x \right\rangle + \sum_k \frac{1}{\eta_k^{(t)}} D_{\psi_k}(x_k, x_k^{(t)}) \\
&= \arg\min_x f(x^{(t)}) + \left\langle \nabla f(x^{(t)}), x - x^{(t)} \right\rangle + D^{(t)}(x, x^{(t)})
\end{aligned} \tag{3.4.9}$$

where the minimization is taken across x in $(\Delta^{\mathcal{P}_1} \times \cdots \times \Delta^{\mathcal{P}_I})^K$ and we used the expression of the gradient $\nabla f(x^{(t)})$, given in Proposition 12, and defined:

$$D^{(t)}(x, x^{(t)}) = \sum_k \frac{1}{\eta_k^{(t)}} D_{\psi_k}(x_k, x_k^{(t)})$$

The expression (3.4.9) can be interpreted as a local approximation of the potential function f: the first term $f(x^{(t)}) + \left\langle \nabla f(x^{(t)}), x - x^{(t)} \right\rangle$ is simply the linear approximation of f around $x^{(t)}$, and the second term $D^{(t)}(x, x^{(t)})$ is a strongly convex function which penalizes deviation from the previous iterate $x^{(t)}$. By this observation, one can think of the joint dynamics of all populations as implementing a stochastic mirror descent on the Rosenthal potential f.

Suboptimality bounds on stochastic mirror descent

We now review some guarantees of the stochastic mirror descent method. First, we need to make assumptions on the stochastic process $(\hat{\ell}^{(t)})$ and the distance generating functions ψ_k.

Assumption 7. *Throughout the paper, we will assume that:*

1. *For all t, $\hat{\ell}^{(t)}$ is unbiased, that is, $\mathbb{E}\left[\hat{\ell}^{(t)} | \mathcal{F}_{t-1}\right] = \ell^{(t)}$, where $(\mathcal{F}_t)$ is the natural filtration of the process $(\hat{\ell}^{(t)})$.*
2. *$\hat{\ell}^{(t)}$ is uniformly bounded in the squared dual norm, that is, there exists L such that for all t:*
$$\mathbb{E}\left[\|\ell^{(t)}\|_*^2\right] \leq L$$
where $\|\cdot\|_*$ is the dual norm defined as follows:
$$\|\ell\|_* = \sup_{\|x\| \leq 1} \langle x, \ell \rangle$$
3. *For all k, there exists D_k such that D_{ψ_k} is bounded on $\Delta^{\mathcal{P}_1} \times \cdots \times \Delta^{\mathcal{P}_I}$ by D_k.*

Proposition 13 (Theorem 4 in [Kri+15a])**.** *Suppose that each population P_k follows a stochastic mirror descent dynamics as in Algorithm 2, and suppose that the learning rates are given by $\eta_k^{(t)} = c_k t^{-\alpha_k}$ with $c_k > 0$ and $\alpha_k \in (0, 1)$. Then for all $t \geq 1$, it holds that:*

$$\mathbb{E}\left[f(x^{(t)})\right] - f^\star \leq \left(1 + \sum_{\tau=1}^{t} \frac{1}{\tau}\right) \sum_{k=1}^{K} \left(\frac{1}{t^{1-\alpha_k}} \frac{D_k}{c_k} + \frac{c_k L}{2\ell_{\psi_k}(1-\alpha_k)} \frac{1}{t^{\alpha_k}}\right)$$

In particular, the system converges to the set of Nash equilibria in expectation, in the sense that $\mathbb{E}\left[f(x^{(t)})\right] \to f^\star$ at the rate $\mathcal{O}(t^{-\bar{\alpha}} \log t)$ where $\bar{\alpha} = \min_k \min(\alpha_k, 1 - \alpha_k)$.

Sensitivity analysis of the stochastic mirror descent update

In this Section, we study the sensitivity of the stochastic process $\hat{\ell}^{(t)}(x^{(t)})$ to changes in the private parameter θ.

First, we consider how the flow allocations change due to a change in mass on some origin-destination pairs. In this case, we hold the observed loss vector $\hat{\ell}^{(t)}$ fixed and will invoke Corollary 5 afterward.

Proposition 14. *Fix a loss vector $\hat{\ell}^{(t)}$ and consider the stochastic mirror descent update for population P_k*

$$x_k^{(t+1)}(\theta_k) = \arg\min_{x_k} \left\langle \hat{\ell}^{(t)}, x_k \right\rangle_{\theta_k} + \frac{1}{\eta_k^{(t)}} D_{\psi_k}(x_k, x_k^{(t)})$$

where the minimization is taken across $\Delta^{\mathcal{P}_1} \times \cdots \times \Delta^{\mathcal{P}_I}$. Here, x^k is viewed as a function of the mass vector θ_k. Then for all $\theta_k, \theta_k' \in \mathbb{R}_+^I$:

$$\|x_k^{(t+1)}(\theta_k) - x_k^{(t+1)}(\theta_k')\| \leq \frac{\eta_k^{(t)} \|\hat{\ell}^{(t)}\|_*}{\ell_{\psi_k}} \|\theta_k - \theta_k'\|_\infty$$

Proof. The minimized function is differentiable on $\Delta^{\mathcal{P}_1} \times \cdots \times \Delta^{\mathcal{P}_I}$, and its gradient at x_k is given by:

$$((\theta_k)_i \hat{\ell}_{\mathcal{P}_i}^{(t)})_{i \in I} + \frac{1}{\eta_k^{(t)}} \left[\nabla \psi_k(x_k) - \nabla \psi_k(x_k^{(t)}) \right]$$

To simplify the following expressions, we will use the following notation:

- $x_k^{(t+1)}(\theta_k)$ is denoted $x_k^{(t+1)}$, and $x_k^{(t+1)}(\theta_k')$ is denoted $x'^{(t+1)}$.
- $g^{(t)}(\theta_k) = ((\theta_k)_i \hat{\ell}_{\mathcal{P}_i}^{(t)})_{i \in I}$
- $h_k^{(t)}(x_k) = \frac{1}{\eta_k^{(t)}} \left[\nabla \psi_k(x_k) - \nabla \psi_k(x_k^{(t)}) \right]$

Then, by first-order optimality, we must have for all $x_k \in \Delta^{\mathcal{P}_1} \times \cdots \times \Delta^{\mathcal{P}_I}$:

$$\left\langle g^{(t)}(\theta_k) + h_k^{(t)}(x_k^{(t+1)}), x_k - x_k^{(t+1)} \right\rangle \geq 0$$

In particular, for $x_k = x'^{(t+1)}$, we have:

$$\left\langle g^{(t)}(\theta_k) + h_k^{(t)}(x_k^{(t+1)}), x_k'^{(t+1)} - x_k^{(t+1)} \right\rangle \geq 0$$

Permuting the roles of θ_k and θ_k', and summing the resulting inequalities, we have:

$$\left\langle g^{(t)}(\theta_k) - g^{(t)}(\theta_k'), x_k'^{(t+1)} - x_k^{(t+1)} \right\rangle \geq \left\langle h_k^{(t)}(x_k'^{(t+1)}) - h_k^{(t)}(x_k^{(t+1)}), x_k'^{(t+1)} - x_k^{(t+1)} \right\rangle \quad (3.4.10)$$

Furthermore, we have by Cauchy-Schwartz:

$$\begin{aligned}&\left\langle g^{(t)}(\theta_k) - g^{(t)}(\theta'_k), {x'_k}^{(t+1)} - x_k^{(t+1)} \right\rangle \\ &\leq \|g^{(t)}(\theta_k) - g^{(t)}(\theta'_k)\|_* \|{x'_k}^{(t+1)} - x_k^{(t+1)}\|\end{aligned}$$

By strong convexity of ψ_k, we have:

$$\begin{aligned}&\left\langle h_k^{(t)}({x'_k}^{(t+1)}) - h_k^{(t)}(x_k^{(t+1)}), {x'_k}^{(t+1)} - x_k^{(t+1)} \right\rangle \\ &= \frac{1}{\eta_k^{(t)}} \left\langle \nabla\psi_k({x'_k}^{(t+1)}) - \nabla\psi_k(x_k^{(t+1)}), {x'_k}^{(t+1)} - x_k^{(t+1)} \right\rangle \\ &\geq \frac{\ell_{\psi_k}}{\eta_k^{(t)}} \|{x'_k}^{(t+1)} - x_k^{(t+1)}\|^2\end{aligned}$$

Combining these inequalities with (3.4.10), we have:

$$\|g^{(t)}(\theta_k) - g^{(t)}(\theta'_k)\|_* \|{x'_k}^{(t+1)} - x_k^{(t+1)}\| \geq \frac{\ell_{\psi_k}}{\eta_k^{(t)}} \|{x'_k}^{(t+1)} - x_k^{(t+1)}\|^2$$

After simplification, this yields:

$$\|{x'_k}^{(t+1)} - x_k^{(t+1)}\| \leq \frac{\eta_k^{(t)}}{\ell_{\psi_k}} \|g^{(t)}(\theta_k) - g^{(t)}(\theta'_k)\|_*$$

Finally, using the expression of $g^{(t)}(\theta_k) = ((\theta_k)_i \hat{\ell}^{(t)}_{\mathcal{P}_i})_{i \in I}$, we have:

$$\|g^{(t)}(\theta_k) - g^{(t)}(\theta'_k)\|_* \leq \sum_{i \in I} \|\hat{\ell}^{(t)}_{\mathcal{P}_i}\|_* |(\theta_k)_i - (\theta'_k)_i| \leq \|\hat{\ell}^{(t)}\|_* \|\theta_k - \theta'_k\|_\infty$$

which concludes the proof. □

We have bounded how much a change in the private parameter affects the distribution on paths. Now, we analyze how the flows are affected by changes in distribution.

We will use the notation $\phi(x; \theta)$, which makes the dependence of edge flows on the parameter θ explicit. Also, $x^{(t+1)}(\theta)$ will be shorthand for $(x_1^{(t+1)}(\theta_1), \ldots, x_K^{(t+1)}(\theta_K))$. Also, let $\|\cdot\|_a$ denote an arbitrary norm on the space of edge flows.

Lemma 3. *For any* $\mathrm{Adj}(\theta, \theta')$*, we have:*

$$\|\phi(x^{(t+1)}(\theta); \theta) - \phi(x^{(t+1)}(\theta'); \theta')\|_a \leq c A_x \left[A_\Delta + A_\theta \frac{\eta_k^{(t)} \|\hat{\ell}^{(t)}\|_*}{\ell_{\psi_k}} \right]$$

Here, A_θ *is as given in Assumption 3 and:*

$$\begin{aligned}A_x &= \sup_{\|x_k\| \leq 1} \left\| \sum_{i=1}^{I} M_i (x_k)_{\mathcal{P}_i} \right\|_a \\ A_\Delta &= \sup_{x_k \in \Delta^{\mathcal{P}_1} \times \cdots \times \Delta^{\mathcal{P}_I}} \|x_k\|\end{aligned}$$

Proof. Consider any $\mathrm{Adj}(\theta, \theta')$. Note that $x_k^{(t+1)}(\theta) = x_k^{(t+1)}(\theta')$ for any $k \neq k^*$, since with the loss vector given, the update for population k only depends on θ_k.

$$\begin{aligned}&\|\phi(x^{(t+1)}(\theta);\theta) - \phi(x^{(t+1)}(\theta');\theta')\|_a \\ &\quad\leq \|\phi(x^{(t+1)}(\theta);\theta) - \phi(x^{(t+1)}(\theta);\theta')\|_a + \dots \\ &\qquad \|\phi(x^{(t+1)}(\theta);\theta') - \phi(x^{(t+1)}(\theta');\theta')\|_a\end{aligned}$$

For the first term, since θ and θ' are adjacent:

$$\begin{aligned}&\|\phi(x^{(t+1)}(\theta);\theta) - \phi(x^{(t+1)}(\theta);\theta')\|_a \\ &= \left\|\sum_{i=1}^{I}(\theta_{k^*})_i M_i(x_{k^*}^{(t+1)}(\theta))_{\mathcal{P}_i} - \sum_{i=1}^{I}(\theta'_{k^*})_i M_i(x_{k^*}^{(t+1)}(\theta))_{\mathcal{P}_i}\right\|_a \\ &\leq \left\|\sum_{i=1}^{I}|(\theta_{k^*} - \theta'_{k^*})_i| M_i(x_{k^*}^{(t+1)}(\theta))_{\mathcal{P}_i}\right\|_a \\ &\leq c\left\|\sum_{i=1}^{I} M_i(x_{k^*}^{(t+1)}(\theta))_{\mathcal{P}_i}\right\|_a \leq cA_xA_\Delta\end{aligned}$$

For the second term, we invoke Proposition 14:

$$\begin{aligned}&\|\phi(x^{(t+1)}(\theta);\theta') - \phi(x^{(t+1)}(\theta');\theta')\|_a \\ &= \left\|\sum_{i=1}^{I}(\theta'_{k^*})_i M_i(x_{k^*}^{(t+1)}(\theta))_{\mathcal{P}_i} - \sum_{i=1}^{I}(\theta'_{k^*})_i M_i(x_{k^*}^{(t+1)}(\theta'))_{\mathcal{P}_i}\right\|_a \\ &= \left\|\sum_{i=1}^{I}(\theta'_{k^*})_i M_i\left[(x_{k^*}^{(t+1)}(\theta))_{\mathcal{P}_i} - (x_{k^*}^{(t+1)}(\theta'))_{\mathcal{P}_i}\right]\right\|_a \\ &\leq A_\theta\left\|\sum_{i=1}^{I} M_i\left[(x_{k^*}^{(t+1)}(\theta))_{\mathcal{P}_i} - (x_{k^*}^{(t+1)}(\theta'))_{\mathcal{P}_i}\right]\right\|_a \\ &\leq A_\theta A_x\left\|x_{k^*}^{(t+1)}(\theta) - x_{k^*}^{(t+1)}(\theta')\right\| \leq cA_\theta A_x\frac{\eta_{k^*}^{(t)}\|\hat{\ell}^{(t)}\|_*}{\ell_{\psi_{k^*}}}\end{aligned}$$

As desired. □

We have bounded the effect of a change in the private parameter on the flows. Thus, we can state the sensitivity of the loss vector at time $t+1$ due to a small differential in the private parameter θ, when the observed loss vector at time t is held fixed.

Theorem 4. Sensitivity of the loss function: *For any* $\mathrm{Adj}(\theta, \theta')$*:*

$$\begin{aligned}&\|\ell(x^{(t+1)}(\theta);\theta, x^{(t)}, \hat{\ell}^{(t)}) - \ell(x^{(t+1)}(\theta');\theta', x^{(t)}, \hat{\ell}^{(t)})\| \\ &\qquad\leq cA_\ell A_x\left[A_\Delta + A_\theta\frac{\max_{k\in[K]}(\eta_k^{(t)})\|\hat{\ell}^{(t)}\|_*}{\min_{k\in[K]}(\ell_{\psi_k})}\right]\end{aligned}$$

Here, A_x, A_Δ, and A_θ are as defined in Assumption 3 and Lemma 3, and A_ℓ denotes the Lipschitz constant of the function $\ell : \phi \mapsto \ell(\phi)$ with respect to the norm $\|\cdot\|_a$ on the domain and $\|\cdot\|$ on the codomain.

Note that the sensitivity of $\ell^{(t+1)}$ depends on t through the learning rate $\eta_k^{(t)}$.

Differential privacy of the routing game

We use results from the previous sections to give privacy guarantees on the routing game when the loss vectors are observed with Gaussian noise.

Also, recall that the Gaussian mechanism preserves (ϵ, δ) differential privacy, and the privacy values depend on the variance of the mechanism and the sensitivity of the function. At each iteration t, we suppose that the populations observe $\hat{\ell}^{(t)} = \ell(x^{(t)}) + Z_t$ where $(Z_t)_p \overset{iid}{\sim} \text{Gauss}(\sigma^2)$.

We offer a couple of different interpretations of this mechanism. The first is that the data collector adds Gaussian noise before releasing this data to the populations. For example, the Department of Transportation might choose to add noise before transmitting the measurements from inductive-loop detectors in the road for privacy purposes. The second interpretation is that each driver experiences a perturbed version of the nominal loss when driving along the road, and when a population aggregates these perturbations, they obey a central limit theorem and look roughly normal in distribution.

First, we observe that for each path p, since the loss function ℓ_p is continuous on the compact set $\left(\Delta^{\mathcal{P}_1} \times \cdots \times \Delta^{\mathcal{P}_I}\right)^K$, it is bounded. Therefore, there exists $M > 0$ such that for all $x \in \left(\Delta^{\mathcal{P}_1} \times \cdots \times \Delta^{\mathcal{P}_I}\right)^K$, $\|\ell(x)\|_\infty \leq M$.

Theorem 5. *After T iterations, the mapping $\theta \mapsto (\hat{\ell}^{(1)}, \ldots, \hat{\ell}^{(T)})$ is (ϵ, δ) differentially private, where:*

$$\epsilon = \sum_{t=1}^{T} \epsilon_t \qquad \delta = \sum_{t=1}^{T} \exp\left[\sum_{t'=t+1}^{T} \epsilon_{t'}\right] \delta_t + \delta'$$

Here, a is any positive constant and $\delta', \epsilon_t, \delta_t$ are any constants that satisfy the following constraints:

$$1 - \delta' = (1 - 2\exp(-a^2/2\sigma^2))^{T \sum_{i=1}^{I} |\mathcal{P}_i|}$$

$$\epsilon_t > \frac{cA_\ell A_x (2\ln(1.25/\delta_t))^{1/2}}{\sigma^2} \times \left[A_\Delta + A_\theta \frac{\max_{k\in[K]}(\eta_k^{(t)})(\sum_{i=1}^{I} |\mathcal{P}_i|)^{1/2}(M+a)}{\min_{k\in[K]}(\ell_{\psi_k})}\right]$$

A_x, A_Δ, A_θ, and A_ℓ are as defined in Assumption 3, Lemma 3, and Theorem 4.

Proof. We can invoke the Chernoff bound and see that $P(|(Z_t)_p| > a) \leq 2\exp(-a^2/2\sigma^2)$. It follows that the event $E = \{\|Z_t\|_\infty \leq a \text{ for all } t\}$ holds with at least probability $(1 - 2\exp(-a^2/2\sigma^2))^{T\sum_{i=1}^{I}|\mathcal{P}_i|}$. On E, we have that $\|\hat{\ell}^{(t)}\|_2 \leq (\sum_{i=1}^{I}|\mathcal{P}_i|)^{1/2}\|\hat{\ell}^{(t)}\|_\infty \leq (\sum_{i=1}^{I}|\mathcal{P}_i|)^{1/2}(M + a)$ a.s.

Invoking Theorem 4, we can see that, on E:

$$\Delta_2\ell^{(t+1)} \leq cA_\ell A_x\left[A_\Delta + A_\theta\frac{\max_{k\in[K]}(\eta_k^{(t)})(\sum_{i=1}^{I}|\mathcal{P}_i|)^{1/2}(M+a)}{\min_{k\in[K]}(\ell_{\psi_k})}\right]$$

Thus, invoking Proposition 11, Corollary 5, and Lemma 2 yields our desired result. □

Note that a can be chosen to be any positive constant, and, in effect, provides a trade-off between the ϵ and the δ parameters.

Numerical example

Consider the routing game played on the network in Figure 3.2, with the following populations:

1. Population P_1 has mass vector $\theta_1 = (1, 0)$, and follows stochastic mirror descent dynamics with learning rates $\mathcal{O}(t^{-.5})$.

2. Population P_2 has mass vector $\theta_2 = (.2, 1.2)$, and follows stochastic mirror descent dynamics with learning rates $\mathcal{O}(t^{-.2})$.

The losses are taken to be linear. The resulting path loss functions are bounded by $M = 2$. We simulate the game for $T = 200$ iterations, with Gaussian noise with standard deviation $\sigma \in \{.01, .1, .4\}$.

Figure 3.3 shows the values of the potential function for the different values of σ. The asymptotic rate is consistent with $\tilde{\mathcal{O}}(t^{-\min(\alpha_1,\alpha_2)}) = \tilde{\mathcal{O}}(t^{-.2})$ rate predicted by Proposition 13. The variance of the noise σ^2 significantly affects the value of the expected potential. The effect of σ can also be observed in Figure 3.4, which shows the path flows for both populations, for $\sigma \in \{.01, .4\}$. Besides the effect of the noise level, we also observe that because the learning rates of population P_2 have a slower decay rate, its updates are more aggressive, which is reflected in the trajectories of its path flows.

Additionally, we consider the differential privacy of these observable traffic flows. Applying Theorem 5, we plot the differential privacy values as a function of the number of iterations in Figure 3.5. Generally, we are able to mask a small amount of population flow, but should c grow too large, the bounds quickly become trivial, i.e. $\delta = 1$. Furthermore, this value at which we can no longer meaningfully guarantee privacy can be thought of as the rate at which populations must shift origin-destination pairs to retain some level of privacy guarantee.

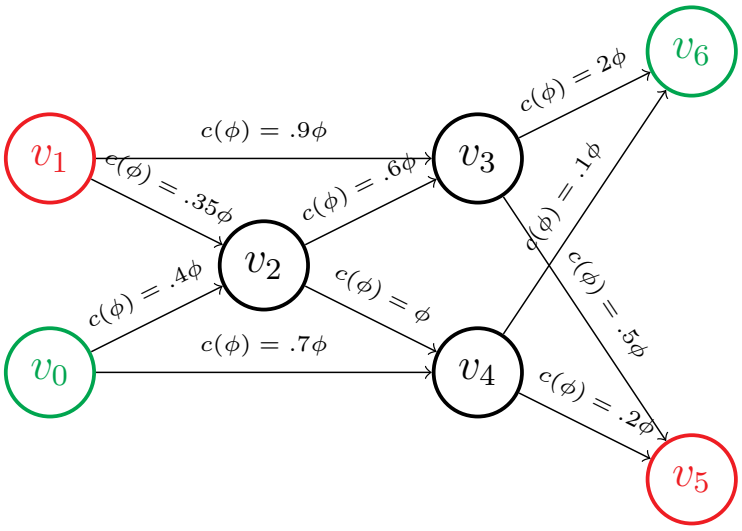

Figure 3.2: Example network with two origin-destination pairs: (v_0, v_6) and (v_1, v_5).

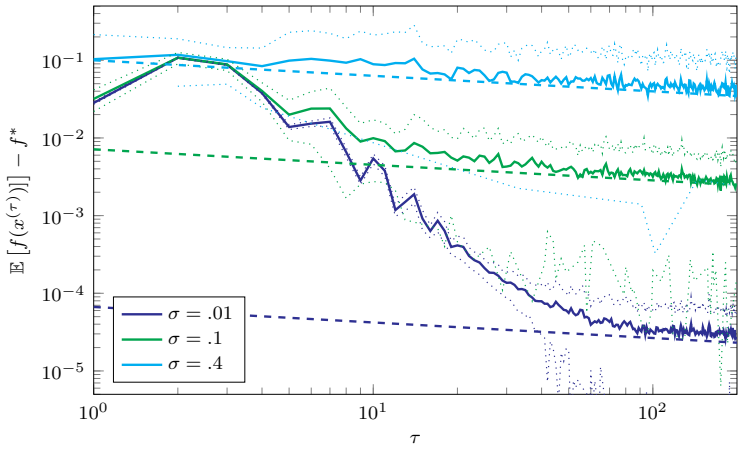

Figure 3.3: Potential function values $f(x^{(\tau)})$ as a function of the iteration τ, for different values of σ. The solid and dotted lines show, respectively, the average and the standard deviation over 150 runs of the simulation. The dashed lines show the $\tilde{\mathcal{O}}(t^{-.2})$ asymptotic rate predicted by Proposition 13.

Conclusion

We considered the privacy of the origins and destinations of drivers when the nominal traffic losses are observable with Gaussian noise. Considering a general online learning model based on stochastic mirror descent, and noting that the routing game is a potential game, we can think of the dynamics of drivers as optimizing the Rosenthal potential.

We analyzed the sensitivity of each update step as a function of the masses for each

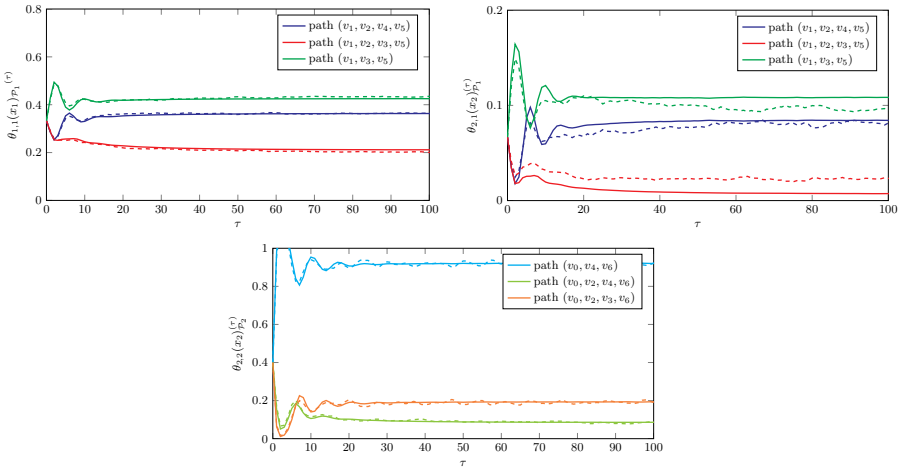

Figure 3.4: Path flows for each population, averaged over 150 runs, for $\sigma = .01$ (solid lines) and $\sigma = .4$ (dashed lines)

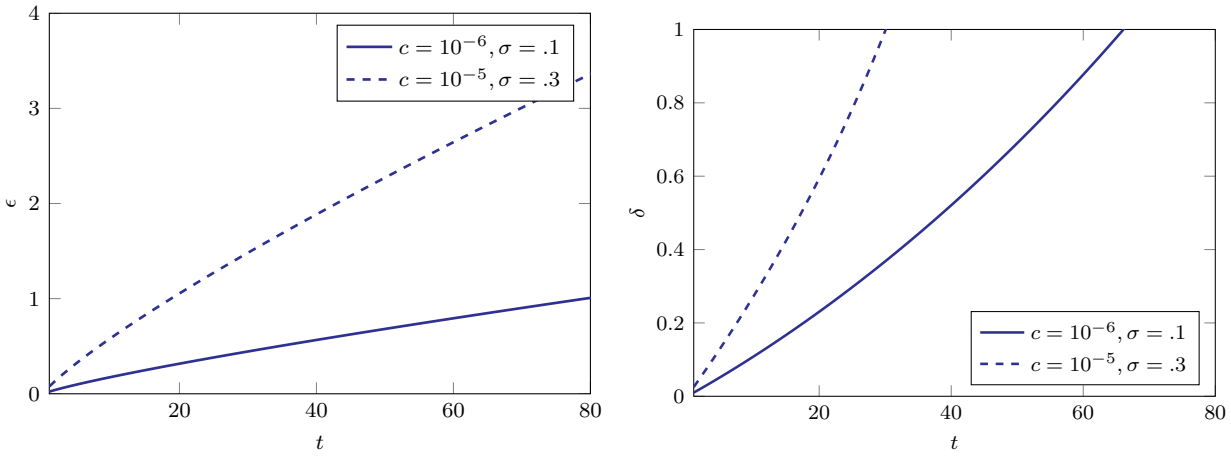

Figure 3.5: A plot of the values of ϵ, δ for which differential privacy holds, as a function of t, the number of iterations. Here, (c, σ) are taken to be $(10^{-6}, .1)$ then $(10^{-5}, .3)$, and a is taken to be 2. For larger values of c, the privacy guarantees are only meaningful for shorter periods of time.

origin-destination pair, which allowed us to bound the influence of this private information on the observable traffic losses. Additionally, we provided bounds on the convergence rates for different levels of noise, which provides insight into the relationship between how long it takes traffic flows to settle at equilibrium and how much is revealed by these observable traffic costs.

3.4.2 Active privacy mechanism example: the utility-privacy tradeoff in IoT at different sampling frequencies

In this section, we consider the tradeoff between the operational value of data collected in IoT and the privacy of consumers. We consider the active privacy mechanism paradigm in a control-theoretic context to quantify this tradeoff in IoT, and focus on a smart grid application for a proof of concept. In particular, we analyze the tradeoff between smart grid operations and how often data is collected by considering a realistic direct-load control example using thermostatically controlled loads, and we give simulation results to show how its performance degrades as the sampling frequency decreases. As a privacy metric, we use inferential privacy with point estimates. This privacy metric assumes a strong adversary model, and provides an upper bound on the adversary's ability to infer the value of a private parameter, independent of the algorithm he uses. Combining these two results allows us to directly consider the tradeoff between better operational performance and consumer privacy.

To successfully understand the utility-privacy tradeoff in these smart grid operations, we must quantify two things. First, we must model the tradeoff between the quality of collected data and performance of smart grid operations. Second, we must understand how data quality affects an adversary's ability to infer a consumer's private information. More generally, in IoT, we'll have to quantify this utility-privacy tradeoff for any data collected to design privacy-aware IoT service models.

The underlying philosophy of our work is that these data transmission policies often unintentionally transmit information about private parameters unrelated to the original control goal: we separate operational parameters from parameters users may consider 'private'. Furthermore, the operational goals of a systems operator are different from the inferential goals of a privacy-breaching adversary. Thus, different types of analyses are needed to understand the tradeoff between data collection and smart grid performance versus the tradeoff between data collection and user privacy.

For example, to quantify how much data is needed for smart grid operations, we consider how the performance of proposed direct load control (DLC) mechanisms change as fewer and fewer measurements are received by the controller. To quantify the privacy risk in these mechanisms, we utilize results in nonintrusive load monitoring (NILM) [Don+14] to give theoretical guarantees on when NILM algorithms will fail: adversaries will not be able to infer the device usage of a consumer from observing the aggregate power consumption of a building. Additionally, we model the private parameters of a consumer, and the inferences that can be made about private parameters from device usage patterns.

Once this analysis is done, we can apply our framework for understanding the utility-privacy tradeoff in IoT. We note that quantifying the operational value of data, picking a useful privacy metric, and applying the chosen metric to the appropriate models to provide a meaningful guarantee of privacy are all highly context-dependent actions. The hopes of a one-size-fits-all privacy solution is likely impossible due not only to formal properties of different IoT technologies, but also social aspects of the system and what is considered 'private' in the technology's application domain. However, we maintain hope for a general

framework by which to begin to devise and analyze privacy solutions.

The contributions of my research presented in this section are as follows.

1. We introduce a control-theoretic framework for quantifying the utility-privacy tradeoff in the Internet of Things. That is, we consider how modifying the *quality* of data, e.g. data with different levels of additive noise or sampled at lower frequencies, affects the operational value and the privacy levels of data.

2. We instantiate this framework in the context of the smart grid, by analyzing the operational value of data as well as the privacy risks inherent in data for different sampling rates in the operation of a recently proposed direct load control method. We are able to calculate the utility-privacy tradeoff for these programs.

3. In the process of applying this framework, we extended a state-of-the-art direct load control mechanism to handle missing measurements.

Utility-privacy tradeoff framework

In this section, we introduce a framework for quantifying the tradeoff between the operational utility of data and the privacy levels of consumers.

Privacy-preserving mechanisms can be divided into two categories: mechanisms which control *access* to data, or mechanisms which vary the *quality* of data.

Access control methods have been researched primarily by the cryptography community, with very strong results [DH79a]. The former can provide strong guarantees of privacy against outside adversaries, but does not protect users from privacy breaches by those who have access to the data. For example, your utility company should have access to your energy consumption, but they may be able to infer aspects of your lifestyle from these patterns [Lis+10; Don+13b]. In the Internet of Things, Google can receive your Nest sensor readings, and Fitbit may receive your GPS and step data, but unless there is an operational justification for the collection of data, it is likely that users will find this data collection invasive and unnecessary.

In contrast, quality-based methods have been researched by several communities. For example, most differential privacy mechanisms add noise to the data [Dwo06; DR14]: as the noise levels increase, the quality of the data decreases and privacy levels increase as well. As another example, systems can sample real-time data less frequently to increase the privacy levels of consumers; these mechanisms are considered in [C+12; Gir+14]. By modifying the quality of the data prior to its transmission, these methods guarantee privacy against both outside adversaries and insiders. However, the modifications to the data's quality must be carefully designed to not erode its original utility; if the data is no longer useful for its intended purposes, then the efficiency and comfort benefits of these novel technologies will be lost.

In this paper, we will focus on privacy-preserving mechanisms that vary the quality of data to achieve different levels of privacy.

The utility of data

In the Internet of Things, the utility of a particular set of data comes from the improvement in the performance of some service due to said data. To model these systems, we follow a control theoretic framework.

We are interested in the performance of our system at some set of times $T \subset \mathbb{R}_+$. This includes the discrete time cases $T = \{0, 1, \dots, N\}$ for some $N \in \mathbb{N}$ or $T = \{0, 1, \dots\}$, as well as the continuous time cases $T = [0, T_f]$ for some $T_f \in \mathbb{R}$ or $T = [0, \infty)$. For simplicity, we will assume operation of the system begins at time $t = 0$ and $0 \in T$.

Our system has some state space $\mathcal{X}$, which represents all possible configurations of the system at one point in time. Usually, we will take $\mathcal{X} = \mathbb{R}^n$ for some $n \in \mathbb{N}$. We will denote the state at time $t \in T$ as $x(t)$. Similarly, the control actions we can take upon the system live in some input space, $\mathcal{U}$, with the input at time t denoted $u(t)$. The dynamics of the system are captured in a function $\phi : \mathcal{X} \times \mathcal{U}^T \to \mathcal{X}^T$ which takes an initial condition and an input signal across all T and specifies which trajectory in $\mathcal{X}^T$ the system will follow. For example, in the context of linear time-invariant systems, if $\phi(x_0, u) = x$, then x is the unique solution to differential equation $\dot{x}(t) = Ax(t) + Bu(t)$ with initial condition $x(0) = x_0$.

The performance of the system is evaluated with respect to a cost function $J : \mathcal{X}^T \times \mathcal{U}^T \to \mathbb{R}$. The system has an initial condition $x_0 \in \mathcal{X}$ and obeys the system dynamics ϕ. The system operator wants to pick a $u \in \mathcal{U}^T$ such that $J(\phi(x_0, u), u)$ is kept low. Ideally, the optimal control problem would be solved: $\min_u J(\phi(x_0, u), u)$. However, this often is difficult and, in practice, we will use controllers that will approximate the optimal control strategy subject to information and tractability constraints.

To attempt to minimize this cost, the system operators will design a controller. This controller will determine the input $u \in \mathcal{U}^T$ that will be given to the system. However, this controller will have a limited amount of data about the system. In our framework, we will consider how variations in the quality of the data affect the system operator's control decisions, and therefore affect the realized cost of the system. For some quality level q and time $t \in T$, we will let $Y(q, t)$ denote the data available to the controller at time t.[2] With this data, the controller will pick a control input $u(t) \in \mathcal{U}$. We let this process be denoted $u_c(Y(q, t), t) \in \mathcal{U}$.

With this controller specified, we can consider the mapping from quality level q to realized cost J. That is, for a particular quality q, the controller will use the controller and issue control command $u_c(Y(q, t), t) \in \mathcal{U}$ at each time $t \in T$. This will cause the realized cost to be $J(\phi(x_0, u_q), u_q)$ where $u_q \in \mathcal{U}^T$ is defined as $u_q(t) = u_c(Y(q, t), t) \in \mathcal{U}$ for every $t \in T$.

Abstractly, this allows us to quantify the utility of data by showing how the control performance of the cyber-physical system erodes for different quality levels of data. As

[2]For generality, we have not included details of what space these objects q and $Y(q, t)$ live in. Formally, q can live in a general space, but we will often think of $q \in \mathbb{R}$. For example, q can denote the sampling period of our system, as we will explore in Section 3.4.2. Similarly, $Y(q, t)$ can live in some arbitrary space for each q and t. In the example in Section 3.4.2, $Y(q, t)$ will be a collection of random variables that the controller can observe at time t.

previously mentioned, we will instantiate this in a concrete example in this section.

The privacy of data

Data is collected from consumers with the intent of improving IoT operations. However, this data also allows the inference of private information about consumers, unrelated to IoT operations. This section quantifies how much information about the private lives of consumers is contained in data.

In the previous section, we fixed a set of time indices $T \subset \mathbb{R}_+$, and defined a data mechanism $Y(q,t)$ for each quality level q and time t. This data mechanism defined what information is collected and transmitted, and we quantified how a controller's performance changes as the quality level q is varied. In this section, we will consider how variations in q affect the privacy levels of consumers in the data mechanism $Y(q,t)$. We take a statistical perspective on privacy: what is the inferential power of these new observations relative to some private parameter? Our model is as follows.

Users have a private parameter $\theta \in \Theta$, which they wish to protect. These private parameters θ live in a space Θ with some particular structure, which depends on the privacy metric in use. In differential privacy, the private parameter space Θ is equipped with an 'adjacency' relationship specifies which pairs $(\theta, \theta') \in \Theta \times \Theta$ which should be indistinguishable. For information theoretic metrics and the inferential privacy metric discussed in Section 3.2, θ is seen as a random variable taking finitely many values, i.e. Θ has finitely many elements and there exist a prior distribution P_θ for the random variable θ.

These privacy metrics should be general enough in definition to allow evaluation for any data mechanism Y under consideration. Additionally, it will depend on the quality q: so our privacy valuations be a function of the structure of our data mechanism, as well as the quality level. This will be denoted $m(Y,q)$. This framework is general enough to capture any quality-varying privacy-preserving mechanisms, and this generality is needed to be able to encompass the spectrum of possible privacy risks and information structures in IoT.

Utility-privacy tradeoff example: Direct load control

In this section, we instantiate our utility-privacy framework in a concrete context. Specifically, we consider the privacy of direct load control (DLC) programs in the smart grid.

DLC has been a promising future direction for the smart grid for a variety of reasons. By controlling loads which can be modified without much impact on consumer satisfaction, we can allay many costs by shifting loads from peak demand and compensating for real-time load imbalances. Additionally, as renewable energy penetration increases, the generation side of power is growing more uncertain and will require demand flexibility. In this section, we will consider the load imbalance signal as exogenous, and use a DLC scheme to try and compensate the imbalance.

Additionally, such DLC policies are being deployed today. For example, Pacific Gas & Electric deployed the SmartAC program in Spring 2007 [Ale+08]. Another provider of

demand response (DR) services has recruited over 1.25 million residential customers in DLC programs, and has deployed over 5 million DLC devices in the United States. In California, they have successfully curtailed over 25 MW of power consumption since 2007 [Cal13]. As these programs are being deployed on a large scale, it is important to consider the privacy aspects of these programs [Lis+10].

In this example, we consider different sampling rates as a method of varying the quality of data q. Our motivations for this are two-fold.

First, there are many cases where noise-free data is required, for practical, regulatory, performance, or economic reasons. For example, suppose random noise is added to your energy consumption signal before being transmitted to the utility company. A consequence of this mechanism is that the energy bill you receive will not be a deterministic function of your energy usage, but rather a random variable with a conditional dependence on your energy usage. Many consumers may be unhappy with this mechanism in which they may be billed for more energy than they used, and a lot of regulatory overhead would be necessary for a utility company to roll out such a mechanism, even in the face of statistical arguments that the effect of such a random mechanism is negligible in the long run.

Second, an analysis of the effect of sampling rates on operational performance is the first step in enacting the *data minimization* principle for dynamical systems. In the United States, the Obama Administration examined privacy issues in its June 2011 smart grid policy framework report [Oba11]. The report recommends that State and Federal regulators should consider, as a starting point, methods to ensure that consumers' detailed energy usage data are protected in a manner consistent with federal Fair Information Practice (FIP) Principles. One of the key principles is data minimization. This principle is consistent with the notion of privacy by design [Cav11].

Similarly, the FIP principle of data minimization appear in smart grid privacy recommendations by the National Institute of Standards and Technology [The], the North American Energy Standards Board [Nor], the Department of Energy [Dep], the Texas Legislature and Public Utility Commission [Pub], and the California Public Utilities Commission (CPUC) [Cal].

The NISTIR 7628 [The] expresses the data minimization principle in the smart grid context as:

> Limit the collection of data to only that necessary for Smart Grid operations, including planning and management, improving energy use and efficiency, account management, and billing.

All these recommendations and policy proposals have been broad in coverage by necessity, as regulators do not want to burden electric utilities with specific limits on what they can collect. However, electric utilities who want to follow these privacy recommendations do not have a sound reasoning principle to help them decide how much data is too little or too much. Our goal in this section is to start discussing scientifically sound principles that can help determine how much data to collect in order to achieve a certain level of functionality of the grid, and how much privacy is granted to consumers under this data collection policy.

By analyzing the effect that sampling rate has on Smart Grid operations, we can begin to quantify the utility of data, a necessary first step to enacting data minimization. Intuitively, there should be a sampling rate where higher sampling frequencies have a negligible effect on the system's performance. For example, this could be due to the ability of the controller to leverage this high frequency data, or the time scales of the system itself. Conversely, there should intuitively be a sampling rate that is so low that the system's performance is comparable to the performance should the controller receive no measurements at all. Finding these regimes of operation is the goal of the first half of our framework.

As mentioned previously, there are several approaches to preserve the privacy of a consumer participating in an advanced metering infrastructure (AMI), including adding noise to data, modifying how data is aggregated, and the duration of data retention [Kur+11; RD11; TG09; Li+10; AC11; San+13]. These quality-varying mechanisms are currently an active topic of research. We note that our work is complementary to these other privacy policies. Our analysis is meant to assist electric utilities in following privacy recommendations: we seek to determine how much data to collect and how often it should be collected. Once this is in place, encryption, anonymization and aggregation techniques can be employed in tandem.

We evaluate the performance of a widely studied DLC scheme as a function of the sampling rate. As we will later show, increasing the sampling period is a means of improving the privacy of consumers. In particular, we focus on a DLC application using thermostatically controlled loads (TCLs) to manage load imbalances.

This section is organized as follows. First, we outline the DLC model. Then, we will define a controller for this DLC model that corrects for load imbalances, and formally specify how its control actions vary with different data sampling frequencies. Next, we estimate the performance of this controller with simulations. Afterward, we consider how privacy levels vary with different data sampling frequencies. Finally, we invoke the framework outlined in Section 3.4.2 to quantify the utility-privacy tradeoff in this application.

DLC model

In this section, we consider one recently proposed DLC program for concreteness. We note that our contribution is a general framework for numerically analyzing the sensitivity of these DLC programs to different information collection policies. We consider this research to be complementary to other research in how parameters affect system performance [Lu12; LZ13].

TCLs, which are often heating, ventilation, and air conditioning (HVAC) systems for buildings, are a promising avenue for the implementation of DLC policies [Cal09; Per+12]. This is due to the fact that buildings have a thermal inertia and can, in essence, store energy. Moreover, power consumption can be deferred and shifted while resulting in an imperceptible change in temperature.

Thermostatically controlled load model

There are several TCL and DLC models in the literature, e.g. [Rui+09; Mou+13; Mat+13], and our analysis can easily be applied to any of these models. For concreteness, we consider the model presented in [Mat+13].

Let $\mathcal{I}$ denote the set of TCLs participating in a DLC program. We model the temperature evolution of each TCL $i \in \mathcal{I}$ as a discrete-time difference equation:

$$x_i(k+1) = a_i x_i(k) + (1-a_i)[T_{a,i}(k) - m_i(k)T_{g,i}] + \epsilon_i(k) \tag{3.4.11}$$

In the above equation, $x_i(k)$ is the internal temperature of TCL i at time k, $T_{a,i}$ is the ambient temperature around TCL i, m_i is the control signal of TCL i, and ϵ_i is a noise process[3]. The term $a_i = \exp(-h_B/(R_iC_i))$, where h_B is the base sampling period[4], R_i is the thermal resistance of TCL i, and C_i is the thermal capacitance of TCL i. The T_g term represents the temperature gain when a TCL is in the ON state, and $T_g = R_i P_{\text{trans},i}$, where $P_{\text{trans},i}$ is the energy transfer rate of TCL i. Let P_i denote the power consumed by TCL i when it is in the ON state.

The local control for TCL i is modeled by the variable m_i. We assume the local controller performs an ON/OFF hysteresis control based on its setpoint and deadband. For a cooling TCL, this is defined as:

$$m_i(k+1) = \begin{cases} 0 & \text{if } x_i(k+1) < T_{\text{set},i} - \delta_i/2 \\ 1 & \text{if } x_i(k+1) > T_{\text{set},i} + \delta_i/2 \\ m_i(k) & \text{otherwise} \end{cases} \tag{3.4.12}$$

In these equations, $T_{\text{set},i}$ and δ_i are the temperature setpoint and deadband of TCL i, respectively. If $m_i(k) = 1$, then we say that TCL i is in the ON state at time k, and similarly $m_i(k) = 0$ means that i is in the OFF state at k.

In the next few sections, we will assume that these local control signals can also be overridden by the direct load controller, replacing Equation 3.4.12. The controller will only intermittently has access to observations $(x_i(k), m_i(k))$, due to a privacy-aware sampling policy.

Direct load control objective

We consider DLC policies that attempt to compensate for load imbalances and defer demands from peak times by switching TCLs between the ON state and the OFF state. The marginal cost of peak loads and unexpected load imbalances is responsible for a large portion of the

[3] Our development focuses on air conditioning for notational simplicity, but similar statements can be made for heaters.

[4] Here, h_B denotes the time scale of the dynamics. Later on, we will introduce how often the direct load controller may receive fewer measurements to preserve privacy, and this subsampling period will be denoted h.

preventable costs in the electricity grid; for a more detailed treatment of the benefits and impact of a DLC policy which can shave demand, we refer the reader to [CH11].

Formally, we consider the load imbalance as an exogenous variable. In particular, the centralized DLC controller is given some desired power trajectory P_{des} for the TCLs[5]. The goal of the controller is to minimize the error between the actual power consumed by the TCLs and the signal P_{des}, i.e. it wishes to minimize $\sum_k \left|\sum_{i\in\mathcal{I}} P_i m_i(k) - P_{\text{des}}(k)\right|$.

Direct load control capabilities

To achieve the DLC objective, we assume the centralized DLC controller has the capability of telling TCLs to switch modes between ON and OFF when the temperature $x(k)$ is between $T_{\text{set},i} - \delta_i/2$ and $T_{\text{set},i} + \delta_i/2$. More explicitly, if the centralized DLC controller issues a command to a TCL to switch from OFF to ON, the TCL turns on its air conditioner earlier than it would have in the absence of a control command. This DLC command will override the local controller. We assume that the centralized DLC controller has no control authority when the temperature is outside of the deadband, with the local controller deterministically in the OFF state when $x(k) < T_{\text{set},i} - \delta_i/2$ and in the ON state when $x(k) > T_{\text{set},i} + \delta_i/2$.

Note that the control policy effectively tightens the deadband. In particular, this control policy maintains customer satisfaction in the sense that the effective deadband is never larger than the user-specified deadband.

Our model of a direct load controller is as follows. We assume the centralized DLC controller has access to the parameters $\beta = (a_i, T_{a,i}, T_{g,i}, T_{\text{set},i}, \delta_i, P_i)$ for each TCL $i \in \mathcal{I}$. In other words, the controller knows the dynamics of each TCL. However, it is only able to observe the signals $(x_i(k), m_i(k))$ for certain values of k, determined by the privacy-aware sampling policy. One of the contributions of this paper is the extension of a DLC controller to situations where measurements are intermittent.

For the rest of this section, we will assume a privacy-preserving sampling policy that considers subsampling rates. In other words, our sampling policy is parameterized by a subsampling period $h \in \mathbb{N}$, and at time k, the centralized controller has access to the measurements $(x(k), m(k))_{k\in T_k}$, where the set $T_k = \{hl : l \in \mathbb{N}, hl \leq k\}$ denotes the time indices in which measurements are available[6].

Direct load controller

In this section, we outline a DLC policy inspired by work in the recent literature [Cal09; Mat+13]. Our model of a direct load controller is as follows. First, the controller maintains an estimate of the thermal state of each TCL. Let $\hat{x}_i(k)$ and $\hat{m}_i(k)$ denote the estimates of $x_i(k)$ and $m_i(k)$, respectively.

[5]We consider this load imbalance signal exogenous. In future work, we hope to examine elements of generation, such as scheduling, and how it is influenced by these programs.

[6]For simplicity, we assume that either all the TCLs transmit their state information at time k or none of them do. More asynchronous transmissions can be handled with some additional notational baggage.

The estimator acts as follows:

$$\hat{x}_k(k) = \begin{cases} x_k(k) & \text{if } k \in T_k \\ a_i \hat{x}_k(k-1) + (1-a_i)[T_{a,i}(k-1) - \hat{m}_i(k-1)T_{g,i}] & \text{if } k \notin T_k \end{cases} \tag{3.4.13}$$

$$\hat{m}_i(k) = \begin{cases} m_i(k) & \text{if } k \in T_k \\ 0 & \text{if } k \notin T_k \text{ and } \hat{x}_i(k) < T_{\text{set},i} - \delta_i/2 \\ 1 & \text{if } k \notin T_k \text{ and } \hat{x}_i(k) > T_{\text{set},i} + \delta_i/2 \\ \hat{m}_i(k-1) & \text{otherwise} \end{cases} \tag{3.4.14}$$

At time k, the estimator uses the observation if it is available. If no measurement is available, it evolves the estimates according to the dynamics with known parameters β, under the assumption that $\epsilon_i(k) = 0$. Similarly, it supposes that a TCL does not switch states under the local controller, unless the estimate of the thermal state of the TCL leaves the deadband.

These estimates are used to issue control commands. Our controller takes a binning approach, as seen in recent research [Cal09; Mat+13]. Each TCL is assigned to a bin based on its thermal state relative to its deadband, and whether or not it is in the ON or OFF state.

More formally, let N_{bin} be a parameter of our centralized DLC controller. N_{bin} is an even number denoting the number of bins our controller uses. For the ON states, we assign $N_{\text{bin}}/2$ bins, and for the OFF states, we assign $N_{\text{bin}}/2$ bins. Then, for each $i \in \mathcal{I}$, we define the following functions. First, we define a normalizing function for each TCL $\varphi_i : \mathbb{R}_+ \to \mathbb{R}$ as:

$$\varphi_i(x) = [x - (T_{\text{set},i} - \delta_i/2)]/\delta_i \tag{3.4.15}$$

This function normalizes x so, if x is in the deadband, then $\varphi_i(x)$ lies in the interval $[0, 1]$, e.g. $\varphi_i(T_{\text{set},i} - \delta_i/2) = 0$ and $\varphi_i(T_{\text{set},i} + \delta_i/2) = 1$.

Next, define the function $\psi_i : \mathbb{R}_+ \to \{0, 1, 2, \ldots, N_{\text{bin}}/2\}$ as:

$$\psi_i(x) = \begin{cases} 1 & \text{if } 0 \le \varphi_i(x) < 1/(N_{\text{bin}}/2) \\ 2 & \text{if } 1/(N_{\text{bin}}/2) \le \varphi_i(x) < 2/(N_{\text{bin}}/2) \\ & \vdots \\ N_{\text{bin}}/2 & \text{if } 1 - 1/(N_{\text{bin}}/2) \le \varphi_i(x) < 1 \\ 0 & \text{otherwise} \end{cases} \tag{3.4.16}$$

This function evenly partitions the interval $[T_{\text{set},i} - \delta_i/2, T_{\text{set},i} + \delta_i/2)$ into $N_{\text{bin}}/2$ bins of length $\delta_i/(N_{\text{bin}}/2)$, and assigns 0 if x lies outside this interval. Bins are indexed by an integer in $\{1, 2, \ldots, N_{\text{bin}}/2\}$ and a state in $\{\text{ON}, \text{OFF}\}$. Thus, if the state estimate of TCL i at time k is $(\hat{x}_i(k), \hat{m}_i(k))$, it will be assigned to bin $(\psi_i(\hat{x}_i(k)), \hat{m}_i(k))$ at time k. The number of TCLs in bin (n, m) at time k is $\sum_{i \in \mathcal{I}} 1\{\psi_i(x_i(k)) = n \text{ and } m_i(k) = m\}$. The estimated number of TCLs in bin (n, m) at time k is $\sum_{i \in \mathcal{I}} 1\{\psi_i(\hat{x}_i(k)) = n \text{ and } \hat{m}_i(k) = m\}$.

Also, note that a TCL may not fall into any bin; this corresponds to when the TCL's thermal state is out of its deadband. Since we are considering deadband tightening strategies to maintain customer satisfaction, if a TCL is outside its deadband, we cannot issue control commands to it.

Based on its estimate of how many TCLs are in each bin, the controller issues a command to each bin, stating what fraction of the TCLs in each bin should switch states. Here, for simplicity, we assume that every TCL consumes the same amount of power when on, i.e. $P_i = P$ for all $i \in \mathcal{I}$.

More concretely, the controller switches TCLs at time k based on the mismatch between the estimated power consumed $\left(\sum_{i\in\mathcal{I}} \hat{m}_i(k)\right) P$ at time k and the desired power consumption $P_{\text{des}}(k)$ at time k. For example, suppose that it is time k. The estimated number of TCLs in the ON state is $\sum_{i\in\mathcal{I}} \hat{m}_i(k)$. If $\left(\sum_{i\in\mathcal{I}} \hat{m}_i(k)\right) P > P_{\text{des}}(k)$, then too many TCLs are on and our controller will issue a command to switch from ON to OFF to some TCLs. It will try to turn off $\lfloor P_{\text{des}}(k)/P \rfloor - \sum_{i\in\mathcal{I}} \hat{m}_i(k)$ TCLs.

To do so, it will issue a probability to each bin, based on how many TCLs are estimated to be in each bin. Since we prefer to switch TCLs that are likely to switch to an OFF state soon, we start by turning off items in bin (1,ON). If there are more than enough TCLs in bin (1,ON), we issue a fraction based on how many TCLs we wish to turn off and the estimated number in a bin. If there are not enough, we command every TCL in the bin to turn off, and move on to the next bin (2,ON). The algorithm is described in more detail in Algorithm 3.4.2.

An analogous process takes place if $\left(\sum_{i\in\mathcal{I}} \hat{m}_i(k)\right) P < P_{\text{des}}(k)$ and the controller must turn TCLs on. This algorithm would be the same, only the variable b in Algorithm 3.4.2 would be initialized with $N_{\text{bin}}/2$ and would decrement across iterations, and ON would be replaced with OFF.

At the level of an individual TCL, the TCL can calculate which bin it is in, based on its true state $(x_i(k), m_i(k))$ and its deadband. When its been receives a command c, it will switch states with probability c. Using a probability allows the centralized controller to issue commands without broadcasting individual TCL identities, and without explicit knowledge of which TCLs will switch. Additionally, a TCL can decide whether or not to switch entirely on its own, without coordination or communication with other members of its bin.

An example of this control algorithm is depicted in Figure 3.6. In the top figure, we see how the TCLs are divided into bins, with $N_{\text{bin}} = 6$. The number in each bin denotes how many TCLs are actually in the bin, the number in parentheses denotes the estimated number of TCLs in the bin. There are an estimated 495 TCLs on, so the estimated total power consumption of the TCLs is 1.2375 MW. Suppose, in an extreme case, we wish to decrease power consumption by 500 kW. Thus, we would have to turn off 200 TCLs. According to the estimate, if we tell every TCL in the (1,ON) bin (the top-left bin), 154 TCLs will turn off. Therefore we must tell 46 TCLs in the bin (2,ON) to turn off as well, where there is estimated to be 170 TCLs. Thus, the control command issued to the bin (1,ON) is 1, to bin (2, ON) is $46/170 = 0.27$, and to all other bins is 0. In the bottom figure, the TCLs actually in each bin switch from the ON state to the OFF state according to a Bernoulli coin flip,

Algorithm 3 The centralized DLC controller's algorithm at time k for issuing commands to bins to reduce power consumption. *Inputs*: the estimated states of each TCL: $(\hat{x}_i(k), \hat{m}_i(k))$, the desired power signal $P_{\text{des}}(k)$, and the power of individual TCLs P. *Outputs*: updated mode estimates $\hat{m}'_i(k)$. (Commands for time k are also issued to each bin.)

procedure IssueDLCCommands($(\hat{x}_i(k), \hat{m}_i(k)), P_{\text{des}}(k), P$)
  $N \leftarrow \lfloor P_{\text{des}}(k)/P \rfloor - \sum_{i \in \mathcal{I}} \hat{m}_i(k)$
  $b \leftarrow 1$ ▷ Initialize the number of TCLs to switch, N, and the bin number b.
  while $N > 0$ and $b \leq N_{\text{bin}}/2$ **do**
    $n \leftarrow \sum_{i \in \mathcal{I}} 1\{\psi_i(\hat{x}_i(k)) = b \text{ and } \hat{m}_i(k) = \text{ON}\}$ ▷ Calculate the estimated number of TCLs in bin (b, ON).
    if $n \geq N$ **then**
      $c \leftarrow N/n$
      $N \leftarrow 0$ ▷ There are enough TCLs. Switch as many as are needed.
    else
      $c \leftarrow 1$
      $N \leftarrow N - n$ ▷ There are not enough TCLs. Switch all of them.
    end if
    issueCommand$(c, (b, \text{ON}))$ ▷ Issue the calculated command to the bin (b, ON).
    for $i \in \mathcal{I}$ **do**
      if $\psi_i(\hat{x}_i(k)) = b$ and $\hat{m}_i(k) = \text{ON}$ **then**
        flip $\sim$ Bernoulli(c) ▷ Update the estimate by having TCL mode estimates switch as necessary.
        if flip $= 1$ **then**
          $\hat{m}'_i(k) \leftarrow 1 - \hat{m}_i(k)$
        else
          $\hat{m}'_i(k) \leftarrow \hat{m}_i(k)$
        end if
      end if
    end for
    $b \leftarrow b + 1$
  end while
  return $\hat{m}'_i(k)$
end procedure

with probability equal to the command issued, and the estimates are updated based on the expected number of TCL switches. The numbers inside the bin represent the actual number of TCLs in each bin after the switching is completed, and the estimated number of TCLs in each bin after the switching is completed.

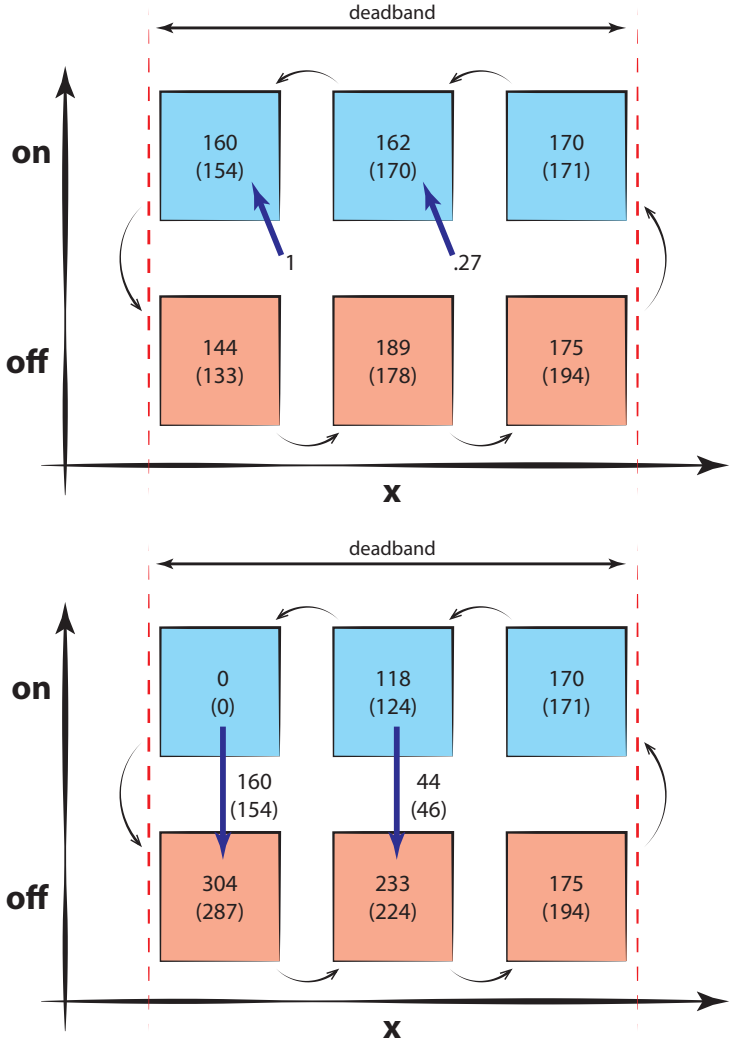

Figure 3.6: An example execution of Algorithm 3.4.2.

Closing the loop on this model development, we have the following model of the TCL with the control actuations by the centralized DLC controller. The closed loop dynamics are given by the following equation:

$$x_i(k+1) = a_i x_i(k) + (1-a_i)[T_{a,i}(k) - \tilde{m}_i(k) T_{g,i}] + \epsilon_i(k) \tag{3.4.17}$$

Here, the parameters are the same as in Equation 3.4.17. Note that the only difference in these dynamics and the open-loop dynamics without DLC is the modification of the $\tilde{m}_i(k)$ term. Furthermore, the mode of the TCL with DLC, $\tilde{m}_i(k)$ is given by:

$$\tilde{m}_i(k) = \begin{cases} 1 - m_i(k) & \text{with probability } c \\ m_i(k) & \text{with probability } 1 - c \end{cases} \tag{3.4.18}$$

$\tilde{m}_i(k)$ will depend on the local control law and the centralized DLC law described in Algorithm 3.4.2, with preference given to the centralized command. Here, $m_i(k)$ is the local control law as defined in Equation 3.4.12.

DLC model simulations

For simulations, we assume each TCL consumes $P_i = 2.5$ kW when in the ON state, and we consider a DLC controller in control of 1000 TCLs. Parameters for each TCL i are drawn independently, from distributions based on recent studies of a 250 m^2 home [Cal09; Mat+13]. The time step h_B was chosen to be $h_B = 1$ minute, and the number of bins $N_{\text{bin}} = 10$.

The ambient temperature $T_a = 32°$C for all TCLs[7], and the noise process $\epsilon_i(k)$ is independent across k and distributed according to a $N(0, 0.0005)$ distribution[8] for each k.

California Independent System Operator (CAISO) market signals are given in 5 minute intervals [Mat+13; Cal14], so for simulations, the signal P_{des} is independently drawn from a $U(875 \text{ kW}, 1.35 \text{ MW})$ distribution[9]. That is, $P_{\text{des}}(k)$ is uniformly drawn for $k \in \{0, 5, 10, \dots\}$. For other values of k, we take the linear interpolation.

Simulations of the aggregate power consumption of all the TCLs is shown in Figure 3.7 for the uncontrolled case, the case where $h = 1$ minute, and the case where $h = 30$ minutes. Comparing the top plot with the middle and bottom plots, we can see that a DLC policy can reduce the load imbalance even when the controller does not always receive measurements. However, small unforeseen temperature deviations can cause the controller's performance to degrade if enough measurements are not provided, as seen by comparing the middle and bottom plots.

Additionally, the thermal state of one TCL is shown in Figure 3.8. We can see that the temperature inside the TCL remains inside the deadband, resulting in no loss of comfort to the consumer, in all three cases.

In Figure 3.9, we plot the error between the actual power consumption and the desired load imbalance compensation signal. First, we randomly drew a P_{des} signal and TCL parameters. Then, for this fixed P_{des} signal and TCL parameters, we ran 500 trials for each

[7] For these simulations, we assumed that the ambient temperature is constant across one hour, which can be reasonable for this short time frame.

[8] This is the variance of the noise for one time step, so 0.0005 models the variance of temperature across $h_B = 1$ minute.

[9] This framework can handle other distributions for the load imbalance signal, but a uniform distribution was chosen as a non-informative prior [Kee10]. The parameters of the distribution were chosen as reasonable values for which energy consumption could be compensated. From simulations, we find that a larger interval is more difficult to track, as expected.

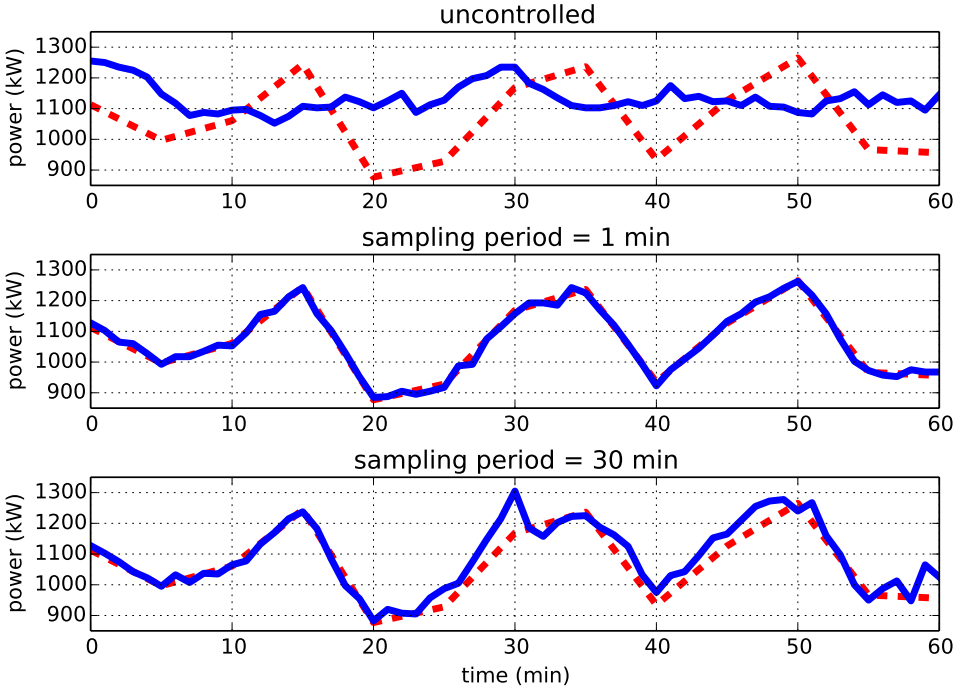

Figure 3.7: A sample simulation of the aggregate power consumption of 1000 TCLs. The solid blue line represents the actual power consumption, and the dotted red line represents the desired power consumption. The top figure shows the power consumption in the absence of any control commands, the middle figure shows the power consumption with a sampling period of $h = 1$ minute, and the bottom figure shows the power consumption with a sampling period of $h = 30$ minutes.

sampling period h, and we consider the empirical distribution of the difference between the actual power consumed by all the TCLs and the desired power signal: $\sum_{i \in \mathcal{I}} P_i m_i - P_{\text{des}}$. We used the ℓ_1 norm on the error signal, so, if we assume a fixed price for spot market electricity purchases/sales throughout the hour interval, this is directly proportional to the cost the utility company must pay.

DLC privacy analysis

In this example, we suppose households consider their income private. However, their income levels will affect their behaviors at home; in this paper, we focus on how their cooking behaviors change. To model this, we use data from the U.S. Energy Information Administration's 2009 Residential Energy Consumption Survey (RECS) [Ber09]. By observing these different cooking behaviors through a household's energy consumption, an adversary may

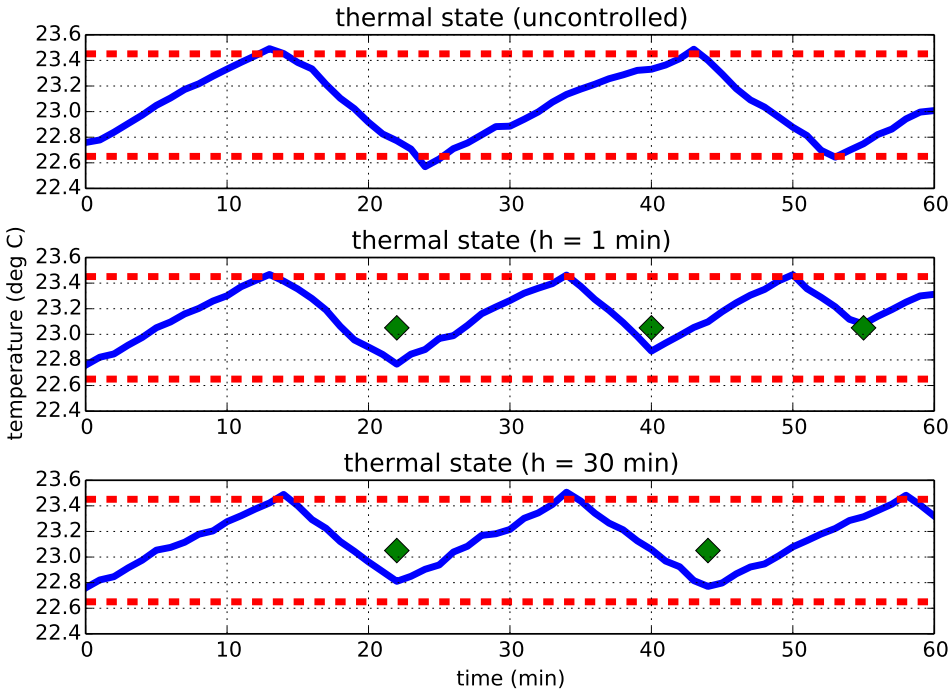

Figure 3.8: The thermal state of one sample TCL. The top graph shows the thermal states of the TCL when there is no control. The middle and bottom graphs show the thermal states based on a controller that receives observations every $h = 1$ minute and $h = 30$ minutes, respectively. The dotted red lines indicate the deadband limits. The diamonds indicate when the DLC policy issued control commands to the TCL.

infer the income of the household.

Formally, let $\Theta = \{\theta_L, \theta_M, \theta_U\}$ denote the private parameter corresponding to lower (less than \$20,000), middle (\$20,000 to \$59,999), and upper (\$60,000 or more) class incomes. Across 113.6 million U.S. homes, 23.7 million households are θ_L, 48.7 million are θ_M, and 41.2 million are θ_U [Ber09]. This will be our prior, P_θ.

Furthermore, we look at the overall energy consumption of each consumer type. This data is shown in Figure 3.10. For each type, we fit a log-normal distribution to the overall energy consumption. To estimate the location parameter μ and scale parameter σ, we used the unbiased, minimum variance estimators [Kee10, Chapter 4] on the log of the data[10]. We assume the scale parameter is the same for all three private parameters, and we can see that these distributions approximate the data quite well. We can see that a household's income

[10]Recall that the log-normal distribution, denoted $\ln N(\mu, \sigma)$, is defined by a location parameter μ and scale parameter σ, with density $x \mapsto \frac{1}{x\sigma\sqrt{2\pi}} \exp\left(\frac{(\ln x - \mu)^2}{2\sigma^2}\right)$ for $x > 0$.

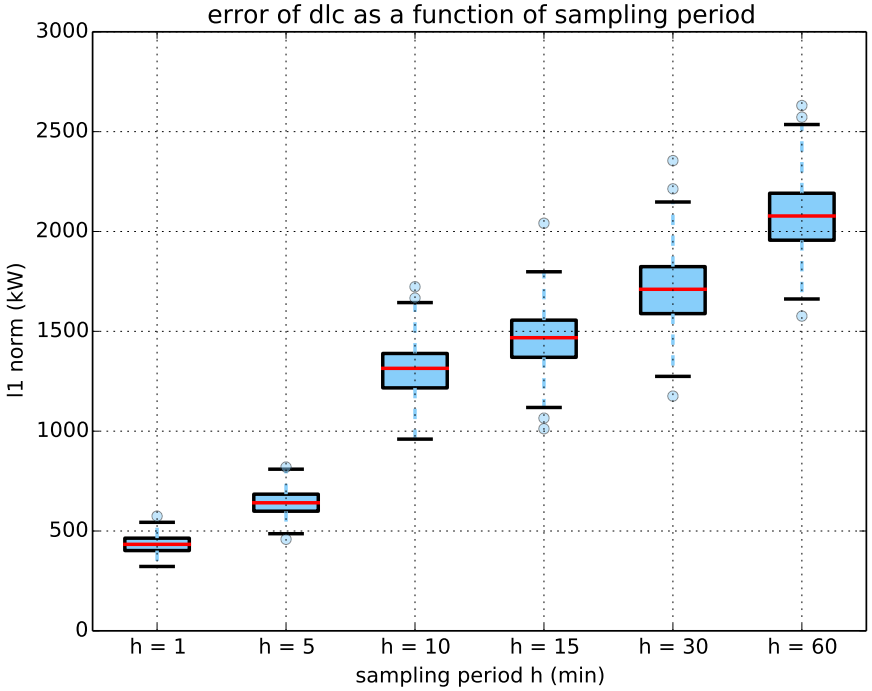

Figure 3.9: A plot of how the error between the actual power consumed by the TCLs and the desired power consumption signal empirically varies with the sampling period h. The value we are plotting is $\| \sum_{i \in \mathcal{I}} P_i m_i - P_{\text{des}} \|_1$. The whiskers indicate all data points within 1.5 times the interquartile range. For reference, the error after 500 simulations of uncontrolled TCLs has an empirical mean of 5.39 MW with a standard error of 302 kW.

level is correlated with the energy consumption.

Thus in this framework, θ determines the distribution across the observable energy consumption y, i.e. $P_{y|\theta}(\cdot \mid \theta)$ is a log-normal distribution. For example, $P_{y|\theta}(\cdot \mid \theta_L)$ is the $\ln N(\mu_L, \sigma^2)$ distribution. We assume that power consumption on smaller time scales is distributed similarly to this annual data, and these distributions are independent across time[11]. In other words, if a household consumed P kWh in a year, then we assume they consumed roughly $P/365$ kWh a day.

With this assumption, we can consider the distribution of energy consumption at different sampling rates. Note that, if we sample at high frequencies, we receive more measurements than in the low frequency case, but each measurement is less informative with regards to the consumer's income level[12].

[11]This assumption is likely valid for certain time scales, but will not hold in general. In future work, we hope to analyze the distributions of energy consumption data at different sampling rates.

[12]Here, we scale the data according to the time scale, and, as before, we used the uniform, minimum

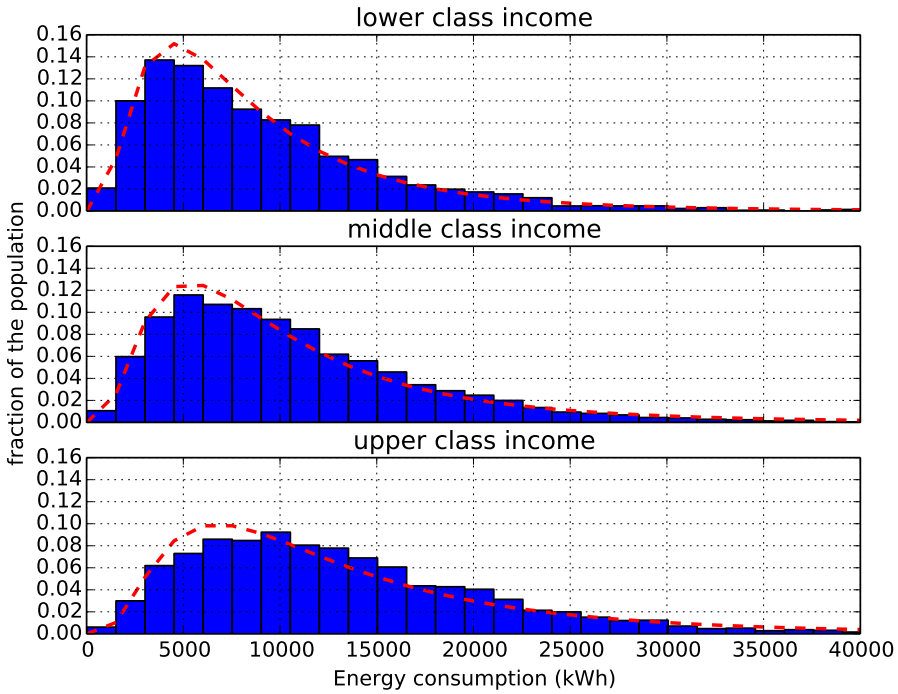

Figure 3.10: Histograms of the United States household total annual energy consumptions in each income level in 2009 [Ber09], corresponding to private parameters θ_L, θ_M, and θ_U. The data roughly follows a log-normal distribution. The location parameters are $\mu_L = 8.88$, $\mu_M = 9.06$, and $\mu_H = 9.31$, and we assumed all three distributions had the sample scale parameter, $\sigma = 0.49$. To model sampling, we assume that this data is representative of energy consumption on smaller time scales as well.

Since the scale parameters are the same for all 3 distributions, we can explicitly calculate the MAP using the theory of exponential families [Kee10, Chapter 2]. Then, using Proposition 2, we can calculate the probability an adversary can infer the private parameters, i.e. income level, from the AMI signals. This is represented in Figure 3.11.

We can see that very high frequency data provides little guarantees of privacy of income level, but this privacy level, α, quickly increases as the sampling period h increases. Furthermore, we can note relationships between time horizons, sampling rates, and privacy. For example, 1 hour of data sampled every 3 minutes is as informative as 6 hours of data sampled every 15 minutes.

Although we focus on a particular example here, this framework can be applied to more

variance estimators on the log of the data [Kee10, Chapter 4]. For example, if we receive measurements every minute, the location parameters for each measurement are $\mu_L = 0.014$, $\mu_M = 0.016$, and $\mu_H = 0.017$, whereas if we receive measurements hourly, the location parameters are $\mu_L = 0.82$, $\mu_M = 0.99$, and $\mu_H = 1.26$.

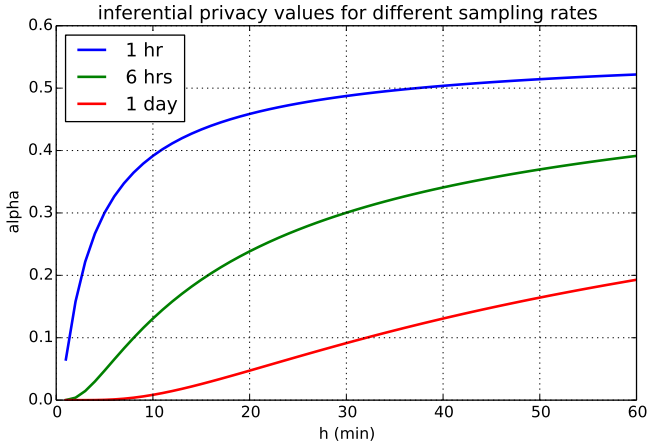

Figure 3.11: A plot of the inferential privacy value α, as a function of the sampling period h. Each line corresponds to a time horizon in which an adversary can receive samples. Intuitively, we would expect the privacy value α to decrease for longer horizons. Note that our framework accounts for the fact that with longer sampling periods, we receive fewer measurements, but each individual measurement is more informative.

detailed models, i.e. more informed adversaries, and other private parameters. For example, we consider the case where an adversary has knowledge of correlation across time and high frequency dynamics across time in [Don+14]. We are also currently examining the effects of sampling for longer time horizons [FC15].

Framework application

In Section 3.4.2, we outlined a formal model of using DLC of TCLs to correct for load imbalances. In Section 3.4.2, we defined the control policy of a direct load controller. Importantly, in our model we accounted for how varying sampling rates affect the performance. This required adapted a proposed control law for various sampling rates. In Section 3.4.2, we quantified the effect of various sampling rates on the operational efficiency of this load imbalance correction program. Additionally, in Section 3.4.2, we consider how energy consumption data can reveal income levels, and quantify this private information leakage using an inferential privacy metric.

Now, we place this in the context of the framework outlined in Section 3.4.2.

The set of time indices we care about is $T = \{0, 1, \dots, N\} \subset \mathbb{N}$. For the utility of data, the state space $\mathcal{X} = (\mathbb{R} \times \{OFF, ON\})^{|\mathcal{I}|}$, where $|\mathcal{I}|$ denotes the number of TCLs participating in the DLC program. At each time step, the controller issues a command between $[0, 1]$ to each bin. The input space is given by $\mathcal{U} = [0, 1]^N_{\text{bin}}$, and the dynamics ϕ are given by Equations 3.4.17 and 3.4.18. The data sampling policy determining the observables

are modeled by $Y(h,t) = (x(k), m(k))_{k \in T_k}$ with $T_k = \{hl : l \in \mathbb{N}, hl \leq k\}$. Note here that the quality parameter is h, the subsampling period. Thus, $Y(h,t)$ takes values in $\cup_{n=1}^{N} \mathbb{R}^n$. This allows us to define the controller u_c as in Algorithm 3.4.2. Additionally, the cost is given by $J(x,u) = \sum_{k \in T} \| \sum_{i \in \mathcal{I}} P_i m_i(k) - P_{\text{des}} \|_1$. The privacy metric used is inferential privacy. Thus, the choice of subsampling period h affects the structure of the observation function Y, and affects the privacy levels $m(Y,q)$.

Using this framework, we can combine the results in Figures 3.9 and 3.11. This is presented in Figure 3.12. As expected, we can see that lower levels of privacy for consumers result in better load imbalance correction for the direct load controller. This framework allows us to quantify this tradeoff: modeling this tradeoff is the first step towards designing systems that account for privacy, and is essential for formulating economic models of data exchange.

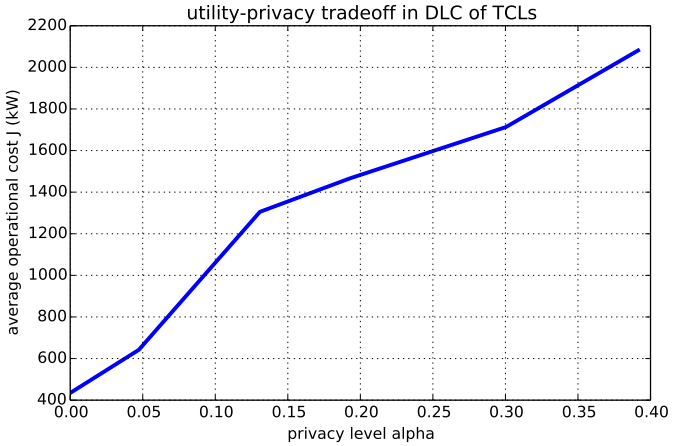

Figure 3.12: The utility-privacy tradeoff in a smart grid application. This depicts a direct load controller's ability to compensate for load imbalances, as a function of the inferential privacy levels of consumers. We chose the time-horizon here to be $N = 6$ hrs.

Closing remarks

In this section, we introduced a control-theoretic framework for quantifying the tradeoff between the utility of data and the privacy loss due to data. Specifically, we considered how variations in the *quality* of data can improve or degrade the operational performance of controllers which utilize this data, and how these variations in quality can change the privacy of users, where privacy is quantified by an appropriately chosen privacy metric.

Additonally, we applied our utility-privacy tradeoff framework in a smart grid application. We considered the direct load control of thermostatically controlled loads, and analyze how its performance degrades as it receives samples less and less frequently—a privacy preserving

metering policy. One of our contributions is a framework for understanding the utility of data in DLC programs, as well as understanding the private information about consumers contained in the data.

As the Internet of Things grows, the potential for privacy breaches is only going to increase. Already, consumers are beginning to become more privacy-aware and sensitive about data transmission policies, and legislative processes are looking into technology-aware policies to handle privacy issues. Moving forward, these technologies need to evolve with a carefully considered privacy component: the utility of collected data must justify the privacy risks involved.

3.4.3 Optimal privacy-by-design example: privacy-enhanced architecture for occupancy-based building control

Large-scale sensing and actuation infrastructures have allowed buildings to achieve significant energy savings; at the same time, these technologies introduce significant privacy risks that must be addressed. In this section, we present a framework for modeling the trade-off between improved control performance and increased privacy risks due to occupancy sensing. More specifically, we consider occupancy-based heating, ventilation, and air conditioning (HVAC) control as the control objective and the location traces of individual occupants as the private variables. Previous studies have shown that individual location information can be inferred from occupancy measurements. To ensure privacy, we design an architecture that distorts the occupancy data in order to hide individual occupant location information while maintaining HVAC performance. Using *mutual information* between the individual's location trace and the reported occupancy measurement as a privacy metric, we are able to optimally design a scheme to minimize privacy risk subject to a control performance guarantee. We evaluate our framework using real-world occupancy data: first, we verify that our privacy metric accurately assesses the adversary's ability to infer private variables from the distorted sensor measurements; then, we show that control performance is maintained through simulations of building operations using these distorted occupancy readings.

Introduction

Large-scale sensing and actuation infrastructures have endowed buildings with the intelligence to perceive the status of their environment, energy usage, and occupancy, and to provide fine-grained and responsive controls over heating, cooling, illumination, and other facilities. However, the information that is collected and harnessed to enable such levels of intelligence may potentially be used for undesirable purposes, thereby raising the question of privacy. To spotlight the value of building sensory data and its potential for exploitation in the inference of private information, we consider as a motivating example the occupancy data, i.e., the number of occupants in a given space over time.

Occupancy data is a key component to perform energy-efficient and user-friendly building management. Particularly, it offers considerable potential for improving energy efficiency of

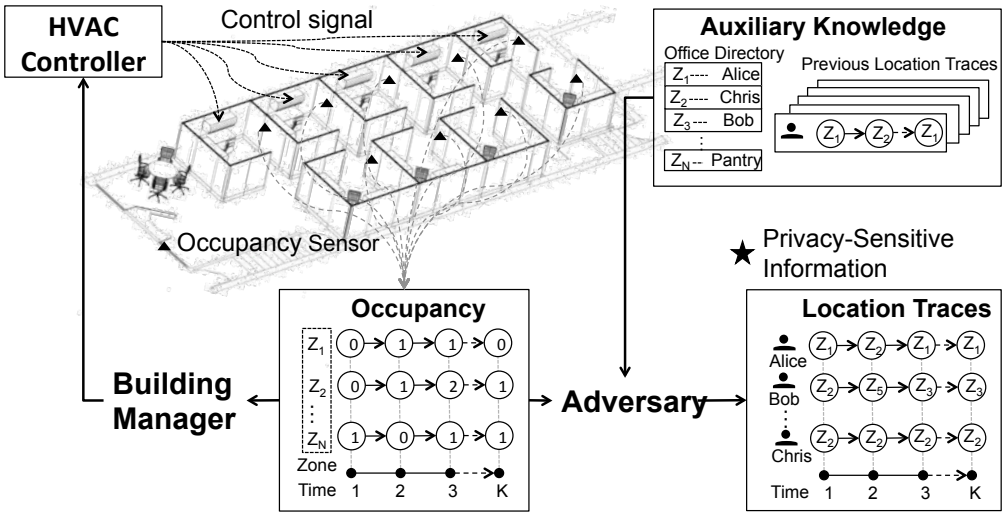

Figure 3.13: An overview of the problem of individual occupant location recovery. The building manager collects occupancy data to enable intelligent HVAC controls adapted to occupancy variations. However, an adversary with malicious intent may exploit occupancy data in combination with the auxiliary information to infer privacy details about indoor locations of building users.

the heating, ventilation, and air conditioning (HVAC) system, a significant source of energy consumption which contributes to more than 50% of the energy consumed in buildings [EIA11]. Recent papers [Bal+13; Kle+14; EC10] have demonstrated substantial energy savings of up to 40% by enabling intelligent HVAC control in response to occupancy variations. The value of occupancy data in building management has also inspired extensive research on occupancy sensing [Don+10; Jin+15; Jin+14; Kha+15; YBG15] as well as a number of commercial products which can provide high accuracy occupancy data.

While people have enjoyed the benefits brought by occupancy data, the privacy risks potentially posed by the data are largely overlooked (Figure 3.13). In effect, location traces of individual occupants can be inferred from the occupancy data with some auxiliary information [WT14]. Throughout this paper, we refer to the individual location trace as the private information to be protected. The contextual information attached to location traces tells much about the individuals' habits, interests, activities, and relationships [Lis+10]. It can also reveal their personal or corporate secrets, expose them to unwanted advertisement and location-based spams/scams, cause social reputation or economic damage, make them victims of blackmail or even physical violence [Sho+11].

At a first glance, it is surprising that occupancy data may incur risks of privacy breach, since it only reports the number of occupants in a given space over time without revealing the identities of the occupants. To illustrate why it is possible to infer location traces from seemingly "anonymized" occupancy data, consider the following scenario. We start by observing two users in one room and then one of them leaves the room and enters another

room. We cannot tell which one of the two made this transition by observing the occupancy change. However, if the one who left entered an private office, the user can be identified with high probability based on the ownership of the office. Although a change in occupancy data may correspond to location shifts of many possible potential users, the knowledge of where the individuals mostly spend their time rules out many possibilities and renders the individual who made the transition identifiable. It has been shown in [WT14] that by simply combining some ancillary information, such as an office directory and user mobility patterns, individual location traces can be inferred from the occupancy data with the accuracy of more than 90%. It is, therefore, the objective of this paper to enable an occupancy-based HVAC control system that provides privacy features for each user on a par with thermal comfort and energy efficiency.

A simple yet effective way to preserve privacy is to obfuscate occupancy data by injecting noise to make the data itself less informative. This approach has been widely used in privacy disclosure control of various databases, ranging from healthcare [DEE13], geolocation [And+13], web-browsing behavior data [Fan+14], etc. While reducing the risk of privacy breach, this approach would also deteriorate the utility of the data. There have been attempts to balance learning the statistics of interest reliably with safeguarding the private information [SC+14]. Cryptography [DH79b] and access control [Wan+14] are also effective means to ease privacy concerns, but they do not provide protection against all privacy breaches. There may be insiders who can access the private, decrypted data, or the building manager may not want to have access to (and responsibility for) the private data.

The objective of this section cannot be attained by simply extending the techniques developed previously. Our task is more challenging. Firstly, as opposed to learning some fixed statistics from static data in most database applications, the data is used for controlling a highly complex and dynamic system in our case, and the control performance relies on the data fidelity. With highly accurate occupancy data, the infrastructure can correctly sense the environment and enable proper response to occupancy variations; nevertheless, the location privacy is sacrificed. On the other hand, the usage of severely distorted occupancy data reduces the risks of privacy leakage, but may lead to even higher levels of energy consumption and discomfort. Essentially, we need to address the trade-off between the performance of a controller on a dynamical system, and, similarly, privacy of a time-varying signal, i.e. the location traces of individual occupants. Secondly, from the perspective of the building manager, the building performance is paramount: adding the privacy feature into the HVAC control system should not impair the performance of HVAC controller in terms of energy efficiency and thermal comfort. To achieve this, the injected noise should be calculated to minimally affect performance of the controller, while maximizing the amount privacy gained from the distortion.

In this section we develop a method which minimizes the privacy risks incurred by collection of occupancy data while guaranteeing the HVAC system operating in a "nearly" optimal condition. Our solution relies on an occupancy distortion mechanism, which can be implemented at the sensor level and "sanitizes" the occupancy data before any form of transport or storage of the data. We draw the inspiration from the information-theoretic approach

in [Raj+11; PCF12a; Ven+15] for characterizing the privacy-utility trade-off, and choose the mutual information (MI) between reported occupancy measurements and individual location traces as our privacy metric. The design problem of finding the optimal occupancy distortion mechanism is cast as an optimization problem where the privacy risk is minimized for a set of constraints on the controller performance. This allows us to find points on the Pareto frontier in the utility-privacy trade-off, and to further analyze the economic side of privacy concerns [Rat+16]. The formulation can be easily generalized to resolve the tension between privacy and data utility in other cases where a control system utilizes some privacy-sensitive information as one of the control inputs, although in this paper we limit our focus to addressing the privacy concern of occupancy-based HVAC controller. In addition, our work here is complementary to the work being done in the cryptography communities: we can use our distortion mechanism to process sensor measurements, and then transmit the processed measurements across secure channels. Our work also serves as a complement for the privacy-preserving access control protocol in [Wan+14], as it provides distortion mechanisms against adversaries who might be able to subvert the protocol while still retaining the benefits for the occupancy data.

The main contributions of this section are as follow:

- We present a systematic methodology to characterize the privacy loss and control performance loss.
- We develop a holistic and tractable framework to balance the privacy pursuit and control performance.
- We evaluate the trade-off between privacy and HVAC control performance using the real-world occupancy data and simulated building dynamics.

The rest of this section is organized as follows. We review the existing work on occupancy-based control algorithms and privacy metrics. We describe the models connecting location and occupancy, and the HVAC system model that will be considered in this paper. We present a framework for quantifying the trade-off between privacy and controller performance. Finally, we evaluate the framework and demonstrate its practical values based on experimental studies and present closing remarks. This work was published in [Jia+16].

Related work in occupancy-based HVAC control

Occupancy-based HVAC systems exploit real-time occupancy measurements to condition the space appropriate to usage. The occupancy-based controllers in the existing work can be categorized into two types: rule-based controller and optimization-based controller or model predictive control (MPC). The rule-based controller uses an "if condition then action" logic for decision making in accordance with occupancy variations [EC10; Bal+13]. MPC is a more advanced control scheme, which employs a model of building thermal dynamics in order to predict the future evolution of the system, and solves an optimization problem

in real-time to determine control actions [Old+12]. A number of papers including [GG10; HK14; Asw+12] analyzed in large-scale simulative or experimental studies the energy saving potential in building climate control by using MPC, which was shown to be well-suited for building applications. This leads to our choice of MPC to exemplify the trade-off between controller performance and privacy.

Occupancy information can be leveraged in different ways in an MPC-based controller. One approach is to build an occupancy model to predict future occupancy based on which the MPC optimizes control actions [BC14]. Another method is to use the instantaneous occupancy measurement and hold it constant during the control horizon of MPC [Goy+13]. This method has been demonstrated to achieve comparable performance with the MPC that exploits occupancy predictions. We will thus without loss of generality follow the latter set-up to avoid explicit modeling of occupancy.

Related work in privacy

Privacy, although not a new topic, has recently developed renewed interest, due in no small part to new technologies and modern infrastructures collecting and storing unprecedented amounts of data. Since privacy is an abstract and subjective concept, it is necessary to develop proper measures for privacy before any privacy protection technique is discussed.

Differential privacy [Dwo06] is one of the most popular metrics for privacy from the area of statistical databases. It is typically assured by adding appropriately chosen random noise to the database output. However, calculating optimal noise for differential privacy is very difficult, and research on the applications of differential privacy mostly assumes the injected noise to be an additive zero-mean Gaussian or Laplacian random variable, which often results in data publication with utility overly sacrificed. As mentioned in the introduction, in our case the performance of HVAC control systems is crucial: as such, our work is an effort to maintain control efficacy by optimally designing noise distribution to maximize privacy subject to a performance guarantee.

Recently, mutual information (MI) has become a popular privacy metric [Raj+11; PCF12a; Jia+15]. Intuitively, MI reflects the change in the uncertainty of a private variable due to the observation of a public random variable. In fact, it is the *only* metric of information leakage that satisfies the data processing inequality [Jia+15]. Unlike differential privacy, this requires some modeling of the adversary's available ancillary information; however, in practice, we can suppose an adversary with access to a large amount of ancillary information, which gives a bound on any weaker adversary's performance. A framework for characterizing privacy-utility trade-off based on MI was proposed in [PCF12a], where the MI between a private variable and a distorted measurement is minimized subject to the bound on the value of an exogenous distortion metric that measures the utility loss from replacing a true measurement with a distorted measurement. Our work is an extension of [PCF12a] to the situations where dynamics at present. We propose a method to abstract out control performance of a dynamical system into a distortion metric, as well as a set of reasonable assumptions for the probabilistic dependencies between occupancy and location data, which allow us to re-

write our privacy metric on time-series data into a static situation akin to that developed in [PCF12a].

Preliminaries

This section collects the concepts we need before introducing the theoretical framework that characterizes the trade-off between privacy and control performance. Two models are described: the *occupancy-location model* that formulates the relationship between occupancy observations and individual location traces, and the model for the HVAC system. We will first consider an occupancy detection system that can collect noise-free or true occupancy, which is then processed by a distortion mechanism into the obfuscated data that the controller observes. We will see the distortion can be similarly applied to noisy occupancy.

Occupancy-location model

Suppose the building of interest consists of N zones represented by $\mathcal{Z} = \{z_0, z_1, \cdots, z_N\}$, where a special zone z_0 is added to refer to the outside of the building. Let $\mathcal{O} = \{o_1, \cdots, o_M\}$ denote the set of occupants. The location of occupant o_m at time k is a random variable denoted by $X_k^{(m)}$ which takes values in the set $\mathcal{Z}$, for $m = 1, \cdots, M$. The true occupancy of zone z_n at time k is denoted by Y_k^n, $n = 0, 1, \cdots, N$. Y_k^n takes values from $\{0, 1, \cdots, M\}$, where M is the total number of occupants in the building. Note that the true occupancy and individual location traces are connected by $Y_k^n = \sum_{m=1}^{M} \mathbb{I}[X_k^{(m)} = z_n]$, where $\mathbb{I}[\cdot]$ is the indicator function.

Additionally, we suppose that the controller observes a distorted version of the true occupancy, denoted by V_k^n which takes values from $\{0, 1, \ldots, M\}$. $\mathbb{P}(V_k^n | Y_k^n)$ represents the distortion mechanism we wish to design. If no distortion on the occupancy data is applied, then $V_k^n = Y_k^n$. We further define some shorthands: $X_k^{(1:M)} := \{X_k^{(1)}, \cdots, X_k^{(M)}\}$, $V_k^{1:N} := \{V_k^1, \cdots, V_k^N\}$.

We make the following assumptions to facilitate the design and analysis of the optimal distortion method. We will show that the optimal distortion proposed works well on the real-world occupancy dataset, which justifies the usage of these assumptions.

Assumption 8. *The location traces for different occupants are mutually independent. That is, we have:* $\mathbb{P}(X_k^{(1:M)}) = \prod_{m=1}^{M} \mathbb{P}(X_k^{(m)})$.

Assumption 9. *The location trace for any given occupant* o_m, $m \in \{1, \ldots, M\}$, *has the first-order Markov property:*

$$\mathbb{P}(X_k^{(m)} | X_{k-1}^{(m)}, X_{k-2}^{(m)}, \ldots, X_1^{(m)}) = \mathbb{P}(X_k^{(m)} | X_{k-1}^{(m)}) \tag{3.4.19}$$

Assumption 10. *The distorted occupancy* V_k^n *depends only on* Y_k^n. *As a result, we can write* $\mathbb{P}(V_k^n | X_k^{(1:M)}) = \mathbb{P}(V_k^n | Y_k^n)$.

These assumptions allow us to model occupancy and location traces via the Factorial Hidden Markov model (FHMM), illustrated in Figure 3.14. The FHMM consists of several independent Markov chains evolving in parallel, representing the location trace of each occupant. Since we only observe the aggregate occupancy information, the location traces are considered to be hidden states.

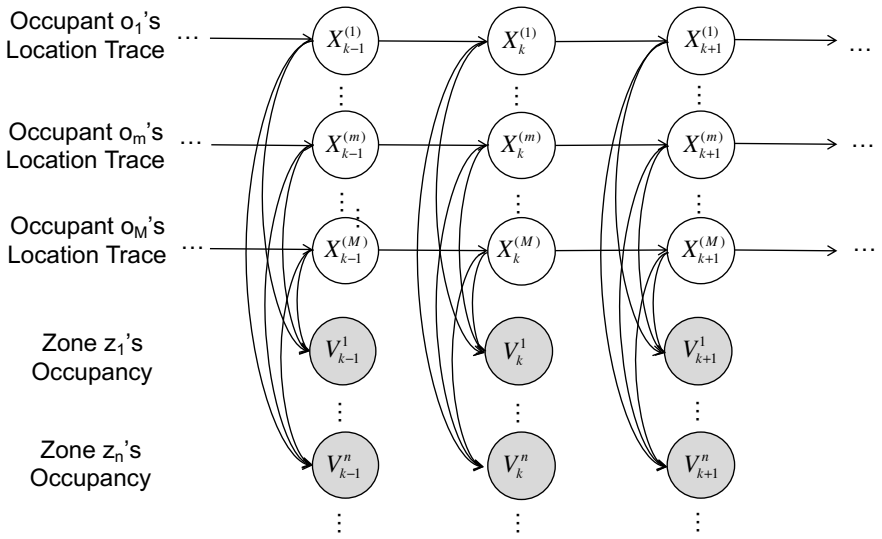

Figure 3.14: The graphical model representation of the FHMM model.

The FHMM model can be specified by the transition probabilities and emission probabilities. The transition probabilities describe the mobility pattern of an occupant, which is denoted as a $(N+1) \times (N+1)$ transition matrix. We define the transition matrix for occupant o_m as $A^{(m)} = [a_{ij}^{(m)}]$, $i, j = 0, 1, \cdots, N$, where $a_{ij}^{(m)} = \mathbb{P}(X_{k+1}^{(m)} = z_j | X_k^{(m)} = z_i)$ for $k = 0, 1, \cdots, K-1$. The transition parameters can be learned from the occupancy data based on maximum likelihood estimation. If the prior knowledge about the past location traces is also available, it can be encoded as the prior distribution of transition parameters from a Bayesian point of view, and then the transition parameters can be learned via *maximum a posteriori* (MAP) estimation. We refer the readers to [WT14] for the details of parameter learning. The emission probabilities characterize the conditional distribution of distorted occupancy given the location of each occupant, defined by

$$\mathbb{P}(V_k^{1:N} | X_k^{(1:M)}) = \prod_{n=1}^{N} \mathbb{P}(V_k^n | X_k^{(1:M)}) = \prod_{n=1}^{N} \mathbb{P}(V_k^n | Y_k^n) \tag{3.4.20}$$

The above equalities result from Assumption 10, which, in other words, indicates that the distorted occupancy depends on individual location traces only via the true occupancy.

HVAC system model

Suppose the thermal comfort of the building space of interest is regulated by the HVAC system shown in Figure 3.15, which provides a system-wide Air Handling Unit (AHU) and Variable Air Volume (VAV) boxes distributed at the zones. In this type of HVAC system, the outside air is conditioned at the AHU to a setpoint temperature T_a by the cooling coil inside. The conditioned air, which is usually cold, is then supplied to all zones via the VAV box at each zone. The VAV box controls the supply air flow rate to the thermal zone, and heats up the air using the reheat coils at the box, if required. The control inputs are temperature and flow rate of the air supplied to the zone by its VAV box. The AHU outlet air temperature setpoint T_a is assumed to be constant in this paper. The HVAC system models described in the subsequent paragraphs will follow [KB11; BC14; Goy+13] closely[13].

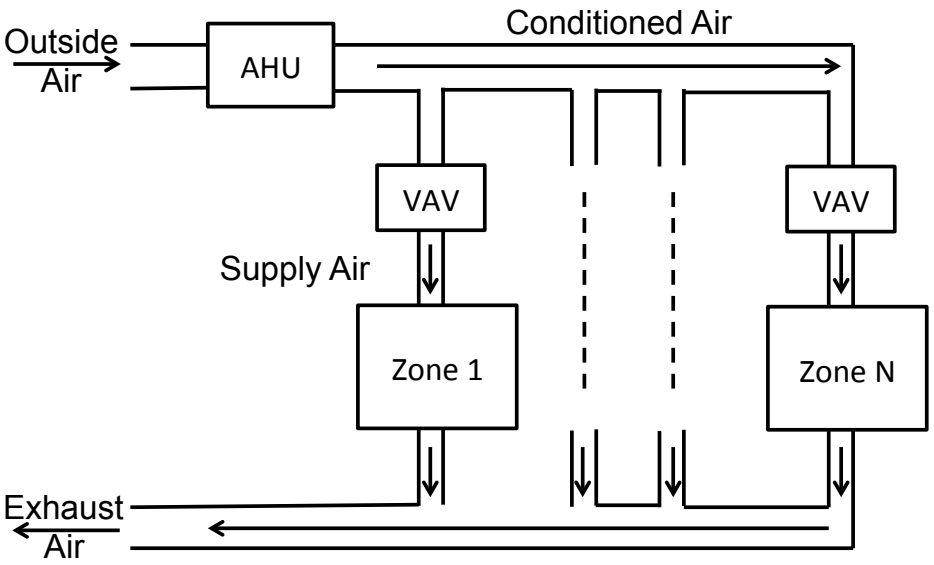

Figure 3.15: A schematic of a typical multi-zone commercial building with a VAV-based HVAC system.

State model. With reference to the notations in Table 3.1, the continuous time dynamics for the temperature T^n of zone z_n can be expressed as

$$C^n \frac{d}{dt} T^n = \mathbf{R}^n \cdot \mathbf{T} + Q^n + \dot{m}_s^n c_p (T_s^n - T^n) \tag{3.4.21}$$

where the superscript n indicates that the associated quantities are attached to zone z_n. $\mathbf{T} := [T^1, \cdots, T^N]$ is a vector of temperature of all N zones. $\mathbf{R}^n$ indicates the heat transfer among different zones and outside. Q^n is the thermal load, which can be obtained by applying

[13]Controlling the flow rate is actually more preferable in building codes in consideration of energy efficiency. Herein, we consider both reheat temperature and flow rate are controllable, while the HVAC model with flow rate as the only control input is a simple application of our model.

Table 3.1: Parameters used in the HVAC controller.

Param.	Meaning	Value & Units
Δt	Discretization step	$60s$
c_p	Thermal capacity of air	$1kJ/(kg \cdot K)$
C^n	Thermal capacity of the env.	$1000kJ/K$
c_o	Thermal load per person	$0.1kW$
R	Heat transfer vector	$0kW/K$
η_h	Heating efficiency	0.9
η_c	Cooling efficiency	4
β	System parameter	$0.5kW \cdot s/kg$
r_e	Electricity price	$1.5 \cdot 10^{-4}\$/kJ$
r_h	Heating fuel price	$5 \cdot 10^{-6}\$/kJ$
$\underline{T}$	Upper bound of comfort zone	$24°C$
$\overline{T}$	Lower bound of comfort zone	$26°C$
T_a	AHU outlet air temperature	$12.8°C$
$\underline{m_s}$	Minimum air flow rate	$0.0084kg/s$
$\overline{m_s}$	Maximum air flow rate	$1.5kg/s$
$\overline{T}_h$	Heating coil capacity	$40°C$

a thermal coefficient c_o to the number of occupants V^n, i.e., $Q^n = c_o V^n$. The control inputs $U^n := [\dot{m}_s^n, T_s^n]$ are the supply air mass flow rate and temperature. Assuming $\dot{m}_s^n$, T_s^n and Q^n are zero-order held at sample rate Δt, we can discretize (3.4.21) using the trapezoidal method and obtain a discrete-time model, which can be expressed as

$$C^n \frac{T_{k+1}^n - T_k^n}{\Delta t} = R^n T_k + c_o V_k^n + \dot{m}_{s,k}^n c_p \left(T_{s,k}^n - \frac{T_{k+1}^n + T_k^n}{2} \right) \tag{3.4.22}$$

where k is the discrete time index and $T_k^n = T_t^n|_{t=k\Delta t}$. Q_k^n, $\dot{m}_{s,k}^n$ and $T_{s,k}^n$ are similarly defined.

Cost function. The control objective is to condition the room while minimizing the energy cost. The power consumption at time k consists of reheating power $P_{h,k}^n = \frac{c_p}{\eta_h} \dot{m}_{s,k}^n (T_{s,k}^n - T_a)$, cooling power $P_{c,k}^n = \frac{c_p}{\eta_c} \dot{m}_{s,k}^n (T_o - T_a)$ and fan power $P_{f,k}^n = \beta \dot{m}_{s,k}^n$, where η_h and η_c capture the efficiencies for heating and cooling side, respectively. β stands for a system dependent constant. We introduce several parameters to reflect utility pricing, r_e for electricity and r_h for heating fuel. These parameters may vary over time.

Therefore, the total utility cost of zone z_n from time $k = 1, \cdots, K$ is:

$$J^n = \sum_{k=1}^{K} \left((r_{e,k} P_{f,k}^n + r_{h,k} P_{h,k}^n + r_{e,k} P_{c,k}^n) \Delta t \right)$$

Constraints. The system states and control inputs are subject to the following constraints:

C1: $\underline{T} \leq T_k^n \leq \overline{T}$, comfort range;

C2: $\underline{\dot{m}_s} \leq \dot{m}_{s,k}^n \leq \overline{\dot{m}_s}$, minimum ventilation requirement and maximum VAV box capacity;

C3: $T_{s,k}^n \geq T_a$, heating coils can only increase temperature;

C4: $T_{s,k}^n \leq \overline{T}_h$, heating coil capacity.

These constraints hold at all times k and all zones $\{z_n\}_{n=1}^N$.

MPC controller. Knitting together the models described above, we present an MPC-based control strategy for the HVAC system to efficiently accommodate for occupancy variations. In this control algorithm, we assume that the predicted occupancy during the optimization horizon to be the same as the instanteneous occupancy observed at the beginning of control horizon. It was shown in [Goy+13] that the control algorithm with this assumption can achieve comparable performance with the MPC that constructs explicit occupancy model to predict occupancy for future time steps.

Let $U_{1:K}^{1:N}$ be the shorthand for $\{U_k^n | k = 1, \cdots, K, \quad n = 1, \cdots, N\}$. The optimal control inputs for the next K time steps are obtained by solving $\min_{U_{1:K}^{1:N}} \sum_{n=1}^N J^n$, subject to the inequality constraints C1-C4 and the equality constraint (3.4.22) and $T_1^n = T_{init}^n$, $\forall n = 1, \cdots, N$, where T_{init}^n is the initial temperature of zone z_n at each MPC iteration. We can see that the optimal control input is a function of the distorted occupancy that the controller sees and the initial temperature. We express this relationship explicitly by denoting the optimal control action at zone z_n as $U_{MPC}^n(V^n, T_{init}^n)$. In addition, the energy cost incurred by applying the optimal control action is denoted by $J_{MPC}^n(U_{MPC}^n(V^n, T_{init}^n), Y^n)$, where the second argument stresses that the actual control cost is dependent on the real occupancy.

Privacy-enhanced control

With the HVAC model established, we can now develop the mathematical framework to discuss a privacy-enhanced architecture. We will first introduce MI as the metric we use throughout the paper to quantify privacy, and then present a method to optimally design the distortion mechanism which minimizes the privacy loss within a pre-specified constraint on control performance.

Privacy metric

Definition 14. *[CT12] For random variables X and V, the* mutual information *is given by:*

$$I(X;V) = H(X) - H(X|V) \tag{3.4.23}$$

where $H(X)$ *and* $H(X|V)$ *represent* entropy *and* conditional entropy*, respectively. Let* $\mathbb{P}_X(x) = \mathbb{P}(X = x)$*,* $H(X)$ *and* $H(X|V)$ *are defined as*

$$H(X) = -\sum_x \mathbb{P}_X(x) \log(\mathbb{P}_X(x)) \tag{3.4.24}$$

$$H(X|V) = -\sum_v \mathbb{P}_V(v) \Big(\sum_x \mathbb{P}_{X|V}(x|v) \log\big(\mathbb{P}_{X|V}(x|v)\big) \Big) \tag{3.4.25}$$

Remark. Entropy measures uncertainty about X, and conditional entropy can be interpreted as the uncertainty about X after observing V. By the definition above, MI is a measure of the reduction in uncertainty about X given knowledge of V. We can see that it is a natural measure of privacy since it characterizes how much information one variable tells about another. It is also worth noting that inference technologies evolve and MI as a privacy metric does not depend on any particular adversarial inference algorithm [Raj+11] as it models the statistical relationship between two variables.

In this paper, we will be using the MI between location traces and occupancy observations, i.e., $I(X_k^{(1:M)}; V_k^{1:N})$, as a metric of privacy loss. This metric reflects the reduction in uncertainty about location traces $X_k^{(1:M)}$ due to observations of $V_k^{1:N}$. As a proof of concept, we will verify that this metric serves as an accurate proxy for an adversary's ability to infer individual location traces in the experiments. We further introduce some assumptions which allow us to simplify the expression of the privacy loss and obtain a form of MI that has direct relationship with the distortion mechanism $P(V_k^n|Y_k^n)$ we wish to design.

Based on results in ergodic theory [Kal02], we know that the probability distribution of individual location traces will converge to a unique stationary distribution under very mild assumptions[14]. For more details on stationary distributions, we refer the reader to [Kal02]. This observation justifies the following:

Assumption 11. *The Markov chains* $X_k^{(m)}$ *have a unique stationary distribution for all occupants* o_m *and are distributed according to those stationary distributions for all time steps* k*.*

Combining this assumption and the occupancy-location model we presented in the preceding section, we present a proposition that allows us to greatly simplify the form of the privacy loss:

Proposition 15. *By Assumption 10, we have that:*

$$I(X_k^{(1:M)}; V_k^{1:N}) = I(Y_k^{1:N}; V_k^{1:N}) \tag{3.4.26}$$

By Assumption 11, we have that $I(Y_k^{1:N}; V_k^{1:N})$ *is a constant for all* k*, so we will drop the subscript:* $I(Y^{1:N}; V^{1:N})$*.*

[14]Since there are only finitely many zones, a sufficient condition is the existence of a path from z_i to z_j with positive probability for any two zones z_i and z_j.

Finally, by the various conditional independences introduced in Assumption 10:

$$I(Y^{1:N}; V^{1:N}) = \sum_{n=1}^{N} I(Y^n; V^n) \tag{3.4.27}$$

Remark. The result that $I(Y_k^{1:N}; V_k^{1:N})$ is a constant value for all k allows us to design a single distortion mechanism $P(V^n|Y^n)$ for all time steps (note that we drop the subscript k to indicate the time-homogeneity of the distortion mechanism). By Proposition 15, minimization of privacy loss $I(X_k^{(1:M)}; V_k^{1:N})$ can be conducted by minimizing a simpler expression $\sum_{n=1}^{N} I(Y^n; V^n)$.

Optimal distortion design

We wish to find a distortion mechanism $P(Y^n|V^n)$ that can produce some perturbed occupancy data with minimum information leakage, while the performance of the controller using the perturbed occupancy data is on a par with that using true occupancy. To be specific, we will bound the difference of energy costs incurred by the controllers seeing distorted and real occupancy data.

Let T_{init1} and T_{init2} be initial temperature of the controller using distorted and real occupancy, respectively. Recall that $U_{MPC}^n(V^n,\, T_{init}^n)$ and $J_{MPC}^n(U_{MPC}^n(V^n, T_{init}^n), Y^n)$ stand for the optimal control actions and the associated cost based on the distorted occupancy; correspondingly, if the controller sees the real occupancy data, the optimal control action and the associated cost will be $U_{MPC}^n(Y^n, T_{init}^n)$ and $J_{MPC}^n(U_{MPC}^n(Y^n, T_{init}^n), Y^n)$, respectively. We denote the resulting temperature after applying optimal control actions as $T_{MPC}^n(U_{MPC}^n(V^n, T_{init}^n), Y^n)$, where the second argument emphasizes that the temperature evoluation depends on the true occupancy. We introduce the following constraints: $\forall |T_{init1} - T_{init2}| \leq \Delta_T'$, $y = 0, \cdots, M$, $n = 1, \cdots, N$,

C5: Cost difference constraint

$$E_{\mathbb{P}(V^n|Y^n=y)}\Big[J_{MPC}^n\big(U_{MPC}^n(T_{init1}, V^n), y\big) - J_{MPC}^n\big(U_{MPC}^n(T_{init2}, y), y\big)\Big] \leq \Delta \tag{3.4.28}$$

C6: Resulting temperature constraint

$$E_{\mathbb{P}(V^n|Y^n=y)}\Big[\big|T_{MPC}^n\big(U_{MPC}^n(T_{init1}, V^n), y\big) - T_{MPC}^n\big(U_{MPC}^n(T_{init2}, y), y\big)\big|\Big] \leq \Delta_T \tag{3.4.29}$$

C5 states that the cost difference between using the distorted occupancy measurements V^n and using the ground truth occupancy measurements Y^n is bounded by Δ in expectation, for any possible value of Y^n. The cost difference can be regarded as the control performance loss due to the usage of distorted data, and Δ stands for the tolerance on the control performance loss. C5 alone is a one-step performance guarantee, that is, it only bounds the cost difference associated with a single MPC iteration. In practice, MPC is repeatedly solved

from the new initial temperature, yielding new control actions and temperature trajectories. In order to offer a guarantee for future cost difference, we introduce another constraint C6 on the resulting temperature difference of one MPC iteration. The idea is that the resulting temperature will become the new initial temperature of the next MPC iteration. If the resulting temperature difference between using distorted occupancy data and using true occupancy data is bounded within a small interval Δ_T, in the next MPC iteration C5 will provide a bound on cost difference for new initial temperatures that do not differ too much, since the cost difference constraint C5 is imposed to hold for all $|T_{init1} - T_{init2}| \leq \Delta'_T$. Typically, Δ'_T is set to be similar to Δ_T, but a small value of Δ'_T is preferred in order to assure the feasibility of the optimization problem (since the number of constraints increases with Δ'_T).

Now, we are ready to present the main optimization for privacy-enhanced HVAC controller by combining the privacy metric and performance constraint just presented. Suppose the assumptions of Proposition 15 hold. Given the control performance loss tolerance Δ, the *optimal distortion mechanism* is given by solving:

$$\min_{\substack{\mathbb{P}(V^n|Y^n) \\ n=1,\cdots,N}} \sum_{n=1}^{N} I(Y^n; V^n) \tag{3.4.30}$$

subject to the constraint C5-C6. Δ serves as a knob to adjust the balance between privacy and the controller performance loss. Increasing Δ leads to larger feasible set for the optimization problem, and thus a smaller value of MI (or privacy loss) is expected. Using the methodology presented previously, we are able to calculate the terms inside the expectation in (3.4.28) and (3.4.29) for all $|T_{init1} - T_{init2}| \leq \Delta'_T$ and $y = 0, \cdots, M$. Treating these as constants, calculating the optimal privacy-aware sensing mechanism is a convex optimization program, and can be efficiently solved. Additionally, since the constraints are enforced for each zone, the optimization (3.4.30) can actually be decomposed to N sub-problems and thus we can solve the optimal distortion scheme separately for each zone.

Remark on noisy occupancy data. In the preceding privacy-enhanced framework, we consider the occupancy can be accurately detected. In practice, the occupancy data may be noisy itself, and thereby the distortion mechanism will be designed based on noisy occupancy W_k^n instead of true occupancy Y_k^n. In effect, the distortion designed using noisy occupancy provides an upper bound on the privacy loss. That is, in practice we could use noisy occupancy to design the distortion mechanism and the realized privacy loss can only be lower than the minimum privacy loss obtained from the optimization. Note that we have the Markov relationship: $Y_k^n \to W_k^n \to V_k^n$ when the distortion is applied to noisy data. Then the proof follows from the data processing inequality [CT12].

Experiment setup

Occupancy dataset. The occupancy data used in this paper is from the Augsburg Indoor Location Tracking Benchmark [Pet04], which includes location traces for 4 users in a office

building with 15 zones. The location data in the benchmark dataset was recorded every second over a period of 4 to 9 weeks. Since the dataset contains some missing observations due to technical issues or the vacation interruption, we finally use the dataset from November 5th to 24th in our experiment, during which the location traces of all the 4 users are complete, and subsample the dataset with 1-minute resolution. The ground truth occupancy data was synthesized by aggregating the locations trace of each user. Table 3.2 shows two statistics of the benchmark dataset. Notably, of all transitions per day, 66.7% to 84.6% either start from or end at one's own office, and office location can divulge one's identity. This sheds light on why location traces of individual users can be actually inferred from the "anonymized" occupancy data.

Table 3.2: The average number of transitions each user made in each workday, and the average percentage of transitions from or to one's office.

User	avg # of transitions per day	avg % of transitions from/to office per day
1	9.3	84.6%
2	20.2	75.4%
3	9.9	66.7%
4	7.6	75.5%

Adversary inference. We consider the adversary to be an *insider* with authorized building automation system access. One can think of it as the worst case of privacy breach, because insiders not only learn the ancillary information that is public-available, but are familiar with building operation policies. To be specific, the following auxiliary information is assumed to be available to the adversary: (1) Building directory and occupant mobility patterns, encoded by the transition matrix of each occupant[15]; (2) Occupancy distortion mechanism designed by building manager.

The adversary attempts to reconstruct the most probable location trace given the occupancy data and the auxiliary information. That is, the attack is to find the MAP of location traces given the other information. The approach to finding MAP is well known as Viterbi algorithm in HMM. However, Viterbi is infeasible in the FHMM case as the location traces to be solved reside in a exponentially large state space ($N^M \times K$). We propose a fast inference method based on Mixed Integer Programming, and thus more efficiently evaluate the adversary's inference attack. The interested readers are referred to the code implementation of this paper for the details of the fast inference algorithm.

Controller parameters. Without loss of generality, we consider the zones have the same thermal properties. The comfort range of temperature in the zones is defined to be within $24 - 26°C$ as in [Nag+15]. The minimum flow rate is set to be $0.084kg/s$ to fulfill the minimum ventilation requirement for $25m^2$-sized zone as per ASHRAE ventilation

[15] In the experiment, we use 4 days' occupancy data and 2 days' location traces to learn these parameters and the rest for evaluating our framework.

standard 62.1-2013 [Ame]. The optimization horizon of the MPC is 120 min, and the control commands are solved for and updated every 15 min [Goy+13]. Other design parameters are shown in Table 3.1, which bascially follows the choices in [KB11].

Platform. The algorithms are implemented in MATLAB; The interior-point algorithm is used to solve the bilinear optimization problem in MPC. To encourage the research on the privacy-preserving controller, the codes involved in this paper will be open-sourced in `http://ruoxijia.github.io/4_code`.

Results: MI as proxy for privacy

We solve the MI optimization for different tolerance levels of control performance deterioration due to the usage of the distorted data, i.e., Δ, and obtain a set of optimal distortion designs and corresponding optimal values of MI. We then randomly perturb the true occupancy data using the different distortion designs, and infer location traces from the perturbed occupancy data. Monte Carlo (MC) simulations are carried out to assess results under the random distortion design. The inference accuracy is defined to be the ratio between the counts of correct location predictions over the total time steps. Figure 3.16 demonstrates the monotonically increasing relationship between adversarial location inference accuracy and MI, which justifies the usage of MI as a measure of privacy loss. When the adversary has perfect occupancy data, individual location traces can be inferred with accuracy of 96.81%. On the contrary, when the MI approaches zero, the adversary tends to estimate the location of each user to be constantly outside of the building, which is the best estimate the adversary can generate based on the uninformative occupancy data since people spend most of their time in a day outside. In this case, the inference accuracy is 77% but the adversary actually has no knowledge about users' movement. This serves as a baseline of the adversarial location inference performance.

Results: optimal utility-privacy tradeoff

Figure 3.17 shows the variation of privacy loss and controller performance loss with respect to different choices of Δ, which is the theoretical guarantee on controller performance loss. It is evident that privacy loss and control performance loss exhibit opposite trends as Δ changes. The privacy loss, measured by MI, monotonically decreases as Δ gets larger. This is the manifestation of the intrinsic utility-privacy trade-off embedded in the main optimization problem (3.4.30). As the performance constraint Δ is more relaxed, a smaller value of MI can be attained and thus privacy can be better preserved. The actual performance loss, measured by the HVAC control cost difference (between using distorted and true data) averaged across different MPC iterations and difference zones, generally increases with Δ and is upper bounded by Δ. This indicates that the theoretical constraint on controller performance loss in our framework is effective and can actually provide a guarantee on the actual controller performance. We can see that the bound is far from tight, since the framework enforces the constraints on the controller performance for every possible true

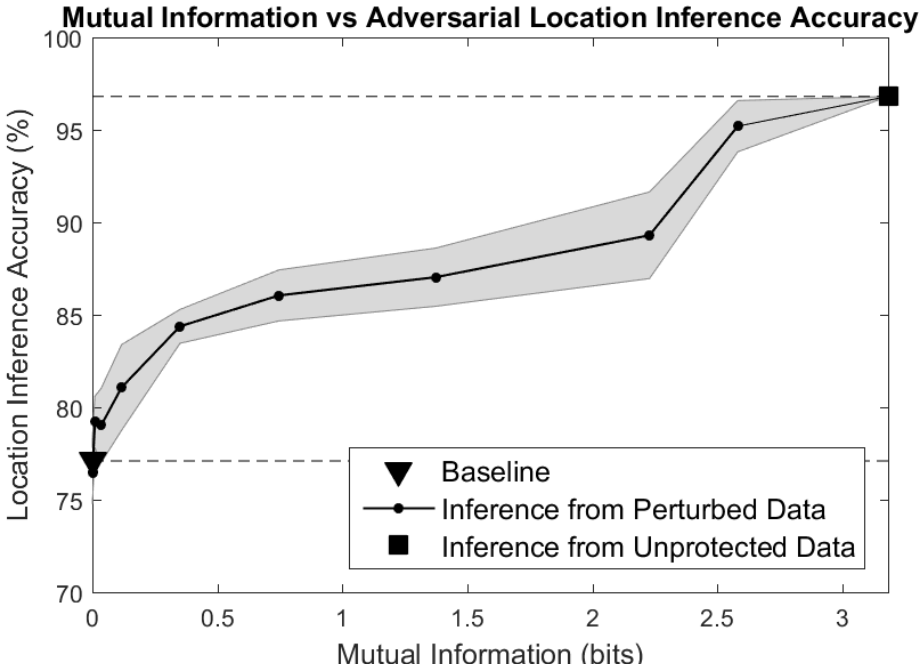

Figure 3.16: The adversary location inference accuracy increases as MI increases. The black line and the band around it show the mean and standard deviation of inference accuracy across ten MC simulations, respectively. The black square shows the location inference accuracy if the adversary sees true occupancy data. The black triangle gives the accuracy when the adversary outputs a constant location estimate.

occupancy value to ensure the robustness while in practice the occupancy distribution is very spiked about the mean occupancy.

Figure 3.18 visualizes the distortion mechanism obtained by solving the MI under different choices of the tolerance on the control performance loss Δ. It can be clearly seen that the mechanism creates a higher level of distortion as Δ increases. When Δ is small, the resulting distortion matrix assigns most probability mass on the diagonal, i.e., the occupancy is very likely to keep unperturbed. As Δ gets larger, the distortion mechanism tends to have the same rows, in which case the distribution of distorted occupancy data is invariant under the change of true occupancy and MI between true occupancy and perturbed occupancy, i.e., the privacy loss, tends to be zero. We also plot the temperature evolution under different distortion levels. Since we enforce a hard constraint on temperature, we can see that the zone temperature stays within the comfort zone for all Δ's. However, larger Δ would lead to a larger deviation from the temperature controlled using the true occupancy.

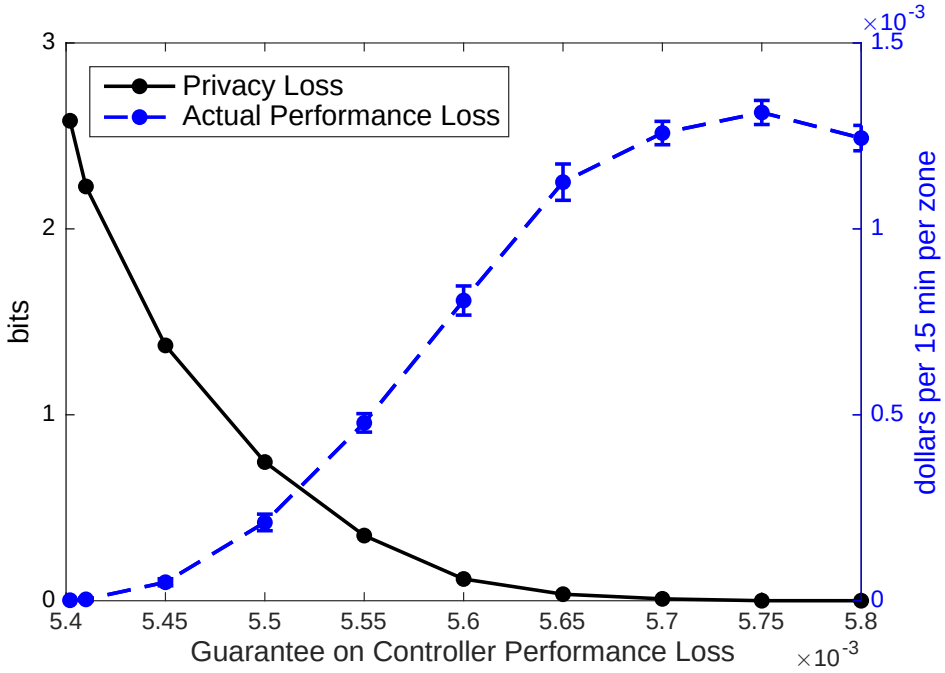

Figure 3.17: The changes of MI and actual control cost difference between using true and perturbed occupancy as the theoretical control cost difference changes. The blue dot line and errorbar demonstrate the mean and standard deviation of actual control cost difference across ten MC simulations, respectively.

Comparison with other methods

We compare the performance of the HVAC controller using our optimally perturbed data against using unperturbed occupancy data, fixed occupancy schedule as well as randomly perturbed data by other distortion methods. In Figure 3.4.3 we plot the privacy loss and control cost for controllers that use the various forms of occupancy data. Fixed occupancy schedule (assuming maximum occupancy during working hours and zero otherwise) exposes zero information about individual location traces, but cannot adapt to occupancy variations and thus incurs considerable control cost. The controller based on clean occupancy data is most cost-effective but discloses maximum private information. One of the random distortion method to be compared is uniform distortion scheme in which the true occupancy is perturbed to some value between zero to maximum occupancy with equal probability. We carry out 10 MC simulations to obtain the control cost incurred under this random perturbation scheme. It can be seen that the uniform distortion scheme protects the private information with compromised controller performance.

A natural question arising is if the current occupancy sensing systems provide intrinsic privacy-preserving features as there always exists occupancy estimation errors. Can we use a cheaper and inaccurate occupancy sensor to acqiure privacy? As is suggested by

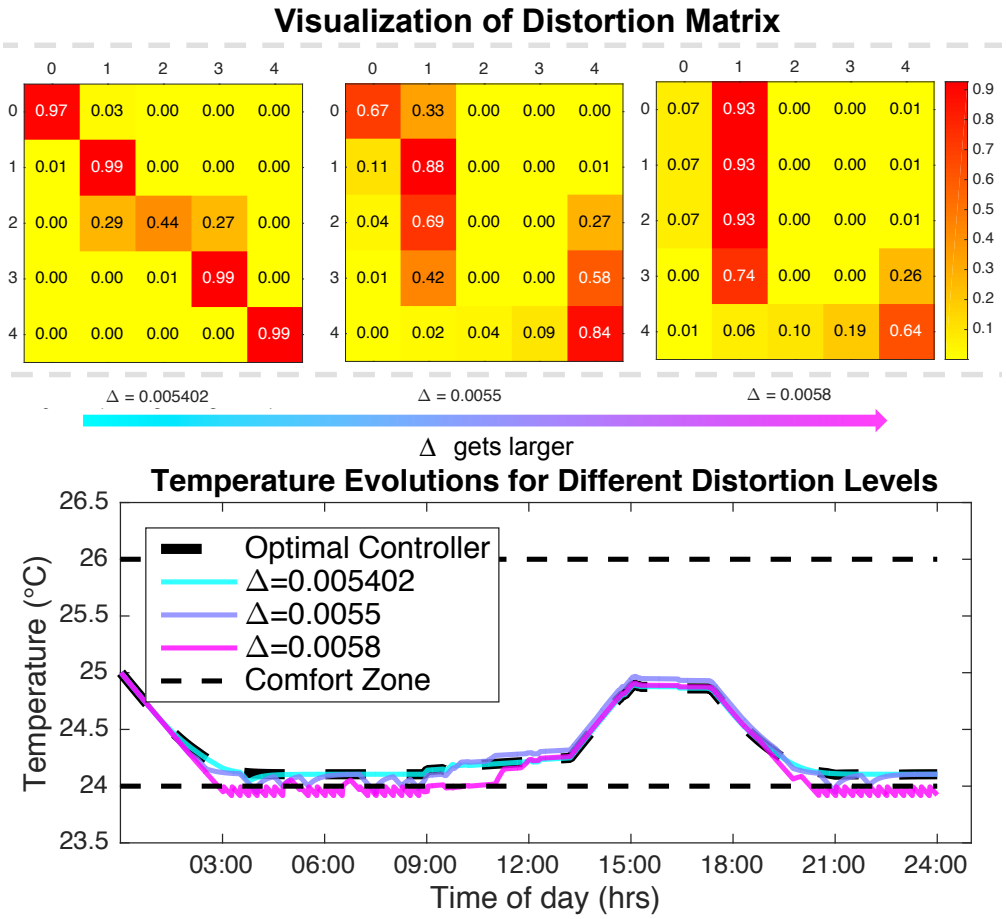

Figure 3.18: Illustration of distortion matrix $P(V|Y)$ under different controller performance guarantees. The row index corresponds to the value of Y, while colomn index corresponds to V. The zone temperature traces resulted from the controllers using occupancy data that is randomly distorted by different distortion matrices are also shown.

the occupancy sensing results in [Jin+15], the estimation noise of a real occupancy sensing system can be modeled by a multinomial distribution which has most probability mass at zero. Inspired by this, we use the following multinomial distortion schemes to imitate a real occupancy sensing system with disparate accuracies acc,

$$P(V^n|Y^n = y) = \begin{cases} acc, & V = y \\ \frac{1-acc}{2}, & V = y-1 \text{ or } y+1 \text{ if } y \neq 0 \\ \frac{1-acc}{2}, & V = 1 \text{ or } 2, \text{ if } y = 0 \end{cases} \tag{3.4.31}$$

Again, MC simulations are performed to evaluate the control performance under this random perturbation, and the results are shown in Figure 3.4.3. It can be seen that when the privacy loss is relatively large (or data is slightly distorted), the control cost of our optimal

noising scheme and the multinomial noising scheme do not differ too much. This is because at this level of privacy loss the two distortion schemes behave similarly, as shown in Figure 3.18, where the occupancy keeps untainted with high probability. But as the privacy loss decreases, our optimal noising scheme's intelligent noise placement begins to significantly improve control performance. In addition, our optimal distortion Pareto dominates the other schemes.

To investigate the scalability of our proposed scheme, we create synthetic data that simulates location traces for 15 occupants based on the Augsburg dataset. We extract the occupants' movement profile, i.e., transition parameters, from the original dataset and randomly assign the profiles to synthesized occupants. An occupant randomly chooses the next location according to the movement profile. The privacy-utility curve evaluated on this larger synthesized dataset is illustrated in Figure 3.4.3, which demonstrates that the optimality of our distortion scheme is preserved when the experiment is scaled up. We can see that the privacy loss of the controller using the unperturbed occupancy gets lower when incorporating more occupants. Although privacy risks are lower as we scale up the experiment since with more people sharing the space it will be more difficult to identify each individuals, adding distortion to occupancy measurements can preserve the privacy even further as shown in Figure 3.4.3.

Closing remarks on the optimal privacy-by-design example

In this section, we present a tractable framework to model the trade-off between privacy and controller performance in a holistic manner. We take occupancy-based HVAC controller as an example where the objective is to utilize occupancy data to enable smart controls over the HVAC system while protect individual location information from being inferred from the occupancy data. We use MI as the measure of privacy loss, and formulate the privacy-utility trade-off by a convex optimization problem that minimizes the privacy loss subject to a pre-specified controller performance constraint. By solving the optimization problem, we can obtain a mechanism that injects optimal amount of noise to occupancy data to enhance privacy with control performance guarantee. We verify our framework using real-world occupancy data and simulated building dynamics. It is shown that our theoretical framework is able to provide guidelines for practical privacy-enhanced occupancy-based HVAC system design, and reaches a better balance of privacy and control performance compared with other occupancy-based controllers.

3.5 Conclusion

In this chapter, we built on the ideas presented in Chapter 2. Namely, now that IoT sensors allow the disaggregation and inference of so many facets of our previously unmeasured lives, what are the privacy risks that arise? First, we presented different methods for quantifying privacy. Then, we discussed different design paradigms by which privacy can be incorporated

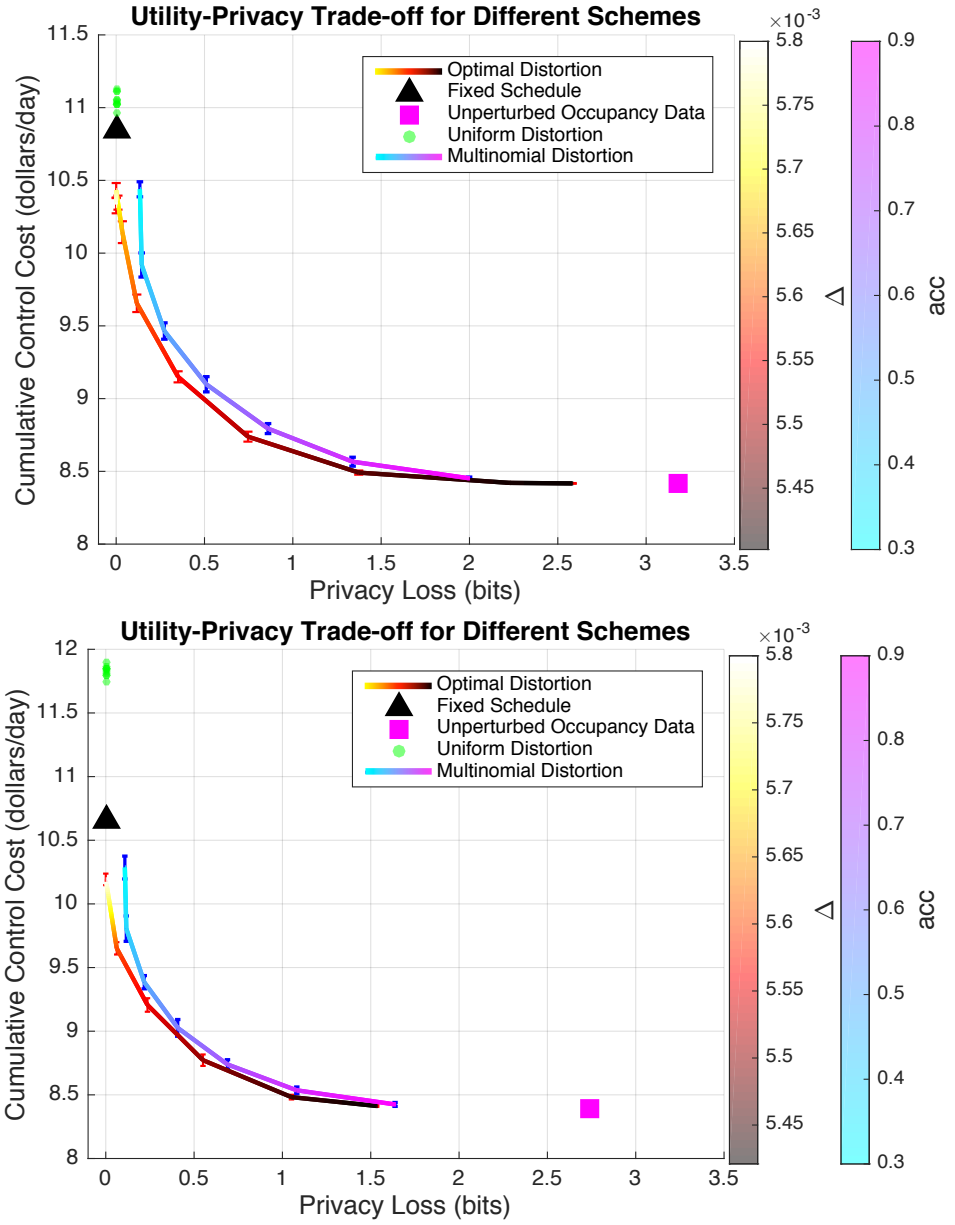

Figure 3.19: Comparison of the privacy-utility trade-off of controllers using different forms of occupancy data, evaluated based on (a) real-world occupancy data and (b) synthesized data.

into our systems. Finally, we presented some of our work on implementing these design paradigms in practice: when we quantify the level of privacy of an existing system, when we vary design parameters to achieve a certain tradeoff between privacy and control, and when we optimally design our noising schemes to maximize privacy while maintaining a pre-specified level of control performance. This work lays the foundation for incorporating privacy as a part of the design process, rather than a constraint added later in the pipeline.

Chapter 4

The Value of Information, Data Markets, and New Service Models in Sample IoT Applications

The Internet of Things (IoT) is an ecosystem that thrives on the abundance of available data: this information flows into analytics engines and estimation algorithms and forms the knowledge base from which new incentives, mechanisms, and controllers are designed. We've already discussed new estimation algorithms for IoT in Chapter 2, and also outlined methods for the analysis of user privacy in Chapter 3. However, most of this work abstracts out the preferences and behavior of the user. For example, when we considered privacy-by-design, we ignored the question: 'How much privacy do users want?'

Users are becoming increasingly privacy-aware and privacy-conscious. In this chapter, we consider models for when users become more aware of the value of their data, and begin to interact with the market for the data in a more strategic fashion. As users become more strategic in their data-sharing behaviors, companies will have to adapt their incentives accordingly. In other words, once we acknowledge that our data comes from human sources, we cannot treat them as independent and identically distributed samples from a distribution anymore; rather, their reported data is the result of a strategic consideration of what maximizes their own benefit. This is discussed in detail in Section 4.1.

Additionally, the Internet of Things introduces novel means by which to influence our systems. These means of influence often do not fall into the traditional analysis of control. For example, when a governmental agency issues monetary rebates for eco-friendly appliances, this is not obviously the input to a linear system. Rather, it is an action taken in hopes of having a causal effect on the behaviors of users and the dynamics of the underlying physical system. We refer to these actions as *causal imputations*. We present the problem of optimal causal imputations and solve it in a few special cases in Section 4.2.

This work can be thought of as the dual to the current work on causal inference and estimation: rather than fixing the causal actions and identifying the causal structure, we fix the causal structure as known and find causal actions that influence the system in desirable

directions for minimal cost.

4.1 Statistical estimation with strategic data sources in competitive settings

In this section, we introduce a preliminary model for interactions in the data market. Recent research has shown ways in which a data aggregator can design mechanisms for users to ensure the quality of data, even in situations where the users are effort-averse (i.e. prefer to submit lower-quality estimates) and the data aggregator cannot observe the effort exerted by the users (i.e. the contract suffers from the principal-agent problem). However, we have shown that these mechanisms often break down in more realistic models, where multiple data aggregators are in competition. Under minor assumptions on the properties of the statistical estimators in use by data aggregators, we show that there is either no Nash equilibrium, or there is an infinite number of Nash equilibrium. In the latter case, there is a fundamental ambiguity in who bears the burden of incentivizing different data sources. We are also able to calculate the price of anarchy, which measures how much social welfare is lost between the Nash equilibrium and the social optimum, i.e. between non-cooperative strategic play and cooperation.

4.1.1 Introduction

The proliferation of smart sensors in recent years has introduced the possibility of accurately detecting and estimating a large new class of phenomena that affect society. These sensors, ranging from smart personal devices to more traditional purpose-built sensors, may be owned by a multitude of sources, and can produce qualitatively different readings which can be combined to make inferences about an event of interest.

In turn, this has led to the advent of crowd sensing, wherein a central data collector accrues the measurements made by a multitude sources, using these data points to generate a single cohesive estimate for some phenomena of interest to the data collector. However, the quality of this central estimate, and thus its value to the data collector, depends fundamentally on the ability, and moreover the willingness, of the data sources to produce accurate readings which are relevant to the phenomena the data collector wishes to study.

Unfortunately, there may be instances where data sources have some aversion to providing the data collector with the quality of estimates she desires. Take as an example, the case where the sensor must exert significant resources to produce an accurate reading (e.g. time or network bandwidth), or a situation where the source views the information she is sharing as private, and has incentive to obfuscate the data she shares [Bak+04; DR14]. Consequently, in order to ensure she consistently receives high quality measurements from the data sources, the central data collector must design an incentive mechanism which:

1. allows her to metricize the quality of the reading each data source provides, and

2. provides incentive for the data sources to produce readings which are considered "high quality" under this metric.

Given the wide range of applications and industries this problem affects, many different compensation mechanisms have been proposed to promote the production of high quality readings from a collection of data sources. An overview of such mechanisms is given in [Gao+15].

The contribution of this section can be seen as an extension of [Cai+15], in which the authors design a general payment mechanism, by which a central data collector may induce each data source in the marketplace to exert precisely the level of effort in collecting data that the central data buyer desires. The goal of the data buyer in this case is to obtain a high quality estimator for some phenomena using the readings from the data sources, while reducing the payments needed to incentivize the necessary exertion of effort from the sensors. Several other papers [Dob+16; Far+15] further investigate mechanisms of this sort, proposing several extensions.

However, it has yet to be studied how such mechanisms perform in situations where more than one central data buyer wishes to purchase readings from data sources in the marketplace. A number of important questions arise when such *data markets* are considered. If the central data buyers are competing companies, will they permit data sources to also sell information to their competitors? If the data buyers do purchase readings from the same set of data sources, who will foot the bill to incentivize the effort the data sources exert? Will the data buyers who provide larger payments to the data sources be compensated with higher quality readings than their competitors?

Most significantly, this section demonstrates that if all the data buyers design compensation schemes as proposed in [Cai+15], each of the data buyers will receive the same quality of reading from a particular data source, regardless of how much each data buyer personally compensates the data source for her effort. This leads to conflicting objectives for each of the data buyers on several fronts. If a data buyer wishes to induce a data source to exert a high level of effort, she must reconcile the fact that her competitors will also receive a high quality reading from this data source. Even in the case where the data buyers care little about the success of the other buyers in the marketplace, each data buyer still wants to incentivize the data sources to produce high quality readings, but wants to force the other data buyers to offer the lion's share of the necessary compensation.

In this section, we analyze the competitive outcomes that arise in such a marketplace by formulating a game between the buyers wherein they

1. compete by designing pricing mechanisms to affect the behavior of the data sources, and

2. design these mechanisms so as meet the personal objectives enumerated above.

We derive conditions for the existence of Nash Equilibria in this game when a particular form is assumed, and analyze the efficiency and equity of these outcomes. We demonstrate

through both analytical and numerical exercises that the outcomes of these games are often highly inefficient from a social standpoint, which motivates future work to design incentive mechanisms which more effectively handle competition between data buyers.

The rest of this section proceeds as follows. In Section 4.1.2, we lay out explicit mathematical structures for the data markets, strategic data sources, strategic data buyers, and the class of contracts we will consider between the sources and buyers. In Section 4.1.3, we analyze the game that forms between the buyers in the data market, and demonstrate that the outcome of this game is in many cases socially inefficient, and often times. Section 4.1.4 provides a numerical example which highlights the issues presented in Section 4.1.3. And finally, Section 4.1.5 prescribes an agenda for future work, with the aim of developing more refined incentive mechanism which do not suffer from the same shortcomings in the competitive setting.

4.1.2 Mathematical formulation

In this section we formulate our model for data markets. We first present our model for strategic data sources, and then strategic buyers who issue incentives to strategic data sources. Based on recent research [Cai+15], we use incentives with a particular payment structure. Then, we define our overall game, as well as a generalized Nash equilibrium for this game.

Data market

At a high level, a *data market* consists of a set $\mathcal{S} = \{1, ..., N\}$ of strategic data sources, and a set $\mathcal{B} = \{1, ..., M\}$ of strategic data buyers. Each data source i is equipped to generate an estimate of the function $f : \mathcal{D} \to \mathbb{R}$ at some data point $x_i \in \mathcal{D}$, and each data buyer $j \in \mathcal{B}$ wishes to use these readings to generate a personal estimator of f, which we will denote $\hat{f}^j$. Each buyer b_j is willing to form a contract with each data source $i \in \mathcal{S}$, which monetarily compensates i for the readings she produces, and we assume it is under the purview of j to define the structure of this contract.

One may think of $\mathcal{D}$ as a set of features or events the data buyers are capable of observing, in order to make a prediction about some phenomena. The value returned by the mapping f encapsulates the relationship between the observable features and the outcome of interest. We further assume that each of the data sources and buyers acts *strategically*; that is, each of these agents acts to maximize some expected personal return from her transactions in this marketplace. The following two subsections of the document provide an explicit mathematical formulation describing the behavior of the data sources and data buyers. The basis for these definitions comes directly from [Cai+15].

Strategic data sources

In this subsection, we define our model for strategic data sources. Intuitively, data sources provide data samples (x, y) whose variance depends on their effort. Thus, the more effort

exerted, the better the statistical estimation for any data buyer who receives the data. Additionally, we assume the data sources are effort-averse, i.e. all else equal, they prefer to exert minimal effort. Furthermore, the buyer has no direct way to verify the amount of effort exerted by the data source. Thus, we have an issue commonly referred to as *moral hazard*.

More formally, all data sources share some function $f : \mathcal{D} \to \mathbb{R}$, where f is the function which data buyers wish to estimate. One may think of $\mathcal{D}$ as a set of features or events the data buyers are capable of observing, in order to make a prediction about some phenomena. The value returned by the mapping f encapsulates the relationship between the observable features and the outcome of interest.

Each data source i has their own feature $x_i \in \mathcal{D}$ and their own cost-of-effort function $\sigma_i^2 : \mathbb{R} \to \mathbb{R}_+$. When data source i exerts effort $e_i \in \mathbb{R}$, they produce an estimate of the form:

$$y_i(e_i) = f(x_i) + \epsilon_i(e_i) \quad \epsilon_i(e_i) \sim N(0, \sigma_i^2(e_i))$$

Both x_i and σ_i^2 are common knowledge, but the effort e_i is private, as well as the the value $y_i(e_i)$ produced. We shall design contracts such that the data source i is incentivized to exert the 'correct' amount of effort (to be defined), and report y_i truthfully.

Data source i will receive a payment from each buyer for their data. For buyer j, let this payment, potentially random, be denoted p_i^j. We assume that the data source has a utility function of the following form, should they opt-in:

$$\mathbb{E}\left(\sum_{j \in \mathcal{B}} p_i^j\right) - e_i \tag{4.1.1}$$

If they opt-out, they will receive utility 0.

Note that this assumes that the data sources are risk-neutral, effort-averse, and must opt-in ex ante. Additionally, we assume the effort e_i can be normalized to be comparable to the payments.

Throughout the rest of this section, we shall often omit the argument e_i when context makes it evident.

Strategic data buyers

A strategic data buyer $j \in \mathcal{B}$ is an agent who wishes to construct the best estimator $\hat{f}^j$ for a function f. She optimizes a loss function across a class of estimators, which the data buyer is free to select. In general, different buyers need not fit models of the same type; for example, one data buyer may choose to generate her estimator via linear regression, while another data buyer constructs his estimator by fitting the data to a polynomial model of higher degree. Differences in the type of estimator data buyers use may be used to encapsulate competitive advantages one data buyer has over another. For a more thorough review of the technical requirements of these estimators, see [Cai+15].

Additionally, each data buyer j has a distribution F_j across $\mathcal{D}$, which denotes how much they value an accurate estimate at various points in $\mathcal{D}$.

In particular, let $\hat{f}^j_{(\vec{x},\ \vec{y^j})}$ denote the estimator that buyer j constructs, based on the location of the data sources, $\vec{x}$, and the reports she receives from the data sources, $\vec{y^j}$. (Here, $\vec{x} = (x_1, \ldots, x_N)$ and similarly $\vec{y^j}$ is the vector of y values reported to buyer j.)

Beyond any intrinsic utility buyer j experiences from increasing the quality of her estimator, j also wishes to construct an estimator that is better than the estimator constructed by her competitors, the other members of $\mathcal{B}$.

Each data buyer j commits to a payment function p^j_i to each data source $i \in \mathcal{S}$, where $p^j_i : \mathcal{D}^N \times \mathbb{R}^N \to \mathbb{R}$ may depend not only on the reading reported by data source i, but also the readings reported by the other members of $\mathcal{S}$, with consideration given to the location of the data sources. In particular, buyer j constructs her various contracts with the data sources so as to minimize:

$$J^j(\vec{p^j}, \vec{p^{-j}}) = \mathbb{E}\left[\left(\hat{f}^j_{\vec{x},\vec{y^j}}(x^*) - f(x^*)\right)^2 - \sum_{k\in -j} \delta^j_k \left(\hat{f}^k_{\vec{x},\vec{y^k}}(x^*) - f(x^*)\right)^2 + \eta^j \sum_{i\in\mathcal{S}} p^j_i(\vec{x}, \vec{y^j})\right] \tag{4.1.2}$$

The expectation in (4.1.2) is taken across $x^* \sim F_j$ as well as the randomness in the reported data $\vec{y^k}$ for $k \in \mathcal{B}$. (Recall that F_j weighs the importance data buyer j places on an an accurate estimator about different points $x^* \in \mathcal{D}$.) Here, as per typical game theory notation, we will let $-i$ denote $\mathcal{S} \setminus \{i\}$ and $-j$ denote $\mathcal{B} \setminus \{j\}$, and when $-i$ or $-j$ is used as a subscript, this denotes everyone else's variables, e.g. $\vec{p^{-j}}$ denotes the vector of payment plans of all the data buyers that are not j.

Here, $\delta^j_k \in [0, 1]$ parameterizes the level of competition between buyers j and k, and we assume this competition is symmetric so $\delta^j_k = \delta^k_j$. When $\delta^j_k = 0$, j is indifferent to the success of k, and competes with k only insofar as trying to determine who will pay to incentive the data sources. Meanwhile, $\delta^j_k = 1$ denotes a situation akin to a zero-sum game between data buyers j and k.

The parameter $\eta^j > 0$ denotes a conversion between dollar amounts allocated by the payment functions and the utility generated by the quality of the various estimators that are constructed.

In order for the objective expressed in (4.1.2) to be well defined, we assume that buyer j chooses to construct an estimator for which there exists a function g_j such that, for all distributions F^j over $\mathcal{D}$, $\vec{x}$, and $\vec{\sigma^2} \in \mathbb{R}^N$:

$$g_j(\vec{x}, F_j, \vec{\sigma^2}) = \mathbb{E}\left[\left(\hat{f}^j_{\vec{x},\vec{y^j}}(x^*) - f(x^*)\right)^2\right] \tag{4.1.3}$$

Here the $\vec{y^j}$ have variance $\vec{\sigma^2}$.

Finally, we assume that buyer j has knowledge of what class of estimator each of the other data buyers plans to use.[1]

[1]This is a heavy-handed assumption, given that competing data buyers are unlikely to inform their

The data buyers are interested in offering payment contracts to data sources. These contracts must be designed such that, for each data source i, when i selects her effort e_i to maximize to (4.1.1), given the payment contracts from all of the other data buyers:

$$\mathbb{E}\left[\sum_{j\in\mathcal{B}} p_i^j(\vec{y^j}(e_i))\right] - e_i \geq 0 \tag{4.1.4}$$

$$\mathbb{E}\left[p_i^j(\vec{y^j}(e_i))\right] \geq 0 \tag{4.1.5}$$

Note that (4.1.4) is an ex-ante constraint for data source i that i receives non-negative payoff in expectation. This depends on the payments of the other data buyers. The second is an ex-ante constraint that data source i never opts into any contract with negative payments.

We model the resulting competition between the data buyers, subject to these coupled constraints, as a generalized Nash equilibrium problem (GNEP) [Dor+13].

Definition 15. *Each player j from a finite set of players $\mathcal{B}$ aims to solve an optimization problem given by:*

$$BR(p^{-j}) = \arg\min_{p^j}\{J^j(p^j, p^{-j})|p^j \in \mathcal{M}^j(p^{-j})\} \tag{4.1.6}$$

$\mathcal{M}^j(p^{-j})$ is called the feasible set *for player j, which depends on the actions taken by the other players $-j$. A vector $p = (p^1, p^2, \ldots p^M)$ is called a* (generalized) Nash equilibrium (GNE) *if $p^j = BR(p^{-j})$ for all $j \in \mathcal{B}$, i.e. the p^j are simultaneously solutions to each players optimization* (4.1.6).

Having laid out the general formulation for this problem, in the final portion of this section we lay out the form of the payment contracts that we consider between the buyers and sellers.

Structure of payment contracts

In [Cai+15] the particular case where $|\mathcal{B}| = 1$ is analyzed, and no competition between buyers of data must be considered. Their work considers payment plans from the single buyer to each data source i of the form:

$$p_i(\vec{x}, \vec{y}) = c_i - d_i\left(y_i - \hat{f}_{(\vec{x},\vec{y^j})_{-i}}(x_i)\right)^2, \tag{4.1.7}$$

where $\hat{f}_{(\vec{x},\vec{y})_{-i}}(x_i)$, is the optimal estimate for $f(x_i)$ that the data buyer can construct from the readings reported by the data sources other than source i, and $c_i \geq 0, d_i \geq 0$ are scalars to be chosen strategically by the buyer. The authors of [Cai+15] demonstrate an algorithm for selecting c_i and d_i which allows the buyer to:

competitors how they intend to process the data supplied by the sources. However, this is keeping with the goal of the section, as we shall demonstrate that even when there is complete information between the buyers, inefficiencies still arise in the data market.

1. precisely incentive data source i to exert any level of effort $\bar{e}_i$ that the buyer desires (the authors can make $\bar{e}_i$ a dominant strategy for data source i), and

2. precisely compensate data source i for her effort ($\mathbb{E}p_i(y_i(\bar{e}_i), \vec{y}_{-i}(\vec{\bar{e}}_{-i})) = e_i$, making the contract tightly satisfy individual rationality constraints).

Our goal is to study how pricing schemes of this form perform in the more general case where $|\mathcal{B}| > 1$, and competition between multiple data buyers becomes a critical consideration. In particular, we assume the following form for each of the incentive mechanisms offered in the data market.

Assumption 12. *Consider a data buyer j and data source i. It is assumed that j offers i a payment function of the form*

$$p_i^j(\vec{x}, \vec{y}) = c_i^j - d_i^j \left(y_i - \hat{f}^j_{(\vec{x},\vec{y})_{-i}}(x_i) \right)^2, \tag{4.1.8}$$

in exchange for knowledge of y_i,where $c_i^j, d_i^j \geq 0$ are parameters that the buyer j is free to choose.

Note that these payments do not directly depend on the level of effort that any of the data sources exert, since the data buyers do not have a means to directly observe these values. The payments only depend on the data reported to them, and can be calculated by data buyers. Having defined the necessary structures for the data markets we wish to study, we are now ready to study the competitive equilibria that arise in these marketplaces.

First, we note that for any data source i, due to the form of the payment contract, they will report the same value to all data buyers.

Proposition 16. *Fix any data source i. Pick any vector of variances $\vec{\sigma^2}$ (one variance for each data buyer), and let $e = \max \{\tilde{e} : \sigma_i^2(\tilde{e}) = (\vec{\sigma^2})_j\}$, i.e. e is the minimum amount of effort for data source i to generate measurements of variance $\vec{\sigma^2}$. Then, data source i has higher payoff, defined by (4.1.1), by choosing variances $\sigma_i^2(e)$ for all j, than the payoff earned from providing each buyer j with data of variance $(\vec{\sigma^2})_j$.*

In other words, since the payment contract from each data buyer j is increasing (in expectation) with respect to effort, data source i will never have incentive to 'add noise' to a measurement once the effort has been exerted. Note this arises due to a fundamental quirk of the nature of data: once the data has been harvested, it is infinitely reproducible with negligible cost. Thus, for the rest of this section, we shall write $\vec{y}$ to denote the measurement reported to all data sources j.

4.1.3 Results

In this section, we analyze the behavior we can expect from each of the agents in the market place, by considering the game that forms between the members of $\mathcal{B}$ as they select the parameters in the contracts they offer to the data sources.

Adopting standard game-theoretic short-hand notion, we denote the set of pricing parameters buyer k selects by (c^k, d^k) , and we denote the choice of the pricing parameters of the other members of $\mathcal{B}$ by (c^{-k}, d^{-k}). From now on, we use the index k to single out a specific buyer, the index q to single out a data source, the index j to sum over a collection of buyers, and the index i (and sometimes l) to sum over a collection of sources.

We begin our analysis by determining under what conditions the data sources will accept the collection of contracts offered to them by the data buyers. Recall that data source q will accept all of the contracts offered by the data sources if and only if the ex-ante total payments are non-negative (4.1.4) and each data buyer's payment is non-negative ex-ante (4.1.5).

Let δ_x denote the probability measure that puts mass 1 at point x. Then, we may simplify (4.1.4) for a fixed q by noting that:

$$\mathbb{E}\left[\sum_{j\in\mathcal{B}} p_q^j(\vec{x},\vec{y})\right] = \sum_{j\in\mathcal{B}} c_q^j - \mathbb{E}\sum_{j\in\mathcal{B}} d_q^j\left(y_q - \hat{f}^j_{\vec{x}_{-q},\vec{y}_{-q}}(x_q)\right)^2 =$$

$$\sum_{j\in\mathcal{B}} c_q^j - \sum_{j\in\mathcal{B}} d_q^j\left(\sigma_q^2(e_q) + g_j(\bar{x}_{-q}, \delta_{x_q}, \vec{\sigma^2}_{-q})\right)$$

Then, (4.1.4) holds if and only if:

$$\sum_{j\in\mathcal{B}} c_q^j - \sum_{j\in\mathcal{B}} d_q^j\left(\sigma_q^2(e_q) + g_j(\bar{x}_{-q}, \delta_{x_q}, \vec{\sigma^2}_{-q})\right) \geq e_q \tag{4.1.9}$$

Similarly, (4.1.5) holds if and only if:

$$c_q^j \geq d_q^j\left(\sigma_q^2(e_q) + g_j(\bar{x}_{-q}, \delta_{x_q}, \vec{\sigma^2}_{-q})\right) \tag{4.1.10}$$

As our goal is to find situations where the buyers receive data from each of the data sources, we shall include equations (4.1.9) and (4.1.10) as constraints in the game between data buyers. Indeed, given a choice of (c^{-k}, d^{-k}), the objective of buyer k is to optimize the following problem:

$$\min_{c^k, d^k} \quad J^k((c^k, d^k), (c^{-k}, d^{-k})) \tag{4.1.11}$$

$$\text{subject to} \quad \mathbb{E}\left[\textstyle\sum_{j\in\mathcal{B}} p_i^j(\vec{x}, \vec{y}(\vec{e^*}))\right] - e_i^* \geq 0 \text{ for all } i \in \mathcal{S} \tag{4.1.12}$$

$$e_i^* = \arg\max_{e_i} \mathbb{E}\left[\textstyle\sum_{j\in\mathcal{B}} p_i^j(\vec{x}, \vec{y}(\vec{e^*}))\right] - e_i \text{ for all } i \in \mathcal{S} \tag{4.1.13}$$

$$\mathbb{E}\left[p_i^k(\vec{x}, \vec{y}(\vec{e}))\right] \geq 0 \text{ for all } i \in \mathcal{S} \tag{4.1.14}$$

$$c_i^k \geq 0, d_i^k \geq 0 \text{ for all } i \in \mathcal{S} \tag{4.1.15}$$

Recall that J^k was defined in (4.1.2). Note that [Cai+15] showed that the payments induce dominant strategies, so (4.1.13) is an optimization that does not depend on e_{-i}.

In general, this may be a computationally difficult problem for b_k to solve. For illustrative purposes, for the rest of this section, we will assume specific forms for the estimators the buyers employ and the σ functions which define the data sources. We first assume:

Assumption 13. *For each data source i, $\sigma_i(e_i)$ is characterized by the the constant $\alpha_i > 0$ and of the form:*

$$\sigma_i(e_i) = \exp(-\alpha_i e_i) \tag{4.1.16}$$

Note that this implies that σ is convex, strictly decreasing and always positive, which are all desirable properties in our context. Furthermore, note that this is the form of the standard deviation, not the variance.

We next determine the level of effort data sources will exert given the pricing parameters set by the data buyers. Fix a data source q and taking the derivative of (4.1.1) with respect to e_q, we obtain:

$$-2\left(\sum_{j\in\mathcal{B}} d_q^j\right)\sigma_i(e_q)\frac{d}{de_q}\sigma_q(e_q) - 1 = 2\left(\sum_{j\in\mathcal{B}} d_q^j\right)\alpha_q \exp(-2\alpha_q e_q) - 1$$

Setting this derivative equal to 0 yields:

$$e_q^* = \frac{\ln\left(2\left(\sum_{j\in\mathcal{B}} d_q^j\right)\alpha_q\right)}{2\alpha_q} \tag{4.1.17}$$

This is the optimum effort selection for data source q. We can also compute how this optimal point varies with d_i^j:

$$\frac{\partial}{\partial d_q^j} e_q^* = \frac{1}{2\left(\sum_{j\in\mathcal{B}} d_q^j\right)\alpha_q}$$

Also we can easily calculate the optimum variance:

$$\sigma_q^2(e_q^*) = \frac{1}{2\left(\sum_{j\in\mathcal{B}} d_q^j\right)\alpha_q} \tag{4.1.18}$$

Assumption 14. (Separable estimators) *For each buyer $k \in \mathcal{B}$, the estimator for f that buyer k employs, $\hat{f}^k$, is separable. In other words, there exists a function h_k such that:*

$$g_k(\vec{x}, F, \vec{\sigma^2}) = \sum_{i\in\mathcal{S}} h_k(x_i, \vec{x}, F)\sigma_i^2$$

Furthermore, we assume that $h \geq 0$.

Note that linear regression, polynomial regression and finite-kernel regression all produce separable estimators. Applying Assumption 14 for the estimators, we may rewrite the loss function for buyer k as:

$$J^k((c^k,d^k),(c^{-k},d^{-k})) = \sum_{i\in\mathcal{S}} h_k(x_i,\vec{x},F_k)\sigma_i^2(e_i^*) - \sum_{j\in -k}\delta_j^k \sum_{i\in\mathcal{S}} h_j(x_i,\vec{x},F_j)\sigma_i^2(e_i^*) +$$

$$\eta^k \sum_{i\in\mathcal{S}} \left(c_i^k - d_i^k \left[\sigma_i^2(e_i^*) + \sum_{l\in -i} h_k(x_l,\vec{x}_{-i},\delta_{x_i})\sigma_l^2(e_l^*) \right] \right)$$

Recall that each x_i is fixed and common knowledge; thus, we can replace each of the above evaluations of the h functions with constants. Define $\beta_i^j = h_j(x_i,\vec{x},F_j)$, $\xi_{i,l}^j = h_j(x_l,\vec{x}_{-i},\delta_{x_i})$ for $i\neq l$ and $\xi_{i,i}^j = 1$. Note that $\xi \geq 0$. Then, this becomes:

$$J^k((c^k,d^k),(c^{-k},d^{-k})) =$$

$$\sum_{i\in\mathcal{S}} \beta_i^k\sigma_i^2(e_i^*) - \sum_{j\in -k}\delta_j^k\sum_{i\in\mathcal{S}}\beta_i^j\sigma_i^2(e_i^*) + \eta^k\sum_{i\in\mathcal{S}}\left(c_i^k - d_i^k\left[\sigma_i^2(e_i^*) + \sum_{l\in -i}\xi_{i,l}^k\sigma_l^2(e_l^*)\right]\right) =$$

$$\sum_{i\in\mathcal{S}}\left(\beta_i^k - \sum_{j\in -k}\delta_j^k\beta_i^j\right)\sigma_i^2(e_i^*) + \eta^k\sum_{i\in\mathcal{S}}\left(c_i^k - d_i^k\left[\sigma_i^2(e_i^*) + \sum_{l\in -i}\xi_{i,l}^k\sigma_l^2(e_l^*)\right]\right)$$

In efforts towards succinctness, let $\gamma_i^k = \beta_i^k - \sum_{j\in -k}\delta_j^k\beta_i^j$. We will now plug in the expression for $\sigma_i^2(e_i^*)$ in (4.1.18), yielding:

$$J^k((c^k,d^k),(c^{-k},d^{-k})) = \sum_{i\in\mathcal{S}}\gamma_i^k\sigma_i^2(e_i^*) + \eta^k\sum_{i\in\mathcal{S}}\left(c_i^k - d_i^k\left[\sigma_i^2(e_i^*) + \sum_{l\in -i}\xi_{i,l}^k\sigma_l^2(e_l^*)\right]\right) =$$

$$\sum_{i\in\mathcal{S}}\frac{\gamma_i^k}{2\left(\sum_{j\in\mathcal{B}} d_i^j\right)\alpha_i} + \eta^k\sum_{i\in\mathcal{S}}\left(c_i^k - d_i^k\left[\frac{1}{2\left(\sum_{j\in\mathcal{B}} d_i^j\right)\alpha_i} + \sum_{l\in -i}\frac{\xi_{i,l}^k}{2\left(\sum_{j\in\mathcal{B}} d_l^j\right)\alpha_l}\right]\right) =$$

$$\sum_{i\in\mathcal{S}}\frac{\gamma_i^k}{2\left(\sum_{j\in\mathcal{B}} d_i^j\right)\alpha_i} + \eta^k\sum_{i\in\mathcal{S}}\left(c_i^k - d_i^k\left[\sum_{l\in\mathcal{S}}\frac{\xi_{i,l}^k}{2\left(\sum_{j\in\mathcal{B}} d_l^j\right)\alpha_l}\right]\right)$$

(Note here we joyfully take advantage of our convention that $\xi_{i,i}^k = 1$.)

Finally, similar reasoning lets us write for any data source q and data buyer k:

$$\mathbb{E}\left[p_q^k(\vec{x},\vec{y})\right] = c_q^k - d_q^k\left(\sigma_q^2(e_q) + g_k(\bar{x}_{-q},\delta_{x_q},\vec{\sigma^2}_{-q})\right) =$$

$$c_q^k - d_q^k\left(\sigma_q^2(e_q) + \sum_{i\in -q} h_k(x_i,\vec{x}_{-q},\delta_{x_i})\sigma_i^2(e_i)\right) = c_q^k - d_q^k\left(\sum_{i\in\mathcal{S}}\xi_{q,i}^k\sigma_i^2(e_i)\right)$$

At optimum effort levels, this becomes:

$$\mathbb{E}\left[p_q^k(\vec{x},\vec{y})\right] = c_q^k - d_q^k\left(\sum_{i\in\mathcal{S}}\xi_{q,i}^k\sigma_i^2(e_i^*)\right) = c_q^k - d_q^k\left(\sum_{i\in\mathcal{S}}\frac{\xi_{q,i}^k}{2\left(\sum_{j\in\mathcal{B}}d_i^j\right)\alpha_i}\right)$$

Also using the expression for e_i^* given in (4.1.17), buyer k has the following optimization problem:

$$\min_{c^k,d^k} \quad \sum_{i\in\mathcal{S}}\frac{\gamma_i^k}{2d_i^{total}\alpha_i} + \eta^k\sum_{i\in\mathcal{S}}\left(c_i^k - d_i^k\left[\sum_{l\in\mathcal{S}}\frac{\xi_{i,l}^k}{2d_l^{total}\alpha_l}\right]\right) \tag{4.1.19}$$

$$\text{subject to} \quad \sum_{j\in\mathcal{B}}\left[c_i^j - d_i^j\left(\sum_{l\in\mathcal{S}}\frac{\xi_{i,l}^j}{2d_l^{total}\alpha_l}\right)\right] - \frac{\ln\left(2d_i^{total}\alpha_i\right)}{2\alpha_i} \geq 0 \tag{4.1.20}$$

$$c_i^k - d_i^k\left(\sum_{l\in\mathcal{S}}\frac{\xi_{i,l}^k}{2d_l^{total}\alpha_l}\right) \geq 0 \tag{4.1.21}$$

$$d_i^{total} = \sum_{j\in\mathcal{B}}d_i^j \tag{4.1.22}$$

$$c_i^k \geq 0, d_i^k \geq 0 \tag{4.1.23}$$

Every constraint above holds for all $i \in \mathcal{S}$. Here, (4.1.22) is a definitional, rather than binding, constraint. Also, note that without loss of generality, we can take $\eta^k = 1$, by normalizing the γ_i^k accordingly. Additionally, we can remove the constraint $c_i^k \geq 0$, as it is redundant in light of the constraint $c_i^k - d_i^k\left(\sum_{l\in\mathcal{S}}\frac{\xi_{i,l}^k}{2d_l^{total}\alpha_l}\right) \geq 0$, since $\xi \geq 0$ and $d \geq 0$.

This leads to the following result.

Theorem 6. *Consider the game where each buyer's objective is to solve the optimization in (4.1.19), and assume $\gamma_i^j \geq 0$ for all $i \in \mathcal{S}, j \in \mathcal{B}$. Then there are either an infinite number of generalized Nash equilibria, or there is no generalized Nash equilibrium.*

Furthermore, in the case where there are an infinite number of generalized Nash equilibria, there is a unique collection of d parameters, in the sense that if $(\vec{c},\vec{d})$ and $(\vec{c'},\vec{d'})$ are both generalized Nash equilibria, then $\vec{d} = \vec{d'}$. Additionally, the c parameters lie in the convex polytope defined by the following constraints:

$$\sum_{j\in\mathcal{B}}c_i^j = \sum_{j\in\mathcal{B}}d_i^j\left(\sum_{l\in\mathcal{S}}\frac{\xi_{i,l}^j}{2d_l^{total}\alpha_l}\right) + \frac{\ln\left(2d_i^{total}\alpha_i\right)}{2\alpha_i}$$

$$c_i^k \geq d_i^k\left(\sum_{l\in\mathcal{S}}\frac{\xi_{i,l}^k}{2d_l^{total}\alpha_l}\right)$$

The effort exerted by each data source is the same in each generalized Nash equilibrium.

Before proving this theorem, we discuss the assumption that $\gamma_i^j \geq 0$. This implies that, for each data buyer, the penalty for other data buyer's successful estimation does not outweigh

the benefit of having a good estimator. This assumption means that no data buyer will have incentive to drive the variance of one data source up towards infinity.

We prove the following useful lemma, and then prove our theorem.

Lemma 4. *Suppose $(\vec{c}, \vec{d})$ is a GNE for the game defined by (4.1.19). The following equality holds for all i and k:*

$$c_i^k = \sum_{j \in \mathcal{B}} d_i^j \left(\sum_{l \in \mathcal{S}} \frac{\xi_{i,l}^j}{2 d_l^{total} \alpha_l} \right) + \frac{\ln\left(2 d_i^{total} \alpha_i\right)}{2\alpha_i} - \sum_{j \in -k} c_i^j$$

In other words, (4.1.20) is always tight in equilibrium.

Proof. To prove this, note that, by the cost function of buyer k, c_i^k will always be chosen such that at least one of (4.1.20) and (4.1.21) is tight. Suppose (4.1.21) is exclusively active, i.e.

$$c_i^k - d_i^k \left(\sum_{l \in \mathcal{S}} \frac{\xi_{i,l}^k}{2 d_l^{total} \alpha_l} \right) = 0$$

$$c_i^k > \sum_{j \in \mathcal{B}} d_i^j \left(\sum_{l \in \mathcal{S}} \frac{\xi_{i,l}^j}{2 d_l^{total} \alpha_l} \right) + \frac{\ln\left(2 d_i^{total} \alpha_i\right)}{2\alpha_i} - \sum_{j \in -k} c_i^j \tag{4.1.24}$$

Note that (4.1.24) is the same constraint for every data buyer. In other words, if it is loose for k, it is loose for all other j. Thus, some other buyer j can reduce their c_i^j and lower their cost, and thus $(\vec{c}, \vec{d})$ cannot be an equilibrium.

This argument does fall apart in one situation, however. No buyer can reduce their cost just by modifying c if (4.1.21) is tight for all buyers k, i.e. for all k:

$$c_i^k - d_i^k \left(\sum_{l \in \mathcal{S}} \frac{\xi_{i,l}^k}{2 d_l^{total} \alpha_l} \right) = 0$$

In this case, (4.1.20), which we assumed held loosely, becomes $2 d_i^{total} \alpha_i < 1$. Let buyer k increase d_i^k such that $2 d_i^{total} \alpha_i = 1$, and then choose a new c^k such that (4.1.21) holds tightly, i.e. $c^k = d_i^k \left(\sum_{l \in \mathcal{S}} \frac{\xi_{i,l}^k}{2 d_l^{total} \alpha_l} \right)$. Note that this decreases their cost:

$$\sum_{i \in \mathcal{S}} \frac{\gamma_i^k}{2 d_i^{total} \alpha_i} > \sum_{i \in \mathcal{S}} \gamma_i^k$$

(This uses the fact that, since (4.1.21) holds for all buyers j, the second term disappears.) Additionally, all the constraints of the original optimization are still satisfied, so (c_i^k, d_i^k) was not an optimizer for buyer k.

This concludes our proof. □

Proof. (Theorem 6) We invoke Lemma 4 and substitute this into the objective function, (4.1.19), for buyer k. This yields:

$$\begin{aligned} \min_{c^k, d^k} \quad & \textstyle\sum_{i\in\mathcal{S}} \left(\frac{\gamma_i^k}{2d_i^{total}\alpha_i} + \sum_{j\in -k} \left(d_i^j \left(\sum_{l\in\mathcal{S}} \frac{\xi_{i,l}^j}{2d_l^{total}\alpha_l} \right) - c_i^j \right) + \frac{\ln\left(2d_i^{total}\alpha_i\right)}{2\alpha_i} \right) \\ \text{subject to} \quad & c_i^k - d_i^k \left(\textstyle\sum_{l\in\mathcal{S}} \frac{\xi_{i,l}^k}{2d_l^{total}\alpha_l} \right) \geq 0 \\ & d_i^{total} = \textstyle\sum_{j\in\mathcal{B}} d_i^j \\ & d_i^k \geq 0 \end{aligned}$$

We quickly manipulate the cost function a little to a more desirable form:

$$\sum_{i\in\mathcal{S}} \left(\frac{\gamma_i^k}{2d_i^{total}\alpha_i} + \sum_{j\in -k} \left(d_i^j \left(\sum_{l\in\mathcal{S}} \frac{\xi_{i,l}^j}{2d_l^{total}\alpha_l} \right) - c_i^j \right) + \frac{\ln\left(2d_i^{total}\alpha_i\right)}{2\alpha_i} \right) =$$

$$\sum_{i\in\mathcal{S}} \left(\frac{\gamma_i^k}{2d_i^{total}\alpha_i} + \frac{\ln\left(2d_i^{total}\alpha_i\right)}{2\alpha_i} \right) + \sum_{i\in\mathcal{S}} \sum_{j\in -k} \sum_{l\in\mathcal{S}} \frac{d_i^j \xi_{i,l}^j}{2d_l^{total}\alpha_l} - \sum_{i\in\mathcal{S}} \sum_{j\in -k} c_i^j =$$

$$\sum_{i\in\mathcal{S}} \left(\frac{\gamma_i^k}{2d_i^{total}\alpha_i} + \frac{\ln\left(2d_i^{total}\alpha_i\right)}{2\alpha_i} \right) + \sum_{i\in\mathcal{S}} \sum_{j\in -k} \sum_{l\in\mathcal{S}} \frac{d_l^j \xi_{l,i}^j}{2d_i^{total}\alpha_i} - \sum_{i\in\mathcal{S}} \sum_{j\in -k} c_i^j =$$

$$\sum_{i\in\mathcal{S}} \left(\frac{\gamma_i^k}{2d_i^{total}\alpha_i} + \sum_{j\in -k} \left(\sum_{l\in\mathcal{S}} \frac{d_l^j \xi_{l,i}^j}{2d_i^{total}\alpha_i} - c_i^j \right) + \frac{\ln\left(2d_i^{total}\alpha_i\right)}{2\alpha_i} \right)$$

Note the index swap on the ξ terms in the second equality. Then define:

$$J_i^k(d_i^k, c^{-k}, d^{-k}) = \frac{\gamma_i^k}{2d_i^{total}\alpha_i} + \sum_{j\in -k} \left(\sum_{l\in\mathcal{S}} \frac{d_l^j \xi_{l,i}^j}{2d_i^{total}\alpha_i} - c_i^j \right) + \frac{\ln\left(2d_i^{total}\alpha_i\right)}{2\alpha_i} =$$

$$\frac{\gamma_i^k + \sum_{j\in -k} \sum_{l\in\mathcal{S}} d_l^j \xi_{l,i}^j}{2d_i^{total}\alpha_i} - \sum_{j\in -k} c_i^j + \frac{\ln\left(2d_i^{total}\alpha_i\right)}{2\alpha_i}$$

Thus, the overall optimization can again be re-written:

$$\begin{aligned} \min_{c^k, d^k} \quad & \textstyle\sum_{i\in\mathcal{S}} J_i^k(d_i^k, c^{-k}, d^{-k}) \\ \text{subject to} \quad & c_i^k - d_i^k \left(\textstyle\sum_{l\in\mathcal{S}} \frac{\xi_{i,l}^k}{2d_l^{total}\alpha_l} \right) \geq 0 \\ & d_i^{total} = \textstyle\sum_{j\in\mathcal{B}} d_i^j \\ & d_i^k \geq 0 \end{aligned}$$

We differentiate the cost with respect to d_q^k:

$$\frac{\partial}{\partial d_q^k}\sum_{i\in\mathcal{S}} J_i^k(d_i^k, c^{-k}, d^{-k}) = \frac{\partial}{\partial d_q^k} J_q^k(d_q^k, c^{-k}, d^{-k}) = -\frac{\gamma_q^k + \sum_{j\in -k}\sum_{l\in\mathcal{S}} d_l^j \xi_{l,q}^j}{2(d_q^{total})^2\alpha_q} + \frac{1}{2d_q^{total}\alpha_q} =$$

$$\frac{-\gamma_q^k - \sum_{j\in -k}\sum_{l\in\mathcal{S}} d_l^j \xi_{l,q}^j + d_q^{total}}{2(d_q^{total})^2\alpha_q} = \frac{-\gamma_q^k - \sum_{j\in -k}\sum_{l\in -q} d_l^j \xi_{l,q}^j + d_q^k}{2(d_q^{total})^2\alpha_q}$$

Note that we use the fact that $\xi_{q,q}^j = 1$ for all j. It is easy to see that:

$$\frac{\partial}{\partial d_q^k} J_q^k(d_q^k, c^{-k}, d^{-k}) \begin{cases} < 0 & \text{if } 0 \le d_q^k < \gamma_q^k + \sum_{j\in -k}\sum_{l\in -q} d_l^j \xi_{l,q}^j \\ = 0 & \text{if } d_q^k = \gamma_q^k + \sum_{j\in -k}\sum_{l\in -q} d_l^j \xi_{l,q}^j \\ > 0 & \text{if } d_q^k > \gamma_q^k + \sum_{j\in -k}\sum_{l\in -q} d_l^j \xi_{l,q}^j \end{cases}$$

Thus, the maximizing d_q^k is given by:

$$d_q^k = \gamma_q^k + \sum_{j\in -k}\sum_{l\in -q} d_l^j \xi_{l,q}^j \tag{4.1.25}$$

Performing this analysis for all combinations of $q \in \mathcal{S}$ and $k \in \mathcal{B}$ yields a system of $M \times N$ equations with $M \times N$ unknowns, of the form (4.1.25).

As we have before, let $\vec{d}$ denote a column vector with entries d_i^j for each $i \in \mathcal{S}$ and $j \in \mathcal{B}$. Similarly, let $\vec{\gamma}$ denote a column vector containing all the terms of the form γ_i^j. Then, we may represent this system of equations with the following matrix equation:

$$\vec{d} = A\vec{d} + \vec{\gamma} \tag{4.1.26}$$

Here, A is a non-negative matrix whose entries are the values of the various ξ parameters at the appropriate places, such that (4.1.26) expresses the set of equality constraints defined by (4.1.25) for all $q \in \mathcal{S}$ and $k \in \mathcal{B}$. To find an GNE of this game, it suffices to find a solution to (4.1.26) such that $d_i^j \ge 0$ for all i and j.

Systems of equations of this form are well studied in the economics literature, as they are of the form specified by the celebrated Leontief input-output model. It has been shown that such systems of equations have a non-negative solution if and only if $\rho(A) < 1$, where $\rho(A)$ is the spectral radius of A [Sta+06]. Moreover, if such a solution exists, it must be unique.

Thus, if $\rho(A) < 1$, inversion of this A matrix yields the equilibrium $\vec{d}$, and, by Lemma 4, we can pick any $\vec{c}$ that satisfies:

$$\sum_{j\in -\mathcal{B}} c_i^j = \sum_{j\in\mathcal{B}} d_i^j \left(\sum_{l\in\mathcal{S}} \frac{\xi_{i,l}^j}{2d_l^{total}\alpha_l}\right) + \frac{\ln\left(2d_i^{total}\alpha_i\right)}{2\alpha_i}$$

$$c_i^k \ge d_i^k \left(\sum_{l\in\mathcal{S}} \frac{\xi_{i,l}^k}{2d_l^{total}\alpha_l}\right)$$

If $\rho(A) \geq 1$, there will not exist a non-negative solution and there is no point $(\vec{c}, \vec{d})$ that simultaneously optimizes (4.1.19) for all k. It follows that there is either a unique $\vec{d}$ that will constitute a Nash solution for the game, which produces a convex polytope of potential GNE, or there this no solution to the game, as desired. □

It is interesting to note that the existence of GNE depends solely on the value of the ξ parameters; it does not depend on the magnitude of the γ parameters. This implies that the existence or non-existence of GNE in this game is simply an artifact of the incentive mechanisms we have chosen to analyze, and does not depend on whether or not there are solutions that are beneficial to all parties involved. Note that we chose this incentive mechanism based on several desirable properties in the single-buyer case; whether or not there exist mechanisms that extend to multi-buyer games in a fashion that provides good efficiency properties is an open problem that we are currently investigating. In Section 4.1.4, we calculate the ξ parameters for a specific example, and see how equilibrium solutions in these marketplaces collapse as the characteristics are varied.

Additionally, note that, in the case where there is a continuum of GNE, the effort exerted by data sources and $\vec{d}$ parameters are the same across all equilibria. The ambiguity arises in the $\vec{c}$ parameters. In other words, the ambiguity arises in determining which data buyers will pay to ensure that each data source's total compensation covers the cost of their effort. In the extreme case, it is possible for one firm to pay for the entirety of the expected compensation offered to the data sources, while the the firms pay nothing on expectation. That is, for some $k \in \mathcal{B}$, $\sum_{i\in\mathcal{S}} p_i^k(\vec{x}, \vec{y}) = \sum_{i\in\mathcal{S}} e_i^*$, and for all $j \neq k$, $\sum_{i\in\mathcal{S}} p_i^j(\vec{x}, \vec{y}) = 0$. In Section 4.1.5 we discuss possible mechanisms to alleviate the disparity that may arise in these situations.

We next turn to analyzing the total utility experienced in the marketplace for a given outcome of the game. We begin with the following definition.

Definition 16. (Ex-ante social loss of the data market) *Suppose that $\eta^j = 1$ for all buyers j. Let $\vec{e}$ be the vector denoting the level of effort the data sources exert. Then, we define the* ex-ante social loss *the marketplace to be the sum of the utility functions of all the data buyers and data sources:*

$$\mathcal{L}(\vec{e}) = \sum_{j\in\mathcal{B}} \left(\mathbb{E}\left[\left(\hat{f}^j_{\vec{x},\vec{y^j}}(x^*) - f(x^*) \right)^2 - \sum_{k\in -j} \delta_k^j \left(\hat{f}^k_{\vec{x},\vec{y^k}}(x^*) - f(x^*) \right)^2 \right] \right) + \sum_{i\in\mathcal{S}} e_i$$

Note that this sum does not include any of the payments made in the marketplace, as they are simply lossless transfers of wealth. We require the additional assumption that $\eta^j = 1$ for all buyers j to ensure that these transfers of wealth are lossless from a utility perspective, i.e. the buyers and sources value the payment equally. This assumption allows us to isolate the social loss due to the mechanism, and ignore any losses due to differential preferences in payment currency.

Theorem 7. *Suppose that Assumptions 13 and 14 hold. Further, assume that $\gamma_i^j > 0$, for all $i \in \mathcal{S}$ and $j \in \mathcal{B}$, and that $\xi_{i,l}^j > 0$ for some $i, l \in \mathcal{S}$, $j \in \mathcal{B}$. Finally, suppose GNE*

solutions exist for the game, and let $\vec{e^*}$ denote the unique level of effort exerted by the data sources across each of these GNE solutions, as stipulated by Theorem 6. Then, there exists $\vec{\hat{e}} \in \mathbb{R}^N$ such that $\mathcal{L}\left(\vec{\hat{e}}\right) < \mathcal{L}\left(\vec{e^*}\right)$. Furthermore, the socially optimal levels of effort, $\vec{\hat{e}}$, are always less than the induced levels of effort at equilibrium $\vec{e^*}$.

Proof. We begin by calculating solving for the value of $\vec{\hat{e}}$ which minimizes the value of $\mathcal{L}\left(\vec{\hat{e}}\right)$. Invoking Assumption 14 and our definition of γ, we may write:

$$\mathcal{L}\left(\vec{\hat{e}}\right) = \sum_{i\in\mathcal{S}}\sum_{j\in\mathcal{B}} \gamma_i^j \sigma_i^2(\hat{e}_i) + \sum_{i\in\mathcal{S}} \hat{e}_i \tag{4.1.27}$$

Taking the derivative with respect to $\hat{e}_q$ and repeating our analysis with Assumption 13, and setting the resulting equation to zero we obtain:

$$-2\alpha_q \left(\sum_{j\in\mathcal{B}} \gamma_q^j\right) \exp(-2\alpha_q \hat{e}_q) + 1 = 0$$

We can re-arrange this to yield:

$$\hat{e}_q = \frac{\ln(2\alpha_q \sum_{j\in\mathcal{B}} \gamma_q^j)}{2\alpha_q} \tag{4.1.28}$$

Note, that $\mathcal{L}$ is strictly convex with respect to $\hat{e}_q$, so choosing the entries of $\hat{\vec{e}}$ according to equation (4.1.28) must produce the unique minimizer of $\mathcal{L}\left(\vec{\hat{e}}\right)$.

Next we compare this to the level of effort the sources produce in the GNE of the game between the buyers. By (4.1.25), we obtain that in the GNE for all $i \in \mathcal{S}$ and $k \in \mathcal{B}$:

$$d_i^k = \gamma_i^k + \sum_{j\in -k}\sum_{l\in -i} d_l^j \xi_{l,i}^j \geq \gamma_i^k > 0$$

Furthermore, since there exists at least one $\xi > 0$, we know that $d_q^k > \gamma_q^k$ for some data source q and buyer k. It follows that, for this particular q:

$$\sum_{j\in\mathcal{B}} d_q^j > \sum_{j\in\mathcal{B}} \gamma_q^j$$

Thus, by (4.1.17), we see that

$$e_q^* = \frac{\ln(2\alpha_q \sum_{j\in\mathcal{B}} d_q^j)}{2\alpha_q} > \frac{\ln(2\alpha_q \sum_{j\in\mathcal{B}} \gamma_q^j)}{2\alpha_q} = \hat{e}_q. \tag{4.1.29}$$

Thus the theorem is proved, since it must be the case that $\mathcal{L}\left(\vec{\hat{e}}\right) < \mathcal{L}\left(\vec{e^*}\right)$ since we chose $\vec{\hat{e}}$ to be the unique minimizer of $\mathcal{L}$. □

Theorem 7 shows that there is always some social loss, ex-ante, from a Nash solution compared to the social optimum. Furthermore, the proof provides a way to identify where this loss is incurred, and how to calculate how much is lost. Note that the social welfare is always lost because the effort induced in equilibrium is higher than is socially optimal. This captures the intuition that each data buyer has a negative externality: they wish to improve their estimates without considering how their improved estimates hurt other data buyers.

The proof itself also provides strong intuition on the ξ parameters between buyers. Loosely speaking, these ξ parameters can be thought of as a measure of each buyer's 'market power', in the sense that it quantifies how much one buyer can influence the payment contracts of other buyers in the data market to his advantage. As an extremal case, when $\xi^j_{i,l} = 0$ for all $i, l \in \mathcal{S}$ and $j \in \mathcal{B}$, there is no coupling between the payments the buyers make, and the social optimum coincides with the Nash solution.

4.1.4 Example: Between two firms

In this section, we present an example which demonstrates how a data market may collapse as the parameters of the system are varied. This example will also demonstrate how the efficiency of the data market, in terms of the ex-ante social loss function $\mathcal{L}$, changes as the market approaches this collapse. In particular, we consider the case where there are two data sources (s_1 and s_2) and two firms acting as data buyers (b_1 and b_2). Each of the data sources is capable of estimating the function

$$f\colon [-1,\ 1] \to \mathbb{R}. \tag{4.1.30}$$

Let $x_1,\ x_2 \in [-1, 1]$ denote the locations where s_1 and s_2 sample f, respectively. Assume that each of the data sources are as defined in Assumption 13, with the characteristic parameters $\alpha_1 = \alpha_2 = 1$.

Next, we assume that each of the data buyers is performing linear regression on f, using the samples reported by the data source. In this case [Cai+15]:

$$g_j(\vec{x}, F_j, \sigma^2(\vec{e})) = \mathbb{E}_{x^* \sim F_j}\left[\begin{bmatrix} x^* \\ 1 \end{bmatrix}^T (X^TX)^{-1}X^T \cdot \operatorname{diag}(\sigma_1^2(e_1), \sigma_2^2(e_2)) \cdot X(X^TX)^{-1}\begin{bmatrix} x^* \\ 1 \end{bmatrix}\right] =$$

$$\gamma_1^j\sigma_1^2(e_1) + \gamma_2^j\sigma_2^2(e_2)$$

In this example, we assume $F_1 = F_2$ as the uniform distribution on the domain of f, $[-1, 1]$. Thus, for $i \in \{1, 2\}$:

$$\gamma_i^1 = \gamma_i^2 = \frac{(x_1 - x_2)^2/3 + (x_i^2 - x_1x_2)^2}{(x_1^2 + x_2^2 - 2x_1x_2)^2}$$

Note that, by these assumptions, $g_1 = g_2$, and furthermore:

$$\xi_{1,2}^1 = \xi_{1,2}^2 = g(x_2, \delta_{x_1}, \sigma_2^2(e_2)) = \frac{(x_1x_2 + 1)^2}{(x_2^2 + 1)^2}$$

$$\xi^1_{2,1} = \xi^2_{2,1} = g(x_1, \delta_{x_2}, \sigma^2_1(e_1)) = \frac{(x_1x_2+1)^2}{(x_1^2+1)^2}$$

For illustrative purposes, we fix $x_2 = 1$, and see what happens to the data market as we vary x_1 along the interval $[-1, 1]$.

Note that, when $x_1 = x_2 = 1$, it is no longer possible to construct a linear estimator of f because there is insufficient data. Thus, the example shows how the game between buyers behaves as it becomes increasingly difficult to construct good estimators. The $\vec{d}$ parameters of any Nash solution can be found by solving:

$$\begin{bmatrix} \gamma^1_1 \\ \gamma^2_1 \\ \gamma^1_2 \\ \gamma^2_2 \end{bmatrix} = \underbrace{\begin{bmatrix} 1 & 0 & 0 & -\xi^2_{2,1} \\ 0 & 1 & -\xi^1_{2,1} & 0 \\ 0 & -\xi^2_{1,2} & 1 & 0 \\ -\xi^1_{1,2} & 0 & 0 & 1 \end{bmatrix}}_{B} \begin{bmatrix} d^1_1 \\ d^2_1 \\ d^1_2 \\ d^2_2 \end{bmatrix} \tag{4.1.31}$$

Note that this B matrix is equal to $I - A$ as defined in the proof. We numerically solve this system of equations for varying values of $x_1 \in [-1, 1]$, and the results are shown in Figures 4.1 through 4.4.

Figures 4.1 and 4.2, demonstrate how the γ and ξ parameters of the game change as a function of x_1. Figure 4.3 demonstrates the d parameters that the buyers will offer the data sources as x_1 varies. And finally, Figure demonstrates the price of anarchy in the data market, as a function of x_1 which is given by:

$$\frac{\mathcal{L}\left(\vec{e^*}\right)}{\mathcal{L}\left(\vec{\hat{e}}\right)}$$

Here, $\vec{e}^*$ is the induced effort of the sensors in the Nash solution of the game between data buyers, and $\vec{\hat{e}}$ is the socially optimal effort for data sources to exert. Further comments in the captions of Figures 4.3 and 4.4 demonstrate the inefficiencies that arise in this example.

4.1.5 Closing remarks

We've analyzed the game that forms between a set of data buyers when they wish to communally incentivize a collection of strategic data sources, using a mechanism that has been proposed in the literature. We derived, for a particular form of the game, conditions for the existence of GNE, and demonstrated that these solutions are frequently socially inefficient. This motivates future work to develop a richer class of incentive mechanisms which alleviate these issues. Possible solutions include more complex pricing mechanisms, or perhaps the addition of a trusted third party market-maker to mediate socially beneficial transactions in these data markets.

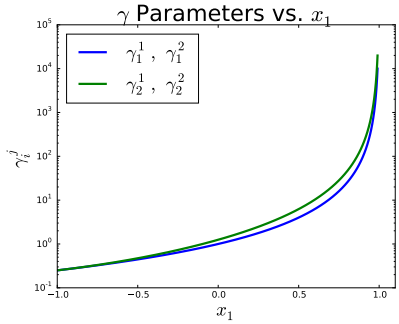

Figure 4.1: This figure depicts how the various γ parameters of the system vary as a function of x_1. Note that as $x_1 \to 1$, γ diverges to infinity, which reflects the fact that as x_1 and x_2 become increasingly close it becomes more difficult to generate a linear estimator from samples at these data points.

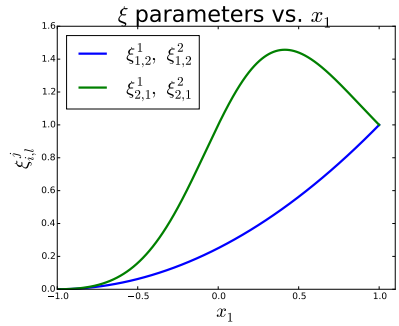

Figure 4.2: This figure depicts how the various ξ parameters vary as a function of x_1. As all of the ξ parameters converge to a value of 1 as $x_1 \to 1$, the matrix B in equation (4.1.31) becomes singular, causing the breakdown of solutions for the d parameters, as is depicted in Figure 4.3.

4.2 Optimal causal imputation for control

The widespread applicability of analytics in cyber-physical systems has motivated research into causal inference methods. Predictive estimators are not sufficient when analytics are used for decision making; rather, the flow of causal effects must be determined. Generally speaking, these methods focus on estimation of a causal structure from experimental data. In this section, we consider the dual problem: we fix the causal structure and optimize over causal imputations to achieve desirable system behaviors for a minimal imputation cost. First, we present the optimal causal imputation problem, and then we analyze the problem in two special cases: 1) when the causal imputations can only impute to a fixed value, 2) when the causal structure has linear dynamics with additive Gaussian noise. This optimal causal imputation framework serves to bridge the gap between causal structures and control.

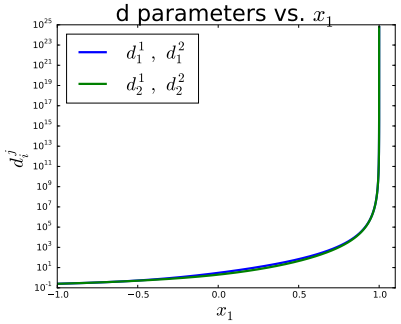

Figure 4.3: This figure depicts the Nash equilibrium $\vec{d}$ parameters for the game between the buyers as a function of x_1. Note that, as $x_1 \to 1$, the $\vec{d}$ parameters go off to infinity, and the Nash equilibria between the buyers breaks down. Comparing these results to Figure 4.1, we see that the $\vec{d}$ parameters diverge much more quickly than the γ parameters, meaning that in the Nash equilibria to the game between the two buyers becomes increasingly inefficient as $x_1 \to 1$.

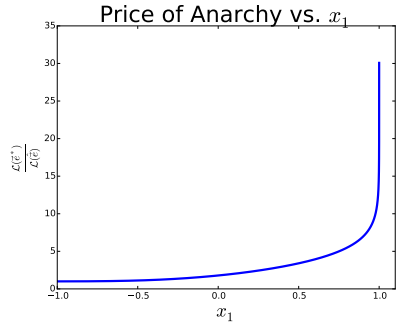

Figure 4.4: This figure depicts the price of anarchy for the marketplace as a function of x_1. When $x_1 = -1$, the payments in the marketplace are decoupled and the ξ parameters are all zero; in this instance the price of anarchy is 1, and the market is perfectly efficient. However, as $x_1 \to 1$, the price of anarchy diverges asymptotically to infinity, and the marketplace becomes increasingly inefficient as it becomes more difficult for the buyers to construct the estimators they desire.

4.2.1 Introduction

Recently, data analytics have achieved amazing levels of success. As analytics penetrate more and more industrial applications, they are increasingly used for decision-making and planning. In these applications, it is important to use estimators that are not only predictive, but estimate the causal structure of the underlying processes.

Correlation is not the same as causation. However, in practice, it is not always easy to apply this principle. In many real-life applications, machine learning is used to determine the

relationship between two variables. This analysis is often used as the basis for determining which actions to take. However, an algorithm with low test error does not necessarily mean that the causal effect has been estimated.

For example, one may train a classifier to estimate the energy consumption of a household given the presence and absence of eco-friendly devices, and this may provide guidelines for which devices should be discounted through rebate programs. Unless the causal structures are explicitly accounted for, there could easily be confounding variables or incorrect causal relationships that change the behavior of the system under consideration.

This has motivated new interest in causal inference techniques. Generally speaking, these techniques take experimental data and attempt to uncover the causal structure. (We defer a literature review of these methods to Section 4.2.3, when a more formal model of causality has been developed.) In this section, we consider the dual problem: we fix the causal structure and attempt to determine what causal actions will lead to system behaviors we desire at a minimal cost.

4.2.2 Background

There are three main paradigms for the mathematical modeling of causality:

1. Rubin causality
2. Granger causality
3. Pearl's structural equation modeling (SEM)

Each of these paradigms has a vast literature in its own right; we will try to present a few representative samples from each field here. Note that each paradigm uses its own notation, so we will change notation as we switch from approach to approach.

It should be noted that these paradigms are not mutually exclusive: for example, a problem that is modeled using Granger causality can be put into Pearl's SEM if the underlying processes operate in discrete time. Rubin causality can often be phrased as an SEM problem, but in applications this will require more structural assumptions to learn the causal structure. A full exposition of the intersections and non-intersections of these three paradigms is outside the scope of this section, but we note that these paradigms can often model the same phenomena and shed different insights on the causal behaviors observed.

Rubin causality was first introduced in [Rub74]. In the basic formulation of Rubin causality, we are given some control variable X taking values in $\{0, 1\}$. There are also two distinct random variables Y_0 and Y_1. If $X = 0$, then we observe Y_0 and not Y_1. If $X = 1$, then we only observe Y_1, and not Y_0. Another way to write this notationally is that we observe Y_X but do not observe Y_{1-X}, which is often called the counterfactual. The fact that we can only observe one or the other, but not both, is the fundamental misery of causality.

One of the key results that the Rubin causality paradigm provides is that if X is independent of Y_0 and Y_1, then randomly assigning $X \in \{0, 1\}$ yields a dataset that can

provide valid estimates of the counterfactuals; thus, Rubin causality provides the theoretical foundation for randomized control trials. This paradigm has also been extended to consider many covariates [IR15], handle confounding variables and incorporate instrumental variables [IR15], and incorporate some machine learning approaches [AI15b]. Sample applications include estimating the causal effect of residential demand response in the Western United States [Zho+16] or the causal effects of providing money, healthcare and education to the very poor in Ethiopia, Ghana, Honduras, India, Pakistan, and Peru [Ban+15].

Granger causality was first introduced in [Gra69]. In this paradigm, we are given data from two stationary random processes X and Y, both indexed by time. First, let U_t denote all the information available in the universe at time t, and let $(U - X)_t$ denote all the information available at time t except for X. Then, let $\sigma^2(Y|U)$ denote the error variance of the unbiased, least-squares estimator of Y_t using U_t, and similarly let $\sigma^2(Y|U - X)$ denote the error variance of the unbiased, least-squares estimator of Y_t using $(U - X)_t$. Then, X Granger-causes (or G-causes, for short) Y if $\sigma^2(Y|U) < \sigma^2(Y|U - X)$, i.e. the estimator that utilizes X has lower variance on its error than the one that cannot. In other words, X has explanatory power for Y.

Granger causality essentially relies on the relationship between causal effects and the arrow of time to distinguish it from general correlations. Although this framework does not address many of the more pernicious philosophical aspects of causality, oftentimes prior knowledge allows us to make the inductive leap from time-lagged correlations to causality. This paradigm is particularly appealing because it is easy to calculate in practice. Sample applications include determining which neuron assemblies Granger-cause other neuron assemblies to fire synapses [Bro+04] or finding that exchange rates Granger-cause stock market prices in Asia [Gra+00].

Pearl's SEM approach to causality models the statistical relationship between random elements with a Bayesian network [Pea09]. Bayesian networks are directed acyclic graphs, such that the distribution of a random element at node i only depends on the values taken at the parent nodes. This is meant to model causal relationships between nodes in the graph. Pearl defines the imputation operator as follows: if one imputes at a node i, one disconnects i from all its parents and deterministically sets its value to some fixed, predetermined constant. We will be building on this approach in this section, so we will defer the formal development of Pearl's SEM until Section 4.2.3.

At a high level, the imputation operator captures a lot of our intuitions about how the subjunctive conditional should function. When one says *If it had rained today, I would have brought my umbrella*, what does one mean? Intuitively, one often means: 'If everything else were the same, only it is the case that it is raining today instead of sunny, these are the actions I would have taken.' One does not mean that the world is structured in a way such that the necessary processes to induce rain today were instead the case. In other words: causal imputation does not travel upstream, e.g. backwards through time. This is captured in Pearl's SEM.

More practically, consider the question: *What are the causal effects of this medication?* If we wish to estimate this, we should 'set' medication taken to `TRUE`, and see the consequences

of this imputation. If we do not explicitly 'set' this value, then the decision to take medication is a consequence of preceding factors. This makes it difficult to determine if the observed effects are a result of the medication or some other confounding variables[2]. Again, this will be more formally discussed in Section 4.2.3.

Thus, we can think of these paradigms in terms of the central phenomenon it is designed to model. In summary:

1. Rubin causality is focused on the estimation of the *counterfactual*.
2. Granger causality is focused on the *explanatory power* one process provides over another process.
3. Pearl's SEM is focused on the causal effects of the *imputation* operator.

Throughout this section, we use Pearl's SEM. However, we note again that oftentimes problems framed in the Rubin causality or Granger causality paradigm often can be translated to an equivalent formulation in SEM.

Notation

For any set A, we denote the powerset of A as 2^A, which can also be thought of as the set of functions mapping $A \to \{0, 1\}$. For a collection of sets $\{A_i\}_{i \in I}$, we denote the Cartesian product as $\prod_{i \in I} A_i$.

Also, I will denote the identity matrix, where context will often be sufficient to determine its dimensions.

We let $U[a, b]$ denote the uniform distribution on the interval $[a, b]$ and $N(\mu, \Sigma)$ to denote the multivariate Gaussian distribution with mean μ and covariance matrix Σ.

4.2.3 Causal framework

In this section, we introduce our framework for modeling causal effects, and then define the problem of optimal causal imputation.

Causal structure

We build on the structural equation modeling framework presented in [Pea09]. First, we will introduce Bayesian networks.

Definition 17. *A* directed graph $G = (V, E)$ *is a set of nodes* V *and a set of edges* $E \subset V \times V$*. Throughout this section we will assume* V *is at most countably infinite.*

A path *from* $v_0 \in V$ *to* $v_N \in V$ *is a finite sequence of edges:*

$$(v_0, v_1), (v_1, v_2), \ldots, (v_{N-1}, v_N) \in E$$

[2]We note that similar reasoning can be done in the Rubin causality formulation as well.

We define the parents *of node i as* $\text{pa}(i) = \{j : (j,i) \in E\}$.

We can iterate this relationship to define the ancestor relationship. Let:

$$\text{pa}^n(i) = \{j : k \in \text{pa}^{n-1}(i), (j,k) \in E\}$$

Here, $\text{pa}^1(i) = \text{pa}(i)$ *is as defined above. Then, the* ancestors *of a node i are given by:*

$$\text{anc}(i) = \cup_{n=1}^{\infty} \text{pa}^n(i)$$

We say j is a descendant *of i if $i \in \text{anc}(j)$.*

A directed graph is acyclic *if $i \notin \text{anc}(i)$ for every $i \in V$. We will refer to such graphs as* directed acyclic graphs (DAGs).

Definition 18. *A* random process *X indexed by a set V is a collection of random elements $(X_i)_{i \in V}$. We will let $\mathcal{X}_i$ denote the possible values of X_i, and $\mathcal{X} = \prod_{i \in V} \mathcal{X}_i$.*

When there is an associated graph $G = (V, E)$, we will use the notation $\text{pa}(X_i)$ *to denote the tuple $(X_j)_{j \in \text{pa}(i)}$.*

Definition 19. *A random process X indexed by V is* Markov relative *to a DAG $G = (V, E)$ if its distribution factorizes:*

$$P(X) = \prod_{i \in V} P(X_i | \text{pa}(X_i))$$

We can also say that X and G are compatible*, or G* represents *X.*

This formalization will serve as our model for causality. The interpretation is that if there is an edge going from i to j, then X_i *causes* X_j.

Throughout this section, we will treat the causal structure $G = (V, E)$ as given. Estimation of this causal structure is a non-trivial task, and an active topic of research. Some approaches to the task of causal inference include: using metrics like directed information to estimate the causal strength between random variables [Gou+87; AM12], graphical-model based methods for estimating structure between random variables [Pea98; Lau01; Li+15; AI15a], and regression based approaches [Daw00; Hec+06; Hoy+09]. Again, this list is far from exhaustive as an extensive literature review of this general field is outside the scope of this section. For a broader overview of various approaches to the problem of causal inference, see [Pea98; Pea09].

Although the estimation of causal structures is never a simple task, the growing field of research promises more and more applications in which accurate estimation of causal structures is feasible.

Previous work has focused on the *estimation* of causal structures. In contrast, our contribution is to consider the problem of *control* of causal structures. In other words, once we are given a causal structure, how can we impute causal effects to drive the overall system into a desirable state?

For example, once we can estimate the causal effects of issuing rebates for energy-efficiency appliances, how do we best distribute these rebates to induce more energy-efficient

consumption patterns? To the best of our knowledge, this is the first work to consider the problem of when and where to impute on a causal structure.

There is an equivalent formulation of the condition in Definition 19 which utilizes *disintegration* results in probability theory. This is referred to as the structural equation modeling framework in [Pea09].

Proposition 17. *[Kal02; Pea09] A random process X indexed by V is Markov relative to $G = (V, E)$ if and only if there exists a collection of functions $(f_i)_{i \in V}$ and independent random elements $(\xi_i)_{i \in V}$ such that:*

$$X_i = f_i(\mathrm{pa}(X_i), \xi_i) \tag{4.2.1}$$

Furthermore, if $\mathcal{X}_i$ are Borel spaces[3], then ξ_i can be taken to be $U[0, 1]$.

We note that Borel spaces are a very general category of measurable spaces: they include Polish spaces equipped with the Borel σ-algebra[4]. This includes finite sets, $\mathbb{R}$, $\mathbb{R}^n$, $L^p(\mathbb{R}^n)$, the set of p-integrable functions defined on $\mathbb{R}^n$. Additionally, the space of probability distributions on any Borel space is also a Borel space.

Assumption 15. *Throughout the rest of this section, we will always use X to denote a random process indexed by V that is Markov relative to a DAG $G = (V, E)$, where X_i takes values in $\mathcal{X}_i$. Similarly, f_i shall denote the functions as specified in Equation 4.2.1, and similarly ξ_i.*

Causal imputation

In this section, we will formally define the *causal imputation* operation. Intuitively, imputation of X produces a new random process Y. This random process Y is equal to X prior to the causal imputation, is forced to some value at the node of imputation, and experiences causal effects after the node of imputation. This is formally defined below.

Definition 20. *[Pea09] A random process Y indexed by V is the* imputation *of X at $i \in V$ to a constant $x_i \in \mathcal{X}_i$ if:*

- $Y_i = x_i$.
- *For any j that is not a descendant of i, $Y_j = X_j$.*
- *For any j that is a descendant of i, $Y_j = f_j(\mathrm{pa}(Y_j), \xi_j)$.*

If this is the case, we will write $Y = \mathrm{do}(X; i, x_i)$.

[3] A measurable space S is Borel if there exists a measurable function $S \to [0, 1]$ with a measurable inverse.

[4] A topological space T is Polish if it is separable and completely metrizable. The Borel σ-algebra of a topological space is the smallest σ-algebra containing all the open sets.

The imputation operator produces a copy of the original process that is exactly equal at all nodes that do not causally depend on the node of imputation X_i. At the point of imputation, the node is disconnected from its parents and forced a constant value x_i. The nodes X_j that causally depend on X_i are replaced with new values that depend on the causal effects of X_i, keeping the innovation terms ξ constant throughout.

Referring back to the discussions in Section 4.2.2, this can be thought of as manually setting the value of X_i to x_i. This should be something that is done exogenously, as a control variable, rather than as a consequence of endogenous factors: this is why Y_i is disconnected from $\text{pa}(Y_i)$.

From this definition, it immediately follows that the imputation operator commutes.

Proposition 18. *Let $i, j \in V$ such that $i \neq j$ and $x_i \in \mathcal{X}_i$ and $x_j \in \mathcal{X}_j$. Then:*

$$\text{do}(\text{do}(X; i, x_i); j, x_j) = \text{do}(\text{do}(X; j, x_j); i, x_i) \text{ almost surely}$$

This allows us to define imputation on any set of nodes, rather than just at a single node.

Definition 21. *For any $I \subset V$ and $x_I \in \prod_{i \in I} \mathcal{X}_i$, we define the imputation $Y = \text{do}(X; I, x_I)$ as the sequential application of element-wise* do *operations. This is almost surely unique by Proposition 18.*

Optimal causal imputation

In the previous section, we defined the causal imputation operator. We can think of our system designer as having the capacity of issuing control commands that have causal effects on the system downstream. When we can define the cost of imputation as well as a control objective, we can formulate the optimal causal imputation problem.

We suppose we are given a collection of functions $(c_I)_{I \subset V}$ where each $c_I : \prod_{i \in I} \mathcal{X}_i \to \mathbb{R}$. These functions can be interpreted as the cost of imputation at a set of nodes $I \subset V$. Drawing on our running example, c represents the cost of issuing rebates for eco-friendly refrigerators at a set of households.

Furthermore, we suppose we are given an operational objective in the form of a cost function $g : \mathcal{X} \to \mathbb{R}$. For example, g can be a penalty on energy-wasting consumption patterns.

Definition 22. *The problem of* optimal causal imputation *is given by:*

$$\min_{I \subset V} \min_{x_I \in \prod_{i \in I} \mathcal{X}_i} \quad c_I(x_I) + \mathbb{E}_Y[g(Y)] \tag{4.2.2}$$

$$\text{subject to} \quad Y = \text{do}(X; I, x_I) \tag{4.2.3}$$

4.2.4 Applications

In Section 4.2.3, we defined the optimal causal imputation problem in its full generality. In this section, we shall provide methods to solve the optimal causal imputation problem in

special cases. In particular, we consider two contexts: 1) situations where imputation is only allowed to a single value, 2) situations where the dynamics are linear-Gaussian. In both instances, we shall assume $\mathcal{X}_i = \mathbb{R}^{n_i}$ for some n_i.

Single-value case

In many applications where we can causally impute values, we can only impute to one particular value. For example, when issuing incentives, we may be able to only offer one form of rebate to consumers. Motivated by this context, we consider situations where the optimal causal imputation problem can be reduced to one of submodular optimization.

Assumption 16. *In this section, we assume V is a finite set and that for each $I \subset V$, there exists an x_I such that $c_I(x_I) < \infty$ and $c_I(x'_I) = \infty$ for any $x'_I \neq x_I$. We shall refer to this as the single-value case.*

In the single-value case, we use the shorthand $F(I) = c(I) + \mathbb{E}[g(\text{do}(X; I))]$, where we drop dependencies on x as it can only take a single value.

Submodular minimization

Definition 23. *The set mapping $F : 2^V \to \mathbb{R}$ is* submodular *if for any $I_1 \subset I_2 \subset V$ and $i \in G \setminus I_2$, we have:*

$$F(I_1 \cup \{i\}) - F(I_1) \geq F(I_2 \cup \{i\}) - F(I_2) \tag{4.2.4}$$

Intuitively, this definition is motivated by economies of scale. We often expect economies of scale from these imputations, e.g. the per-customer cost of a rebate is non-increasing as the number of customers increases, due to bulk-purchase discounts. In our running example, the additional cost of issuing a rebate to customer i is higher when you have issued few rebates than when you have issued a lot of rebates. (When $I_1 \subset I_2$, then I_2 corresponds to the situation where you have issued more rebates than I_1.)

From a combinatorial optimization perspective, submodularity is a very well-behaved property that makes optimization, or approximate optimization, very tractable. We shall quickly outline the details now, but we refer the interested reader to [Sch03] for more details.

First, note that there is a very direct correspondence between a subset $I \subset V$ and a tuple in $\{0,1\}^V$. For example, if $V = \{0,1,2\}$, then $(0,1,1)$ corresponds to the subset $\{1,2\}$. Thus, we can think of $F : \{0,1\}^V \to \mathbb{R}$. Now, we define the Lovász extension [Lov83].

Definition 24. *Let $\lambda \sim U[0,1]$. Then, for any set mapping $F : \{0,1\}^V \to \mathbb{R}$, we define the* Lovász extension *$f : [0,1]^V \to \mathbb{R}$ as:*

$$f(z) = \mathbb{E}_\lambda[F(\{i : z_i > \lambda\})]$$

For the rest of this section, an unindexed f will denote the Lovász extension of F.

We note two nice properties of the Lovász extension immediately.

Proposition 19. *[Lov83] For any $z \in \{0,1\}^V$, we have $f(z) = F(z)$.*

Proposition 20. *[Lov83] F is submodular if and only if f is convex.*

Note that the optimal causal imputation problem can be written as:

$$\min_{z \in \{0,1\}^V} F(z)$$

The Lovász extension provides us with an easy solution to the problem.

Proposition 21. *[Lov83] If F is submodular, then the following is a convex optimization program.*

$$\min_{z \in [0,1]^V} f(z) \tag{4.2.5}$$

Furthermore, there exist minimizers of (4.2.5) in $\{0,1\}^V$.

In other words, the combinatorial optimization problem can be solved tractably with convex optimization if F is submodular. Thus, we are motivated in searching for conditions under which $F(I) = c(I) + \mathbb{E}[g(\text{do}(X; I))]$ is submodular. We provide a common sufficient condition for submodularity of F in the following theorem:

Theorem 8. *If:*

- *$g(Y) = \|Y_i - \mathbb{E}Y_i\|_2^2$ for some $i \in V$.*
- *There exists functions f_j^ξ such that, if X_j has no parents, $X_j = f_j^\xi(\xi_j)$ and otherwise $X_j = \text{pa}(X_j) + f_j^\xi(\xi_j)$.*
- *For each $j \in \text{anc}(i)$, there exists one unique path from j to i.*
- *$c(I)$ is submodular.*

Then $F(I) = c(I) + \mathbb{E}[g(\text{do}(X; I))]$ is submodular.

Note here that we treat $\text{pa}(X_i)$ as a vector in $\mathbb{R}^{n_i}$, where n_i is the appropriate dimension. These assumptions encompass many graphical models where a node's parents set a location parameter, and the control objective is the second moment of some feature.

Proof. Note that the desired result will follow if we show that the set mapping $G : I \mapsto \mathbb{E}[g(\text{do}(X; I))]$ is submodular, since the sum of submodular functions is submodular. Throughout this proof, we use i to refer to the index i pulled out by the function g.

We can see that $G(\emptyset) = \mathbb{E}[g(X)]$. By the independence of the $(\xi_i)_{i \in V}$ and the form of the $(X_i)_{i \in V}$ relationships, we can write this as $\mathbb{E}[g(X)] = \sum_{j \in \text{anc}(i)} \|f_j^\xi(\xi_j) - \mathbb{E}f_j^\xi(\xi_j)\|_2^2$. (Note that the unique path assumption ensures that each variance is only counted once in this sum.)

Algorithm 4 The greedy approach for combinatorial maximization.

$I \leftarrow \emptyset$
while $\max_{i:I\cup\{i\}\in S} F(I \cup \{i\}) - F(I) \geq 0$ **do**
  Pick $i^* \in \arg\max_{i:I\cup\{i\}\in S} F(I \cup \{i\})$
  $I \leftarrow I \cup \{i^*\}$
end while
  return I

More generally, we can write an expression for $G(I)$. Note that if we impute at a node j, all the uncertainty due to node j, and the ancestors of j, is zeroed out. Thus, we can write $G(I) = \mathbb{E}[g(X)] - \sum_{j\in(I\cup\text{anc}(I))} \|f_j^\xi(\xi_j) - \mathbb{E}f_j^\xi(\xi_j)\|_2^2$, where we define $\text{anc}(I) = \cup_{j\in I}\text{anc}(j)$.

Now, we can verify the submodularity condition on G. Pick $I_1 \subset I_2$ and $i' \in V \setminus I_2$. Then:

$$G(I_1 \cup \{i'\}) - G(I_1) =$$
$$\sum_{j\in(I_1\cup\text{anc}(I_1))} \|f_j^\xi(\xi_j) - \mathbb{E}f_j^\xi(\xi_j)\|_2^2 - \sum_{j\in(I_1\cup\{i'\}\cup\text{anc}(I_1\cup\{i'\}))} \|f_j^\xi(\xi_j) - \mathbb{E}f_j^\xi(\xi_j)\|_2^2 =$$
$$- \sum_{j\in\{i'\}\cup(\text{anc}(i')\setminus\text{anc}(I_1))} \|f_j^\xi(\xi_j) - \mathbb{E}f_j^\xi(\xi_j)\|_2^2$$

In words, the change in G due to adding i' to I_1 is the variances due to the terms related to i' and the ancestors of i' that have not already been zeroed out due to imputation, i.e. the ancestors of i' that are not already ancestors of I_1. A similar derivation can be done for I_2.

Thus, we can verify that $G(I_1 \cup \{i'\}) - G(I_1) \geq G(I_2 \cup \{i'\}) - G(I_2)$ by noting that $\text{anc}(i') \setminus \text{anc}(I_2) \subset \text{anc}(i') \setminus \text{anc}(I_1)$, so the right-hand side of the inequality adds more negative terms. This concludes our proof. □

Submodular maximization

Alternatively, suppose we are attempting to maximize a submodular function subject to a constraint, i.e. $F(I) = c(I) + \mathbb{E}[g(\text{do}(X; I))]$ subject to a constraint that $I \in S \subset 2^V$ and our objective is to solve $\max_{I\in S} F(I)$.[5]

First, consider the greedy method for submodular maximization. This is presented as Algorithm 4. At each iteration, it simply adds an element to I which maximizes $F(I \cup \{i\})$, if one exists. If one does not exist, it terminates and returns I. Under certain structural conditions, this algorithm yields approximate optimizers.

Definition 25. *A set mapping $F : 2^V \to \mathbb{R}$ is nondecreasing if $F(S) \leq F(T)$ whenever $S \subset T$.*

[5]Strictly speaking, to remain consistent with the problem in Section 4.2.3, we should be solving $\min_{I\in S} -F(I)$, but we express it as a maximization for clarity of presentation.

The monotonicity condition effectively prevents the algorithm from straying too far from the optimum when taking the greedy approach, as shown in [Nem+78]. Note that if F is non-decreasing, then the condition $\max_{i:I\cup\{i\}\in S} F(I\cup\{i\}) - F(I) \geq 0$ is equivalent to the existence of $i \in V$ such that $I \cup \{i\} \in S$.

Proposition 22. *[Nem+78] If F is nondecreasing and submodular, then the greedy method presented in Algorithm 4 will return $I^* \in S$ such that $F(I^*) \geq \left(\frac{e-1}{e}\right) \max_{I\in S} F(I)$.*

We now present a quick corollary of Theorem 8, which provides conditions under which we can leverage the existing results for maximization of nondecreasing submodular functions.

Corollary 7. *If:*

- $g'(Y) = -\|Y_i - \mathbb{E}Y_i\|_2^2$ *for some* $i \in V$.
- *There exists functions* f_j^ξ *such that, if* X_j *has no parents,* $X_j = f_j^\xi(\xi_j)$ *and otherwise* $X_j = \text{pa}(X_j) + f_j^\xi(\xi_j)$.
- *For each* $j \in \text{anc}(i)$, *there exists one unique path from* j *to* i.
- $c(I)$ *is nondecreasing and submodular.*

Then $F(I) = c(I) + \mathbb{E}[g'(\text{do}(X; I))]$ *is nondecreasing and submodular.*

Proof. This follows from Theorem 8 if we can show that $G' : I \mapsto \mathbb{E}[g(\text{do}(X; I))]$ is non-decreasing. Let $Y = \text{do}(X; I)$, and note that adding elements to I can only decrease the variance of Y_i. This can be formalized by noting, similar to the arguments in the proof of Theorem 8, $G'(I) = \mathbb{E}[g'(X)] + \sum_{j\in(I\cup\text{anc}(I))} \|f_j^\xi(\xi_j) - \mathbb{E}f_j^\xi(\xi_j)\|_2^2$. Thus, G', the additive inverse of the variance of Y_i, is nondecreasing. □

Note the minus sign in g' in Corollary 7: in most instances where you are maximizing a submodular cost, you would still wish to reduce uncertainty, i.e. have a lower variance.

Linear-Gaussian case

In this section, we consider causal imputation on a discrete-time linear dynamical system with Gaussian noise. That is, we analyze the special case of a random process with the form:

$$X_{t+1} = AX_t + \epsilon_t$$

Where $X_t \in \mathbb{R}^n$, $\epsilon_t \sim N(0, \sigma^2 I)$ independently for $t = 0, ..., T$, and $A \in \mathbb{R}^{n\times n}$ is a matrix representing the dependencies.

This process can be represented as a causal graph in the form of a trellis, where the random variables are all Gaussian. More specifically, each node has its expected value equal to a linear combination of their parents, as described by a matrix A, and additive noise of the distribution $N(0, \sigma^2)$.

To analyze our optimal casual imputation problem, we first redefine the indices for this problem. Since our causal graph represents a process over time, we index into the process by state k, for $k = 1, ...n$ as well as a time t for $t = 0, ..., T$. Thus X_{kt} indicates the value of state k at time t, and our graph has vertices $V = \{1, \dots, n\} \times \{0, \dots, T\}$. As before, X_t represents the value of the vector of all the states of X at time t, and we can think of X as a vector in $\mathbb{R}^{nT}$. We assume that the cost of imputation $c_I(x_I)$ has the following form for some parameters $\delta_i, q_i \geq 0$:

$$c_I(x_I) = \sum_{i \in I} \delta_i + q_i x_i^2$$

Further, we look at the case where the system cost of interest is minimizing the expected distance of the the random process from some target trajectory $\bar{y}$. Thus $g(Y) = \|Y - \bar{y}\|_2^2$.

Our optimal causal imputation problem in this case is thus:

$$\begin{aligned} &\min_{S \subset V} \min_{x_S \in \mathbb{R}^S} \sum_{i \in S} \left(\delta_i + q_i x_i^2\right) + \mathbb{E}\left[\|Y - \bar{y}\|_2^2\right] \\ &\text{subject to } Y = \text{do}(X; S, x_S) \end{aligned} \tag{4.2.6}$$

The summation term can be thought of as a cost of issuing control commands and the expectation term can be thought of as a trajectory tracking objective.

Given our structure on the random process, we can rewrite this optimization problem more concretely.

We first define $Q \in \mathbb{R}^{nT \times nT}$ to be diagonal matrix with the q_i's on the diagonal. We define $\delta \in \mathbb{R}^{nT}$ to be the vector of δ_i's. Further, let $\mathbf{1}_{nT}$ denote the column vector of all ones in $\mathbb{R}^{nT}$. Lastly, we define $\text{diag}(S)$ to be the square matrix with the elements of S on the diagonal, and zeros everywhere else.

The optimization in (4.2.6) now becomes:

$$\begin{aligned} \min_{\substack{S \in \{0,1\}^{nT} \\ \bar{x} \in \mathbb{R}^{nT}}} \quad & \bar{x}^T (Q + D)\bar{x} + \sigma^2 \text{Tr}(D I_S) + \delta^T S - 2\bar{y}^T \bar{x} \\ \text{subject to} \quad & (S_i - 1)\bar{x}_i = 0 \ \text{ for all } i = 1, \dots, nT \\ & P = (1 - \tilde{A})^{-1} \\ & D = P^T P \\ & I_S = I - \text{diag}(S) \end{aligned}$$

$$\tilde{A} = I_S \begin{bmatrix} 0 & 0 & 0 & \dots & 0 & 0 \\ A & 0 & 0 & \dots & 0 & 0 \\ 0 & A & 0 & \dots & 0 & 0 \\ 0 & 0 & A & \ddots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & A & 0 \end{bmatrix}$$

We note that for any matrix A and any S, the matrix $I - \tilde{A}$, with $\tilde{A}$ as defined above, is invertible, so P will always be well-defined.

Additionally, for a fixed S, the optimization across $\bar{x}$ is easy to solve. That is, D is entirely determined by S. If we let $(Q + D)_S$ denote the submatrix of $(Q + D)$ indexed by the non-zero elements of S, and similarly $\bar{x}_S$ and $\bar{y}_S$, then the optimizer is given by $\bar{x}^*_S = (Q + D)^{-1}_S \bar{y}_S$, with the other entries of $\bar{x}^*$ equal to 0.

Thus, we can easily calculate a set mapping $F(S)$ such that optimal causal imputation in the linear-Gaussian case is simply $\min_{S \subset V} F(S)$. We can solve this when nT is relatively small, and are currently investigating properties of $F(S)$ which would allow us to apply combinatorial optimization techniques [Sch03].

Closing remarks

The previous literature on mathematical formulations of causality has been focused on the *estimation* of causal structures. In this section, we presented the problem of *control* of causal structures. We formally defined the problem of optimal causal imputation, and formulate solutions for it in two cases: where imputation is allowed to only a single value, and the case where the dynamics are linear and the noise is Gaussian.

In future work, we hope to apply this framework to real situations which allow both the estimation of causal structures, as well as verification of the consequences and costs of imputation. Additionally, we hope to generalize our results to consider dynamical systems whose behavior are influenced by different features. For example, we can consider the dynamics of the power grid, but also account for frequently used machine learning features as well, such as the zip code of different energy consumers and the age of deployed assets.

We believe that considering the control aspects of causality is increasingly more relevant. In many smart infrastructure applications, we no longer have control commands that directly affect the dynamics, but rather our control actions act more like causal imputations. The optimal causal imputation framework is a promising direction to model these interactions between machine learning and control, and provides a model for closing the loop on analytics in cyber-physical systems.

4.3 Conclusion

In this chapter, we considered two problems: one of estimation in an IoT setting, and another of new control mechanisms in IoT settings.

First, we considered a data market where data is collected by effort-averse agents who need to be incentivized to provide high-quality data. Additionally, we introduce multiple buyers in this data market and analyze the equilibrium properties that arise from this formulation. Loosely speaking, the buyers strategically participate in a economic game against each other. From the joint action of all the buyers, a mechanism arises which is given to

the data sources, who then play a game against each other to maximize their own incentives while minimizing effort exerted.

Second, we considered how many of the new means of actuation available to IoT system operators are actually causal imputations. Rather than driving a dynamical system with an input in the traditional sense, many actions taken (e.g. a demand response event where users are told to curtail their energy usage, a semi-autonomous vehicle's warning light) have the causal effect of modifying distributions and propagating those modifications downstream in ways difficult to model with traditional differential equations. Furthermore, many of these 'actuation points' are often endogenous, and it is not always clear how to mathematically model the 'change' in an endogenous variable. Traditional control frameworks do not directly handle these situations, and we present a new framework for understanding the problem of optimal causal imputation.

The technological infrastructure of the Internet of Things is fundamentally changing many engineered systems. These changes cannot easily be modeled in many of the current paradigms for analysis, and require the development of new theoretical frameworks to capture the essence of the new dynamics and uncertainties. In this chapter, we covered examples of the new problems in estimation and control that arise as a result of the technologies of IoT.

Chapter 5

The End of the Thesis

In this document, we outlined several problems motivated by the Internet of Things (IoT), and their corresponding solutions. To develop these solutions, we have had to unify concepts from control theory, discrete optimization, system identification, probability theory, statistics, behavioral economics, information theory, model-predictive control, game theory, and graphical models. Insofar as IoT is the interconnection of a heterogenous set of devices and functionalities, the theoretical framework for the study and analysis of IoT will require the capacity to accommodate a very diverse set of mathematical tools and models.

Our work in new theoretical models for IoT has been focused on the role of data and the value of information. These new sensing capabilities are essential for painting a complete picture of the behaviors and functions of a interconnected smart city or smart town. However, there is currently a very thin line between a smart city and a surveillance city, and privacy issues naturally arise as IoT technologies are becoming more ubiquitous. The legislation and litigation on the privacy of new technologies is currently in a state of flux and it is very difficult to predict where it will land.

Our work has attempted to address these concerns with two research directions:

1. Designing systems in a fashion that retains the operational objectives of a smart city infrastructure while preserving the privacy of users.

2. Modeling data markets where privacy-conscious users selectively decide whether to honestly share their data, or whether a strategic play of data sharing is more beneficial.

As we researched these emerging problems in IoT settings, we increasingly realized the central role of the effect of human agents. Our vision of the impact of IoT is a new web of information flows, which include personal data, measurements of physical processes, and time-varying preferences. Whereas a lot of the literature takes it as given, our work focuses on considering the recorded and transmitted data as the outcome of human interactions.

We have considered the case where the data is privacy-sensitive or the decision of a economic utility maximizing agent, but there are broader classes of models for how this data comes to be part of the IoT system. That is, we can no longer think of all of IoT data as

independent and identically distributed samples from a fixed distribution. Rather, they are the outcome of games, they are dynamic and time-varying, and they are contextual.

Additionally, one of the contributions of this document is a unifying taxonomy by which to understand the work in privacy metrics and privacy-preserving mechanisms. However, an overarching framework by which to understand the literature on IoT data flows requires more generality and this is something that is a main focus of future work.

We view the Internet of Things as a phenomena in which new service models will emerge. Central to these service models will be the provided data and the conversations surrounding it. A technical analysis of the IoT systems and the statistical properties of their data, a behavioral analysis of the human actors who respond to IoT systems and participate by the revelation of their data (or lack thereof), and a game theoretic analysis of the data analytics companies who drive competitive data markets with market power are all components in a larger picture of the IoT as an emerging data market, and motivates much of the theoretical frameworks we have developed and plan to develop in future work.

Bibliography

[AC11] G. Acs and C. Castelluccia. "I Have a DREAM! (DiffeRentially privatE smArt Metering)". In: *Information Hiding*. Vol. 6958. Lecture Notes in Computer Science. Springer Berlin Heidelberg, 2011, pp. 118–132. DOI: 10.1007/978-3-642-24178-9_9 (cit. on pp. 30, 65).

[AF10] R. Anderson and S. Fuloria. *On the security economics of electricity metering*. Ninth Workshop on the Economics of Information. 2010 (cit. on p. 29).

[AI15a] S. Athey and G. Imbens. "Recursive Partitioning for Heterogeneous Causal Effects". In: *ArXiv e-prints* (Apr. 2015). arXiv: 1504.01132 (stat.ML) (cit. on p. 124).

[AI15b] S. Athey and G. W. Imbens. "Machine Learning Methods for Estimating Heterogeneous Causal Effects". In: *arXiV* (2015) (cit. on p. 122).

[AM+97a] K. Abed-Meraim, P. Loubaton, and E. Moulines. "A subspace algorithm for certain blind identification problems". In: *IEEE Transactions on Information Theory* 43.2 (1997), pp. 499–511. DOI: 10.1109/18.556108 (cit. on p. 19).

[AM+97b] K. Abed-Meraim, W. Qiu, and Y. Hua. "Blind system identification". In: *Proceedings of the IEEE* 85.8 (1997), pp. 1310–1322. DOI: 10.1109/5.622507 (cit. on p. 19).

[AM12] P. Amblard and O. J. J. Michel. "The relation between Granger causality and directed information theory: a review". In: *CoRR* abs/1211.3169 (2012). URL: http://arxiv.org/abs/1211.3169 (cit. on p. 124).

[Acq+15] A. Acquisti, L. Brandimarte, and G. Loewenstein. "Privacy and human behavior in the age of information". In: *Science* 347.6221 (2015), pp. 509–514. DOI: 10.1126/science.aaa1465 (cit. on p. 32).

[Ahm+12] A. Ahmed, B. Recht, and J. Romberg. "Blind Deconvolution using Convex Programming". In: *arXiv* (2012) (cit. on p. 19).

[Ale+08] M. Alexander, K. Agnew, and M. Goldberg. "New Approaches to Residential Direct Load Control in California". In: *2008 ACEEE Summer Study on Energy Efficiency in Buildings*. 2008 (cit. on p. 63).

[Ame] *ANSI/ASHRAE Standard 62.1-2013: Ventilation for Acceptable Indoor Air Quality*. American Society of Heating, Refrigerating and Air-Conditioning Engineers, 2013 (cit. on p. 93).

[And+13] M. E. Andrés, N. E. Bordenabe, K. Chatzikokolakis, and C. Palamidessi. "Geo-indistinguishability: Differential privacy for location-based systems". In: *Proceedings of the 2013 ACM SIGSAC conference on Computer & communications security*. ACM. 2013, pp. 901–914 (cit. on p. 81).

[Arm+13] K. C. Armel, A. Gupta, G. Shrimali, and A. Albert. "Is disaggregation the holy grail of energy efficiency? The case of electricity". In: *Energy Policy* 52 (2013), pp. 213–234. DOI: http://dx.doi.org/10.1016/j.enpol.2012.08.062 (cit. on p. 4).

[Asw+12] A. Aswani, N. Master, J. Taneja, D. Culler, and C. Tomlin. "Reducing transient and steady state electricity consumption in HVAC using learning-based model-predictive control". In: *Proceedings of the IEEE* 100.1 (2012), pp. 240–253 (cit. on p. 83).

[BC14] A. Beltran and A. E. Cerpa. "Optimal HVAC building control with occupancy prediction". In: *Proceedings of the 1st ACM Conference on Embedded Systems for Energy-Efficient Buildings*. ACM. 2014, pp. 168–171 (cit. on pp. 83, 86).

[BCB12] S. Bubeck and N. Cesa-Bianchi. "Regret Analysis of Stochastic and Nonstochastic Multi-armed Bandit Problems". In: *Foundations and Trends in Machine Learning* 5.1 (2012), pp. 1–122 (cit. on p. 51).

[BF02] E.-W. Bai and M. Fu. "A blind approach to Hammerstein model identification". In: *IEEE Transactions on Signal Processing* 50.7 (2002), pp. 1610–1619. DOI: 10.1109/TSP.2002.1011202 (cit. on p. 19).

[BT03] A. Beck and M. Teboulle. "Mirror Descent and Nonlinear Projected Subgradient Methods for Convex Optimization". In: *Oper. Res. Lett.* 31.3 (May 2003), pp. 167–175 (cit. on p. 51).

[BT89] D. P. Bertsekas and J. N. Tsitsiklis. *Parallel and Distributed Computation: Numerical Methods*. Upper Saddle River, NJ, USA: Prentice-Hall, Inc., 1989 (cit. on p. 25).

[Bai+02] E. Bai, Q. Li, and S. Dasgupta. "Blind identifiability of IIR systems". In: *Automatica* 38.1 (2002), pp. 181 –184. DOI: http://dx.doi.org/10.1016/S0005-1098(01)00179-0 (cit. on p. 19).

[Bak+04] D. E. Bakken, R. Rarameswaran, D. M. Blough, A. A. Franz, and T. J. Palmer. "Data obfuscation: anonymity and desensitization of usable data sets". In: *IEEE Security Privacy* 2.6 (2004), pp. 34–41. DOI: 10.1109/MSP.2004.97 (cit. on p. 101).

[Bal+13] B. Balaji, J. Xu, A. Nwokafor, R. Gupta, and Y. Agarwal. "Sentinel: occupancy based HVAC actuation using existing WiFi infrastructure within commercial buildings". In: *Proceedings of the 11th ACM Conference on Embedded Networked Sensor Systems*. ACM. 2013, p. 17 (cit. on pp. 80, 82).

[Ban+15] A. Banerjee, E. Duflo, N. Goldberg, D. Karlan, R. Osei, W. Parienté, J. Shapiro, B. Thuysbaert, and C. Udry. "A multifaceted program causes lasting progress for the very poor: Evidence from six countries". In: *Science* 348.6236 (2015). DOI: 10.1126/science.1260799 (cit. on p. 122).

[Ber+09] M. Berges, E. Goldman, H. S. Matthews, and L. Soibelman. "Learning systems for electric consumption of buildings". In: *ASCI International Workshop on Computing in Civil Engineering*. 2009 (cit. on p. 6).

[Ber+10] M. E. Berges, E. Goldman, H. S. Matthews, and L. Soibelman. "Enhancing electricity audits in residential buildings with nonintrusive load monitoring". In: *Journal of Industrial Ecology* 14 (2010), pp. 844–858 (cit. on p. 6).

[Ber09] C. Berry. *Residential Energy Consumption Survey*. Tech. rep. U.S. Energy Information Administration, 2009 (cit. on pp. 73, 74, 76).

[Blu+06] A. Blum, E. Even-Dar, and K. Ligett. "Routing without regret: on convergence to nash equilibria of regret-minimizing algorithms in routing games". In: *Proceedings of the twenty-fifth annual ACM symposium on Principles of distributed computing*. PODC '06. New York, NY, USA: ACM, 2006, pp. 45–52 (cit. on p. 44).

[Boy+11] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein. "Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers". In: *Found. Trends Mach. Learn.* 3.1 (Jan. 2011), pp. 1–122. DOI: 10.1561/2200000016 (cit. on p. 25).

[Bro+04] A. Brovelli, M. Ding, A. Ledberg, Y. Chen, R. Nakamura, and S. L. Bressler. "Beta oscillations in a large-scale sensorimotor cortical network: Directional influences revealed by Granger causality". In: *Proceedings of the National Academy of Sciences of the United States of America* 101.26 (June 2004), pp. 9849–9854. DOI: 10.1073/pnas.0308538101 (cit. on p. 122).

[But+15] D. J. Butler, J. Huang, F. Roesner, and M. Cakmak. "The Privacy-Utility Tradeoff for Remotely Teleoperated Robots". In: *ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. ACM. 2015, pp. 27–34 (cit. on p. 32).

[CBL06] N. Cesa-Bianchi and G. Lugosi. *Prediction, learning, and games*. Cambridge University Press, 2006 (cit. on p. 51).

[CH11] D. Callaway and I. Hiskens. "Achieving Controllability of Electric Loads". In: *Proc. of the IEEE* 99.1 (2011), pp. 184–199. DOI: 10.1109/JPROC.2010.2081652 (cit. on p. 67).

[CT12] T. M. Cover and J. A. Thomas. *Elements of information theory*. John Wiley & Sons, 2012 (cit. on pp. 88, 91).

[CT91] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley-Interscience, 1991 (cit. on pp. 31, 37, 38, 40).

[C+12] A. A. Cárdenas, S. Amin, G. Schwartz, R. Dong, and S. S. Sastry. "A game theory model for electricity theft detection and privacy-aware control in AMI systems". In: *Proc. of the 50th Allerton Conf. on Communication, Control, and Computing*. 2012, pp. 1830–1837. DOI: http://dx.doi.org/10.1109/Allerton.2012.6483444 (cit. on p. 61).

[Cai+15] Y. Cai, C. Daskalakis, and C. Papdimitriou. "Optimum Statistical Estimation with Strategic Data Sources". In: *JMLR: Workship and Conf. Proc.* Vol. 40. 2015, pp. 1–17 (cit. on pp. 102–104, 106, 108, 117).

[Cal09] D. S. Callaway. "Tapping the energy storage potential in electric loads to deliver load following and regulation, with application to wind energy". In: *Energy Conversion and Management* 50.5 (2009), pp. 1389 –1400. DOI: http://dx.doi.org/10.1016/j.enconman.2008.12.012 (cit. on pp. 65, 67, 68, 72).

[Cav11] A. Cavoukian. *Privacy by Design: Strong Privacy Protection - Now, and Well into the Future*. A Report on the State of PbD to the 33rd International Conference of Data Protection and Privacy Commissioners. Information and Privacy Commissioner of Ontario, 2011 (cit. on p. 64).

[DEE13] F. K. Dankar and K. El Emam. "Practicing Differential Privacy in Health Care: A Review." In: *Transactions on Data Privacy* 6.1 (2013), pp. 35–67 (cit. on p. 81).

[DH79a] W. Diffie and M. E. Hellman. "Privacy and authentication: An introduction to cryptography". In: *Proceedings of the IEEE* 67.3 (1979), pp. 397–427. DOI: 10.1109/PROC.1979.11256 (cit. on p. 61).

[DH79b] W. Diffie and M. E. Hellman. "Privacy and authentication: An introduction to cryptography". In: *Proceedings of the IEEE* 67.3 (1979), pp. 397–427 (cit. on p. 81).

[DK99] S. Drenker and A. Kader. "Nonintrusive monitoring of electric loads". In: *IEEE Computer Applications in Power* 12 (1999), pp. 47–51 (cit. on p. 7).

[DR14] C. Dwork and A. Roth. *The Algorithmic Foundations of Differential Privacy*. Foundations and Trends in Theoretical Computer Science, 2014 (cit. on pp. 33, 45, 48, 61, 101).

[Dal77] T. Dalenius. "Towards a Methodology for Statistical Disclosure Control". In: *Statistisk tidskrift* 15 (1977), pp. 429–444 (cit. on p. 30).

[Daw00] A. P. Dawid. "Causal Inference Without Counterfactuals". In: *Journal of the American Statistical Association* 95.450 (2000), pp. 407–424. URL: http://www.jstor.org/stable/2669377 (cit. on p. 124).

[Dob+16] D. G. Dobakhshari, N. Li, and V. Gupta. "An incentive-based approach to distributed estimation with strategic sensors". In: *Decision and Control (CDC), 2016 IEEE 55th Conference on*. IEEE. 2016, pp. 6141–6146 (cit. on p. 102).

[Don+10] B. Dong, B. Andrews, K. P. Lam, M. Höynck, R. Zhang, Y.-S. Chiou, and D. Benitez. "An information technology enabled sustainability test-bed (ITEST) for occupancy detection through an environmental sensing network". In: *Energy and Buildings* 42.7 (2010), pp. 1038–1046 (cit. on p. 80).

[Don+13a] R. Dong, L. Ratliff, H. Ohlsson, and S. Sastry. "A dynamical systems approach to energy disaggregation". In: *2013 IEEE 52nd Annu. Conf. on Decision and Control (CDC)*. 2013, pp. 6335–6340. DOI: 10.1109/CDC.2013.6760891 (cit. on p. 7).

[Don+13b] R. Dong, L. J. Ratliff, H. Ohlsson, and S. S. Sastry. "Energy disaggregation via adaptive filtering". In: *2013 51st Annu. Allerton Conf. on Communication, Control, and Computing (Allerton)*. 2013, pp. 173–180. DOI: 10.1109/Allerton.2013.6736521 (cit. on pp. 7, 61).

[Don+14] R. Dong, L. Ratliff, H. Ohlsson, and S. S. Sastry. "Fundamental Limits of Non-intrusive Load Monitoring". In: *Proc. of the 3rd Int. Conf. on High Confidence Networked Systems*. HiCoNS '14. Berlin, Germany: ACM, 2014, pp. 11–18. DOI: 10.1145/2566468.2566471 (cit. on pp. 38, 41, 45, 60, 77).

[Don+15] R. Dong, W. Krichene, A. M. Bayen, and S. S. Sastry. "Differential privacy of populations in routing games". In: *2015 54th IEEE Conference on Decision and Control (CDC)*. 2015, pp. 2798–2803. DOI: 10.1109/CDC.2015.7402640 (cit. on pp. 31, 43).

[Don06] D. L. Donoho. "Compressed sensing". In: *IEEE Transactions on Information Theory* 52.4 (2006), pp. 1289–1306. DOI: 10.1109/TIT.2006.871582 (cit. on p. 5).

[Dor+13] D. Dorsch, H. T. Jongen, and V. Shikhman. "On Structure and Computation of Generalized Nash Equilibria". In: *SIAM J. Optimization* 23.1 (2013), pp. 452–474. DOI: 10.1137/110822670 (cit. on p. 106).

[Duc+12] J. C. Duchi, M. I. Jordan, and M. J. Wainwright. "Privacy Aware Learning". In: *arXiv* (2012) (cit. on p. 45).

[Dwo06] C. Dwork. "Differential privacy". In: *Proc. of the Int. Colloq. on Automata, Languages and Programming*. Springer, 2006, pp. 1–12 (cit. on pp. 30, 33, 34, 45, 61, 83).

[EC10] V. L. Erickson and A. E. Cerpa. "Occupancy based demand response HVAC control strategy". In: *Proceedings of the 2nd ACM Workshop on Embedded Sensing Systems for Energy-Efficiency in Building*. ACM. 2010, pp. 7–12 (cit. on pp. 80, 82).

[EIA11] U. EIA. "Annual energy review". In: *Energy Information Administration, US Department of Energy: Washington, DC www. eia. doe. gov/emeu/aer* (2011) (cit. on p. 80).

[EM+10] K. Ehrhardt-Martinez, K. A. Donnelly, and J. A. Laitner. *Advanced Metering Initiatives and Residential Feedback Programs: A Meta-Review for Household Electricity-Saving Opportunities*. Tech. rep. American Council for an Energy-Efficient Economy, 2010 (cit. on p. 5).

[FC15] M. Faisal and A. A. Cárdenas. "How the Quantity and Quality of Training Data Impacts Re-Identification of Smart Meter Users". In: *IEEE Smart Grid Communications Conf.* 2015 (cit. on p. 77).

[FZ99] L. Farinaccio and R. Zmeureanu. "Using a pattern recognition approach to disaggregate the total electricity consumption in a house into the major end-uses". In: *Energy and Buildings* 30 (1999), pp. 245–259 (cit. on p. 7).

[Fan+14] L. Fan, L. Bonomi, L. Xiong, and V. Sunderam. "Monitoring web browsing behavior with differential privacy". In: *Proceedings of the 23rd international conference on World wide web*. ACM. 2014, pp. 177–188 (cit. on p. 81).

[Far+15] F. Farokhi, I. Shames, and M. Cantoni. "Budget-constrained contract design for effort-averse sensors in averaging based estimation". In: *arXiv preprint arXiv:1509.08193* (2015) (cit. on p. 102).

[Faz+01] M. Fazel, H. Hindi, and S. P. Boyd. "A rank minimization heuristic with application to minimum order system approximation". In: *American Control Conference, 2001. Proceedings of the 2001*. Vol. 6. 2001, 4734–4739 vol.6. DOI: 10.1109/ACC.2001.945730 (cit. on p. 21).

[Fro+11] J. Froehlich, E. Larson, S. Gupta, G. Cohn, M. Reynolds, and S. Patel. "Disaggregated End-Use Energy Sensing for the Smart Grid". In: *IEEE Pervasive Computing* 10.1 (2011), pp. 28–39. DOI: 10.1109/MPRV.2010.74 (cit. on p. 11).

[GB08] M. Grant and S. Boyd. "Graph implementations for nonsmooth convex programs". In: *Recent Advances in Learning and Control*. Ed. by V. Blondel, S. Boyd, and H. Kimura. Lecture Notes in Control and Information Sciences. http://stanford.edu/~boyd/graph_dcp.html. Springer-Verlag Limited, 2008, pp. 95–110 (cit. on p. 25).

[GB14] M. Grant and S. Boyd. *CVX: Matlab Software for Disciplined Convex Programming, version 2.1*. http://cvxr.com/cvx. Mar. 2014 (cit. on p. 25).

[GG10] D Gyalistras and M Gwerder. "Use of weather and occupancy forecasts for optimal building climate control (OptiControl): Two Years Progress Report Main Report". In: *Terrestrial Systems Ecology ETH Zurich R&D HVAC Products, Building Technologies Division, Siemens Switzerland Ltd, Zug, Switzerland* (2010) (cit. on p. 83).

[GW95] M. X. Goemans and D. P. Williamson. "Improved Approximation Algorithms for Maximum Cut and Satisfiability Problems Using Semidefinite Programming". In: *J. ACM* 42.6 (Nov. 1995), pp. 1115–1145. DOI: 10.1145/227683.227684 (cit. on p. 21).

[Gao+15] H. Gao, C. H. Liu, W. Wang, J. Zhao, Z. Song, X. Su, J. Crowcroft, and K. K. Leung. "A survey of incentive mechanisms for participatory sensing". In: *IEEE Communications Surveys & Tutorials* 17.2 (2015), pp. 918–943 (cit. on p. 102).

[Gir+14] J. Giraldo, A. Cárdenas, E. Mojica-Nava, N. Quijano, and R. Dong. "Delay and sampling independence of a consensus algorithm and its application to smart grid privacy". In: *IEEE 53rd Annu. Conf. on Decision and Control*. 2014, pp. 1389–1394. DOI: 10.1109/CDC.2014.7039596 (cit. on p. 61).

[Gou+87] C. Gourieroux, A. Monfort, and E. Renault. "Kullback Causality Measures". In: *Annales d'conomie et de Statistique* 6/7 (1987), pp. 369–410. URL: http://www.jstor.org/stable/20075662 (cit. on p. 124).

[Goy+13] S. Goyal, H. A. Ingley, and P. Barooah. "Occupancy-based zone-climate control for energy-efficient buildings: Complexity vs. performance". In: *Applied Energy* 106 (2013), pp. 209–221 (cit. on pp. 83, 86, 88, 93).

[Gra+00] C. W. Granger, B.-N. Huang, and C.-W. Yang. "A bivariate causality between stock prices and exchange rates: Evidence from recent Asian flu". In: *The Quarterly Review of Economics and Finance* 40.3 (2000), pp. 337 –354. DOI: http://dx.doi.org/10.1016/S1062-9769(00)00042-9 (cit. on p. 122).

[Gra69] C. W. J. Granger. "Investigating Causal Relations by Econometric Models and Cross-spectral Methods". In: *Econometrica* 37.3 (1969), pp. 424–438. URL: http://www.jstor.org/stable/1912791 (cit. on p. 122).

[Gre+69] B. G. Greenberg, A.-L. A. Abul-Ela, W. R. Simmons, and D. G. Horvitz. "The Unrelated Question Randomized Response Model: Theoretical Framework". In: *Journal of the American Statistical Association* 64.326 (1969), pp. 520–539. DOI: 10.1080/01621459.1969.10500991 (cit. on p. 30).

[HK14] J. Hu and P. Karava. "Model predictive control strategies for buildings with mixed-mode cooling". In: *Building and Environment* 71 (2014), pp. 233–244 (cit. on p. 83).

[HV94] T. Han and S. Verdú. "Generalizing the Fano inequality". In: *IEEE Trans. Inf. Theory* 40.4 (1994), pp. 1247–1251. DOI: 10.1109/18.335943 (cit. on pp. 37, 40).

[Han+14] S. Han, U. Topcu, and G. J. Pappas. "Differentially Private Distributed Constrained Optimization". In: *arXiv* (2014) (cit. on pp. 31, 45).

[Hec+06] D. Heckerman, C. Meek, and G. Cooper. "A Bayesian Approach to Causal Discovery". In: *Innovations in Machine Learning: Theory and Applications*. Ed. by D. E. Holmes and L. C. Jain. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006, pp. 1–28. DOI: 10.1007/3-540-33486-6_1 (cit. on p. 124).

[Hoy+09] P. O. Hoyer, D. Janzing, J. M. Mooij, J. Peters, and P. B. Schölkopf. "Nonlinear causal discovery with additive noise models". In: *Advances in Neural Information Processing Systems 21*. Ed. by D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou. Curran Associates, Inc., 2009, pp. 689–696 (cit. on p. 124).

[Hsu+14] J. Hsu, Z. Huang, A. Roth, and Z. S. Wu. "Jointly Private Convex Programming". In: *arXiv* (2014) (cit. on pp. 31, 45).

[Hua+14a] Z. Huang, S. Mitra, and N. Vaidya. "Differentially Private Distributed Optimization". In: *arXiv* (2014) (cit. on p. 45).

[Hua+14b] Z. Huang, Y. Wang, S. Mitra, and G. E. Dullerud. "On the Cost of Differential Privacy in Distributed Control Systems". In: *Proc. of the 3rd Int. Conf. on High Confidence Networked Systems*. HiCoNS '14. Berlin, Germany: ACM, 2014, pp. 105–114. DOI: 10.1145/2566468.2566474 (cit. on p. 31).

[Hua02] Y. Hua. "Blind methods of system identification". In: *Circuits, Systems and Signal Processing* 21.1 (2002), pp. 91–108. DOI: 10.1007/BF01211654 (cit. on p. 19).

[IH91] I. Ibragimov and R. Has'minskii. *Statistical Estimation – Asymptotic Theory*. Springer-Verlag New York, 1991 (cit. on p. 40).

[IR15] G. W. Imbens and D. B. Rubin. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge University Press, 2015 (cit. on p. 122).

[JW12] M. J. Johnson and A. S. Willsky. "Bayesian Nonparametric Hidden Semi-Markov Models". In: *arXiv:1203.1365* (Mar. 2012). eprint: 1203.1365 (cit. on pp. 7, 11, 17).

[Jia+15] J. Jiao, T. A. Courtade, K. Venkat, and T. Weissman. "Justification of Logarithmic Loss via the Benefit of Side Information". In: *IEEE Transactions on Information Theory* 61.10 (2015), pp. 5357–5365. DOI: 10.1109/TIT.2015.2462848 (cit. on pp. 36, 83).

[Jia+16] R. Jia, R. Dong, S. S. Sastry, and C. Spanos. "Privacy-Enhanced Architecture for Occupancy-based HVAC Control". In: *8th ACM/IEEE International Conference on Cyber-Physical Systems (ICCPS)*. 2016 (cit. on pp. 31, 82).

[Jin+14] M. Jin, R. Jia, Z. Kang, I. C. Konstantakopoulos, and C. J. Spanos. "Presencesense: Zero-training algorithm for individual presence detection based on power monitoring". In: *Proceedings of the 1st ACM Conference on Embedded Systems for Energy-Efficient Buildings*. ACM. 2014, pp. 1–10 (cit. on p. 80).

[Jin+15] M. Jin, N. Bekiaris-Liberis, K. Weekly, C. Spanos, and A. Bayen. "Sensing by Proxy: Occupancy Detection Based on Indoor CO2 Concentration". In: *UBICOMM 2015* (2015), p. 14 (cit. on pp. 80, 96).

[Jud+11] A. Juditsky, A. Nemirovski, and C. Tauvel. "Solving variational inequalities with stochastic mirror-prox algorithm". In: *Stoch. Syst.* 1.1 (2011), pp. 17–58. DOI: 10.1214/10-SSY011 (cit. on p. 44).

[KB11] A. Kelman and F. Borrelli. "Bilinear model predictive control of a HVAC system using sequential quadratic programming". In: *Ifac world congress*. Vol. 18. 2011, pp. 9869–9874 (cit. on pp. 86, 93).

[KJ11] J. Z. Kolter and M. J. Johnson. "REDD: A Public Data Set for Energy Disaggregation Research". In: *Proc. of the SustKDD Workshop on Data Mining Applications in Sustainability*. 2011 (cit. on pp. 6, 9, 17).

[KJ12] J. Z. Kolter and T. Jaakkola. "Approximate Inference in Additive Factorial HMMs with Application to Energy Disaggregation". In: *Proc. of the Int. Conf. on Artificial Intelligence and Statistics*. 2012 (cit. on pp. 7, 11, 17).

[KN10] J. Z. Kolter and A. Y. Ng. "Energy Disaggregation via Discriminative Sparse Coding". In: *Neural Information Processing Systems*. 2010 (cit. on pp. 7, 17).

[Kal02] O. Kallenberg. *Foundations of Modern Probability*. Springer, 2002 (cit. on pp. 89, 125).

[Kee10] R. W. Keener. *Theoretical Statistics: Topics for a Core Course*. Springer, 2010 (cit. on pp. 37, 72, 74, 76).

[Kha+15] M. A.A. H. Khan, H. Hossain, and N. Roy. "Infrastructure-less occupancy detection and semantic localization in smart environments". In: *proceedings of the 12th EAI International Conference on Mobile and Ubiquitous Systems*. ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering). 2015, pp. 51–60 (cit. on p. 80).

[Kim+11] H. Kim, M. Marwah, M. F. Arlitt, G. Lyon, and J. Han. "Unsupervised Disaggregation of Low Frequency Power Measurements". In: *SDM'11*. 2011, pp. 747–758 (cit. on pp. 7, 11).

[Kle+14] W. Kleiminger, S. Santini, and F. Mattern. "Smart heating control with occupancy prediction: how much can one save?" In: *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication*. ACM. 2014, pp. 947–954 (cit. on p. 80).

[Kri+15a] S. Krichene, W. Krichene, R. Dong, and A. Bayen. "Convergence of Stochastic Mirror Descent and applications to distributed optimization". In: *Internation Conference on Machine Learning (ICML), in review*. 2015 (cit. on pp. 44, 52).

[Kri+15b] W. Krichene, S. Krichene, and A. Bayen. "Convergence of Mirror Descent Dynamics in the Routing Game". In: *European Control Conference (ECC), accepted*. 2015 (cit. on pp. 44, 51).

[Kri+15c] W. Krichene, B. Drighès, and A. Bayen. "Learning Nash equilibria in Congestion Games". In: *SIAM Journal on Control and Optimization (SICON), to appear* (2015) (cit. on p. 44).

[Kur+11] K. Kursawe, G. Danezis, and M. Kohlweiss. "Privacy-friendly Aggregation for the Smart-grid". In: *Proc. of the 11th Int. Conf. on Privacy Enhancing Technologies*. PETS'11. Waterloo, ON, Canada: Springer-Verlag, 2011, pp. 175–191 (cit. on pp. 30, 65).

[LC73] L. Le Cam. "Convergence of Estimates Under Dimensionality Restrictions". English. In: *The Ann. of Statistics* 1.1 (1973), pp. 38–53. URL: http://www.jstor.org/stable/2958155 (cit. on pp. 37, 39).

[LNP14] J. Le Ny and G. Pappas. "Differentially Private Filtering". In: *IEEE Trans. Autom. Control* 59 (2014), pp. 341–354. DOI: 10.1109/TAC.2013.2283096 (cit. on pp. 31, 45, 46).

[LS91] L. Lovsz and A. Schrijver. "Cones of Matrices and Set-Functions and 01 Optimization". In: *SIAM Journal on Optimization* 1.2 (1991), pp. 166–190. DOI: 10.1137/0801013 (cit. on p. 21).

[LZ13] N. Lu and Y. Zhang. "Design Considerations of a Centralized Load Controller Using Thermostatically Controlled Appliances for Continuous Regulation Reserves". In: *IEEE Trans. on Smart Grid* 4.2 (2013), pp. 914–921. DOI: 10.1109/TSG.2012.2222944 (cit. on p. 65).

[Lau01] S. L. Lauritzen. *Causal Inference from Graphical Models*. 2001 (cit. on p. 124).

[Li+10] F. Li, B. Luo, and P. Liu. "Secure Information Aggregation for Smart Grids Using Homomorphic Encryption". In: *2010 1st IEEE Int. Conf. on Smart Grid Communications (SmartGridComm)*. 2010, pp. 327–332. DOI: 10.1109/SMARTGRID.2010.5622064 (cit. on pp. 30, 65).

[Li+15] J. Li, S. Ma, T. D. Le, L. Liu, and J. Liu. "Causal Decision Trees". In: *CoRR* abs/1508.03812 (2015). URL: http://arxiv.org/abs/1508.03812 (cit. on p. 124).

[Lis+10] M. Lisovich, D. Mulligan, and S. Wicker. "Inferring Personal Information from Demand-Response Systems". In: *IEEE Security Privacy* 8 (2010), pp. 11–20. DOI: 10.1109/MSP.2010.40 (cit. on pp. 29, 61, 64, 80).

[Lju99] L. Ljung. *System Identification – Theory for the User*. Prentice Hall, 1999 (cit. on p. 18).

[Lof04] J. Lofberg. "YALMIP : a toolbox for modeling and optimization in MATLAB". In: *Computer Aided Control Systems Design, 2004 IEEE International Symposium on*. 2004, pp. 284–289. DOI: 10.1109/CACSD.2004.1393890 (cit. on p. 25).

[Lov83] L. Lovász. "Submodular functions and convexity". In: *Mathematical Programming The State of the Art: Bonn 1982*. Ed. by A. Bachem, B. Korte, and M. Grötschel. Springer, 1983, pp. 235–257. DOI: 10.1007/978-3-642-68874-4_10 (cit. on pp. 127, 128).

[Lu12] N. Lu. "An Evaluation of the HVAC Load Potential for Providing Load Balancing Service". In: *IEEE Trans. on Smart Grid* 3.3 (2012), pp. 1263–1270. DOI: 10.1109/TSG.2012.2183649 (cit. on p. 65).

[Mat+13] J. Mathieu, S. Koch, and D. Callaway. "State Estimation and Control of Electric Loads to Manage Real-Time Energy Imbalance". In: *IEEE Trans. Power Syst.* 28.1 (2013), pp. 430–440. DOI: 10.1109/TPWRS.2012.2204074 (cit. on pp. 66–68, 72).

[Mou+13] S. Moura, J. Bendtsen, and V. Ruiz. "Observer design for boundary coupled PDEs: Application to thermostatically controlled loads in smart grids". In: *IEEE 52nd Annu. Conf. on Decision and Control (CDC)*. 2013, pp. 6286–6291. DOI: 10.1109/CDC.2013.6760883 (cit. on p. 66).

[NP33] J. Neyman and E. S. Pearson. "On the Problem of the Most Efficient Tests of Statistical Hypotheses". English. In: *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character* 231 (1933), pp. 289–337. URL: http://www.jstor.org/stable/91247 (cit. on pp. 37, 38).

[NS06] A. Narayanan and V. Shmatikov. "How To Break Anonymity of the Netflix Prize Dataset". In: *arXiv* (2006) (cit. on p. 30).

[NY83] A. S. Nemirovsky and D. B. Yudin. *Problem complexity and method efficiency in optimization*. Wiley-Interscience series in discrete mathematics. Wiley, 1983 (cit. on p. 51).

[Nag+15] S. Nagarathinam, A. Vasan, V. Ramakrishna P, S. R. Iyer, V. Sarangan, and A. Sivasubramaniam. "Centralized Management of HVAC Energy in Large Multi-AHU Zones". In: *Proceedings of the 2nd ACM International Conference on Embedded Systems for Energy-Efficient Built Environments*. ACM. 2015, pp. 157–166 (cit. on p. 92).

[Nem+78] G. L. Nemhauser, L. A. Wolsey, and M. L. Fisher. "An analysis of approximations for maximizing submodular set functions". In: *Mathematical Programming* 14.1 (1978), pp. 265–294. DOI: 10.1007/BF01588971 (cit. on p. 130).

[Nes98] Y. Nesterov. "Semidefinite relaxation and nonconvex quadratic optimization". In: *Optimization Methods and Software* 9.1-3 (1998), pp. 141–160. DOI: 10.1080/10556789808805690 (cit. on p. 21).

[Nis04] H. Nissenbaum. "Privacy as Contextual Integrity". In: *Washington Law Review* (2004) (cit. on p. 31).

[Ohl+13] H. Ohlsson, A. Y. Yang, R. Dong, and S. Sastry. "Nonlinear basis pursuit". In: *Signals, Systems and Computers, 2013 Asilomar Conference on*. 2013, pp. 115–119. DOI: 10.1109/ACSSC.2013.6810285 (cit. on p. 25).

[Ohl+14] H. Ohlsson, L. Ratliff, R. Dong, and S. S. Sastry. "Blind Identification via Lifting". In: {*IFAC*} *Proceedings Volumes* 47.3 (2014). 19th {IFAC} World Congress, pp. 10367 –10372. DOI: http://dx.doi.org/10.3182/20140824-6-ZA-1003.02567 (cit. on p. 18).

[Old+12] F. Oldewurtel, A. Parisio, C. N. Jones, D. Gyalistras, M. Gwerder, V. Stauch, B. Lehmann, and M. Morari. "Use of model predictive control and weather forecasts for energy efficient building climate control". In: *Energy and Buildings* 45 (2012), pp. 15–27 (cit. on p. 83).

[PCF12a] F. du Pin Calmon and N. Fawaz. "Privacy against statistical inference". In: *2012 50th Annu. Allerton Conf. on Commun., Control, and Computing (Allerton)*. 2012, pp. 1401–1408. DOI: 10.1109/Allerton.2012.6483382 (cit. on pp. 31, 35, 36, 82–84).

[PCF12b] F. du Pin Calmon and N. Fawaz. "Privacy against statistical inference". In: *Communication, Control, and Computing (Allerton), 2012 50th Annual Allerton Conference on*. IEEE. 2012, pp. 1401–1408 (cit. on p. 36).

[Par+12] O. Parson, S. Ghosh, M. Weal, and A. Rogers. "Nonintrusive load monitoring using prior models of general appliance types". In: *Proc. of the 26th AAAI Conf. on Artificial Intelligence*. 2012 (cit. on pp. 7, 11).

[Pat12] S. Pattem. "Unsupervised Disaggregation for Non-intrusive Load Monitoring". In: *11th International Conference on Machine Learning and Applications (ICMLA), 2012*. Vol. 2. IEEE. 2012, pp. 515–520 (cit. on p. 7).

[Pea09] J. Pearl. *Causality: Models, Reasoning and Inference*. Cambridge University Press, 2009 (cit. on pp. 122–125).

[Pea98] J. Pearl. "Graphical Models for Probabilistic and Causal Reasoning". In: *Quantified Representation of Uncertainty and Imprecision*. Ed. by P. Smets. Dordrecht: Springer Netherlands, 1998, pp. 367–389. DOI: 10.1007/978-94-017-1735-9_12 (cit. on p. 124).

[Per+12] C. Perfumo, E. Kofman, J. H. Braslavsky, and J. K. Ward. "Load management: Model-based control of aggregate power for populations of thermostatically controlled loads". In: *Energy Conversion and Management* 55 (2012), pp. 36 –48. DOI: http://dx.doi.org/10.1016/j.enconman.2011.10.019 (cit. on p. 65).

[Pet04] J. Petzold. "Augsburg Indoor Location Tracking Benchmarks". In: (2004) (cit. on p. 91).

[RD11] A. Rial and G. Danezis. "Privacy-preserving Smart Metering". In: *Proc. of the 10th Annu. ACM Workshop on Privacy in the Electronic Society*. WPES '11. Chicago, Illinois, USA: ACM, 2011, pp. 49–60. DOI: 10.1145/2046556.2046564 (cit. on pp. 30, 65).

[Rah+12] D. Rahayu, B. Narayanaswamy, S. Krishnaswamy, C. Labbé, and D. P. Seetharam. "Learning to be energy-wise: discriminative methods for load disaggregation". In: *Third International Conference on Future Energy Systems: Where Energy, Computing and Communication Meet (e-Energy), 2012*. IEEE. 2012, pp. 1–4 (cit. on p. 7).

[Raj+11] S. R. Rajagopalan, L. Sankar, S. Mohajer, and H. V. Poor. "Smart meter privacy: A utility-privacy framework". In: *Smart Grid Communications (SmartGridComm), 2011 IEEE International Conference on*. 2011, pp. 190–195. DOI: 10.1109/SmartGridComm.2011.6102315 (cit. on pp. 31, 36, 82, 83, 89).

[Rat+16] L. J. Ratliff, C. Barreto, R. Dong, H. Ohlsson, A. Cárdenas, and S. S. Sastry. "Effects of Risk on Privacy Contracts for Demand-Side Management (under review)". In: *ACM Transactions on Internet Technology* (2016) (cit. on pp. 38, 82).

[Rec+10] B. Recht, M. Fazel, and P. A. Parrilo. "Guaranteed Minimum-Rank Solutions of Linear Matrix Equations via Nuclear Norm Minimization". In: *SIAM Review* 52.3 (2010), pp. 471–501. DOI: 10.1137/070697835 (cit. on p. 24).

[Roc70] R. T. Rockafellar. *Convex Analysis*. 1970 (cit. on p. 24).

[Roe+12] F. Roesner, J. Fogarty, and T. Kohno. "User Interface Toolkit Mechanisms for Securing Interface Elements". In: *Proceedings of the 25th Annual ACM Symposium on User Interface Software and Technology*. UIST '12. Cambridge, Massachusetts, USA: ACM, 2012, pp. 239–250. DOI: 10.1145/2380116.2380147 (cit. on p. 32).

[Rou07] T. Roughgarden. "Routing games". In: *Algorithmic game theory*. Cambridge University Press, 2007. Chap. 18, pp. 461–486 (cit. on p. 44).

[Rub74] D. B. Rubin. "Estimating causal effects of treatments in randomized and nonrandomized studies". In: *Journal of Educational Psychology* 66(5) (1974), pp. 688–701. DOI: 10.1037/h0037350 (cit. on p. 121).

[Rui+09] N. Ruiz, I. Cobelo, and J. Oyarzabal. "A Direct Load Control Model for Virtual Power Plant Management". In: *IEEE Trans. Power Syst.* 24.2 (2009), pp. 959–966. DOI: 10.1109/TPWRS.2009.2016607 (cit. on p. 66).

[SC+14] J. Soria-Comas, J. Domingo-Ferrer, D. Sánchez, and S. Martínez. "Enhancing data utility in differential privacy via microaggregation-based k-anonymity". In: *The VLDB Journal* 23.5 (2014), pp. 771–794 (cit. on p. 81).

[Sal+13] S. Salamatian, A. Zhang, F. du Pin Calmon, S. Bhamidipati, N. Fawaz, B. Kveton, P. Oliveira, and N. Taft. "How to hide the elephant-or the donkey-in the room: Practical privacy against statistical inference for large data". In: *IEEE GlobalSIP* (2013) (cit. on pp. 35, 45).

[San+11] L. Sankar, S. Kar, R. Tandon, and H. Poor. "Competitive privacy in the smart grid: An information-theoretic approach". In: *2011 IEEE Int. Conf. on Smart Grid Communications (SmartGridComm)*. 2011, pp. 220–225. DOI: 10.1109/SmartGridComm.2011.6102322 (cit. on pp. 35, 45).

[San+13] L. Sankar, S. Rajagopalan, and H. Poor. "Utility-Privacy Tradeoffs in Databases: An Information-Theoretic Approach". In: *IEEE Trans. Inf. Forens. Security* 8 (2013), pp. 838–852. DOI: 10.1109/TIFS.2013.2253320 (cit. on pp. 31, 65).

[San01] W. H. Sandholm. "Potential games with continuous player sets". In: *Journal of Economic Theory* 97.1 (2001), pp. 81–108 (cit. on p. 44).

[Sat75] Y. Sato. "A Method of Self-Recovering Equalization for Multilevel Amplitude-Modulation Systems". In: *IEEE Transactions on Communications* 23.6 (1975), pp. 679–682. DOI: 10.1109/TCOM.1975.1092854 (cit. on p. 19).

[Sch03] A. Schrijver. *Combinatorial Optimization: Polyhedra and Efficiency*. Springer-Verlag Berlin Heidelberg, 2003 (cit. on pp. 127, 132).

[Sco+15] D. Scobee, L. Ratliff, R. Dong, H. Ohlsson, M. Verhaegen, and S. S. Sastry. "Nuclear norm minimization for blind subspace identification (N2BSID)". In: *2015 54th IEEE Conference on Decision and Control (CDC)*. 2015, pp. 2127–2132. DOI: 10.1109/CDC.2015.7402521 (cit. on p. 27).

[Sha+12] H. Shao, M. Marwah, and N. Ramakrishnan. "A Temporal Motif Mining Approach to Unsupervised Energy Disaggregation". In: *Proceedings of 1st International Non-Intrusive Load Monitoring Workshop*. 2012 (cit. on pp. 7, 17).

[Sha+16] R. Shay, S. Komanduri, A. L. Durity, P. S. Huh, M. L. Mazurek, S. M. Segreti, B. Ur, L. Bauer, N. Christin, and L. F. Cranor. "Designing Password Policies for Strength and Usability". In: *ACM Trans. Inf. Syst. Secur.* 18.4 (May 2016), 13:1–13:34. DOI: 10.1145/2891411 (cit. on p. 32).

[Sho+11] R. Shokri, G. Theodorakopoulos, J.-Y. Le Boudec, and J.-P. Hubaux. "Quantifying location privacy". In: *Security and privacy (sp), 2011 ieee symposium on*. IEEE. 2011, pp. 247–262 (cit. on p. 80).

[Sho87] N. Shor. "Quadratic optimization problems". In: *Soviet Journal of Computer and Systems Sciences* (1987) (cit. on p. 21).

[Smi12] G. Smith. *Marijuana bust shines light on utilities*. Post and Courier. 2012 (cit. on p. 29).

[Sol02] D. J. Solove. "Conceptualizing Privacy". In: *California Law Review* 90 (2002), p. 1087 (cit. on pp. 31, 43).

[Sta+06] S. Stańczak, M. Wiczanowski, and H. Boche. "Chapter 2: On the Positive Solution to a Linear System with Nonnegative Coefficients". In: *Resource Allocation in Wireless Networks: Theory and Algorithms*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006, pp. 51–68. DOI: 10.1007/11818762_2 (cit. on p. 114).

[Sun+99] L. Sun, W. Liu, and A. Sano. "Identification of a dynamical system with input nonlinearity". In: *IEE Proceedings - Control Theory and Applications* 146.1 (1999), pp. 41–51. DOI: 10.1049/ip-cta:19990371 (cit. on p. 19).

[Swe02] L. Sweeney. "k-anonymity: a model for protecting privacy". In: *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems* (2002) (cit. on pp. 30, 32).

[TG09] G. Taban and V. Gligor. "Privacy-Preserving Integrity-Assured Data Aggregation in Sensor Networks". In: *Int. Conf. on Computational Science and Engineering*. Vol. 3. 2009, pp. 168–175. DOI: 10.1109/CSE.2009.389 (cit. on pp. 30, 65).

[TK14] Z. Tufekci and B. King. *We Can't Trust Uber*. New York Times. 2014 (cit. on p. 44).

[Ton+91] L. Tong, G. Xu, and T. Kailath. "A new approach to blind identification and equalization of multipath channels". In: *Signals, Systems and Computers, 1991. 1991 Conference Record of the Twenty-Fifth Asilomar Conference on*. 1991, 856–860 vol.2. DOI: 10.1109/ACSSC.1991.186568 (cit. on p. 19).

[Tsy09] A. B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer New York, 2009 (cit. on p. 39).

[Van+13] S. VanVaerenbergh, J. Va, and I. Santamara. "Blind Identification of SIMO Wiener Systems Based on Kernel Canonical Correlation Analysis". In: *IEEE Transactions on Signal Processing* 61.9 (2013), pp. 2219–2230. DOI: 10.1109/TSP.2013.2248004 (cit. on p. 19).

[Ven+15] P. Venkitasubramaniam, J. Yao, and P. Pradhan. "Information-theoretic security in stochastic control systems". In: *Proceedings of the IEEE* 103.10 (2015), pp. 1914–1931 (cit. on p. 82).

[Ven13] P. Venkitasubramaniam. "Privacy in stochastic control: A Markov Decision Process perspective". In: *2013 51st Annu. Allerton Conf. on Communication, Control, and Computing (Allerton)*. 2013, pp. 381–388. DOI: 10.1109/Allerton.2013.6736549 (cit. on pp. 35, 45).

[WT14] X. Wang and P. Tague. "Non-Invasive User Tracking via Passive Sensing: Privacy Risks of Time-Series Occupancy Measurement". In: *Proceedings of the 2014 Workshop on Artificial Intelligent and Security Workshop*. ACM. 2014, pp. 113–124 (cit. on pp. 80, 81, 85).

[Wan+07] J. Wang, A. Sano, D. Shook, T. Chen, and B. Huang. "A blind approach to closed-loop identification of Hammerstein systems". In: *International Journal of Control* 80.2 (2007), pp. 302–313. DOI: 10.1080/00207170601026505 (cit. on p. 19).

[Wan+10] J. Wang, A. Sano, T. Chen, and B. Huang. "A Blind Approach to Identification of Hammerstein Systems". In: *Block-oriented Nonlinear System Identification*. Ed. by F. Giri and E.-W. Bai. London: Springer London, 2010, pp. 293–312. DOI: 10.1007/978-1-84996-513-2_18 (cit. on p. 19).

[Wan+14] H. Wang, L. Sun, and E. Bertino. "Building access control policy model for privacy preserving and testing policy conflicting problems". In: *Journal of Computer and System Sciences* 80.8 (2014). Special Issue on Theory and Applications in Parallel and Distributed Computing Systems, pp. 1493 –1503. DOI: http://dx.doi.org/10.1016/j.jcss.2014.04.017 (cit. on pp. 81, 82).

[War65] S. L. Warner. "Randomized Response: A Survey Technique for Eliminating Evasive Answer Bias". In: *Journal of the American Statistical Association* 60.309 (1965). PMID: 12261830, pp. 63–69. DOI: 10.1080/01621459.1965.10480775 (cit. on p. 30).

[Wat92] G. Watson. "Characterization of the subdifferential of some matrix norms". In: *Linear Algebra and its Applications* 170 (1992), pp. 33 –45. DOI: http://dx.doi.org/10.1016/0024-3795(92)90407-2 (cit. on p. 24).

[Wu+15] C. Wu, J. Thai, S. Yadlowsky, A. Pozdnoukhov, and A. Bayen. "Cellpath: Fusion of cellular and traffic sensor data for route flow estimation via convex optimization". In: *Transportation Research Part C: Emerging Technologies* 59 (2015). Special Issue on International Symposium on Transportation and Traffic Theory, pp. 111 –128. DOI: http://dx.doi.org/10.1016/j.trc.2015.05.004 (cit. on p. 4).

[YBG15] Z. Yang and B. Becerik-Gerber. "Cross-space building occupancy modeling by contextual information based learning". In: *Proceedings of the 2nd ACM International Conference on Embedded Systems for Energy-Efficient Built Environments*. ACM. 2015, pp. 177–186 (cit. on p. 80).

[Yu97] B. Yu. "Assouad, Fano, and Le Cam". In: *Festschrift for Lucien Le Cam*. Springer, 1997, pp. 423–435 (cit. on pp. 37, 39, 40).

[ZR11] M. Zeifman and K. Roth. "Nonintrusive appliance load monitoring: Review and outlook". In: *IEEE Transactions on Consumer Electronics* 57 (2011), pp. 76–84. DOI: 10.1109/TCE.2011.5735484 (cit. on p. 6).

[Zer+99] A. Zerva, A. Petropulu, and P.-Y. Bard. "Blind deconvolution methodology for site-response evaluation exclusively from ground-surface seismic recordings". In: *Soil Dynamics and Earthquake Engineering* 18.1 (1999), pp. 47 –57. DOI: http://dx.doi.org/10.1016/S0267-7261(98)00005-0 (cit. on p. 19).

[Zho+16] D. Zhou, M. Balandat, and C. J. Tomlin. "Residential Demand Response Targeting Using Machine Learning with Observational Data". In: *55th IEEE Conference on Decision and Control (CDC)*. 2016 (cit. on p. 122).

[Cal] California Public Utilities Commission. *Decision Adopting Rules to Protect the Privacy and Security of the Electricity Usage Data of the Customers of Pacific Gas and Electric Company, Southern California Edison Company, and San Diego Gas & Electric Company* (cit. on p. 64).

[Cal13] California Energy Commission. *Docket No. 13-IEP-1F: Increasing Demand Response Capabilities in California*. 2013 (cit. on p. 64).

[Cal14] California Independent System Operators. *Business Practice Manual for Market Operations*. 2014 (cit. on p. 72).

[Dep] Department of Energy. *Data Access and Privacy Issues Related To Smart Grid Tchnologies* (cit. on p. 64).

[Gov11] Government Accountability Office. *ELECTRICITY GRID MODERNIZATION: Progress Being Made on Cybersecurity Guidelines, but Key Challenges Remain to be Addressed*. 2011 (cit. on p. 29).

[Nor] North American Energy Standards Board. *NAESB Privacy Policy* (cit. on p. 64).

[Oba11] Obama Administration. *A Policy Framework for the 21st Century Grid: Enabling Our Secure Energy Future*. NSTC, 2011 (cit. on p. 64).

[Pub] Public Utility Commission of Texas. *Electric Substantive Rules – Chapter 25* (cit. on p. 64).

[The] The Smart Grid Interoperability Panel - Cyber Security Working Group. *NISTIR7628 – Guidelines for Smart Grid Cyber Security: Vol. 2, Privacy and the Smart Grid* (cit. on p. 64).

[Tra11] Transportation Research Board. *Transportation Research Board 2011 Annual Report*. Tech. rep. The National Academies, 2011 (cit. on p. 43).